Modern Discrete Probability

An Essential Toolkit

Sébastien Roch

November 19, 2020
Contents

Preface v

1 Introduction 1

1.1 Background ................................................................. 1

1.1.1 Review of graph theory ........................................... 1

1.1.2 Review of Markov chain theory ................................. 6

1.2 Percolation ................................................................. 11

1.3 Random graphs ............................................................. 11

1.4 Markov random fields .................................................... 12

1.5 Random walks on graphs ............................................... 13

2 Moments and tails 17

2.1 Background ................................................................. 17

2.1.1 Definitions ............................................................ 17

2.1.2 Basic inequalities ................................................... 18

2.2 First moment method .................................................... 21

2.2.1 The probabilistic method ......................................... 21

2.2.2 Union bound .......................................................... 24

2.2.3 Random permutations: longest increasing subsequence ... 26

2.2.4 Percolation on \( \mathbb{Z}^2 \): non-trivial threshold .......... 28

2.3 Second moment method ................................................ 32

2.3.1 Paley-Zygmund inequality ........................................ 32

2.3.2 Erdős-Rényi: subgraph containment and connectivity .. 35

2.3.3 Percolation on trees: critical value and branching number 42

2.4 Chernoff-Cramér method .............................................. 46

2.4.1 Tail bounds via the moment-generating function .......... 47

2.4.2 Some examples ....................................................... 50

2.4.3 Sub-Gaussian random variables ................................. 52

2.4.4 Sub-exponential random variables ............................. 56
4.3.3 \(\triangleright\) Random walk on trees: speed .................. 212
4.3.4 Correlation inequalities: FKG and Holley’s inequalities . . . 213
4.3.5 \(\triangleright\) Erdős-Rényi graph: Janson’s inequality and application to the containment problem .................. 218
4.3.6 \(\triangleright\) Percolation on \(\mathbb{Z}^2\): RSW theory and a proof of Harris’ theorem .................. 221
4.4 Couplings of Markov chains ................................. 229
4.4.1 Bounding the mixing time via coupling .................... 229
4.4.2 \(\triangleright\) Markov chains: mixing on cycles, hypercubes, and trees . . 231
4.4.3 Path coupling ........................................ 239
4.4.4 \(\triangleright\) Ising model: Glauber dynamics at high temperature . . 242
4.4.5 \(\triangleright\) From approximate sampling to approximate counting . . 245
4.5 Chen-Stein method ........................................ 246
4.5.1 Main bounds and examples .............................. 246
4.5.2 Motivation and proof ................................... 252
4.5.3 \(\triangleright\) Erdős-Rényi: subgraph containment at the threshold . . 261
5 Spectral methods ............................................ 268
5.1 Spectral techniques for reversible Markov chains ............... 268
5.1.1 Spectral gap ........................................ 268
5.1.2 \(\triangleright\) Markov chains: random walk on cycles and hypercubes revisited ........................................ 277
5.1.3 Canonical paths and comparison .......................... 281
5.1.4 Spectral radius ....................................... 281
5.2 Expansion ................................................ 289
5.2.1 Bottleneck ratio ...................................... 289
5.2.2 \(\triangleright\) Markov chains: random walk on trees, cycles, and hypercubes revisited ........................................ 300
5.2.3 \(\triangleright\) Ising model: Glauber dynamics on the complete graph and expander graphs ........................................ 302
6 Branching processes ........................................ 309
6.1 Background ............................................... 309
6.1.1 Basic definitions ..................................... 309
6.1.2 Extinction ........................................... 311
6.1.3 \(\triangleright\) Bond percolation on Galton-Watson trees ........ 316
6.1.4 \(\triangleright\) Random walk on Galton-Watson trees ............ 316
6.2 Random-walk representation ................................ 317
6.2.1 Exploration process .................................. 317
6.2.2 Duality principle .................................... 319
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.3 Hitting-time theorem</td>
<td>320</td>
</tr>
<tr>
<td>6.3 Comparison to branching processes</td>
<td>323</td>
</tr>
<tr>
<td>6.3.1 ⊳ Percolation on trees: critical exponents</td>
<td>323</td>
</tr>
<tr>
<td>6.3.2 ⊳ Random binary search trees: height</td>
<td>326</td>
</tr>
<tr>
<td>6.3.3 ⊳ Erdős-Rényi graph: the phase transition</td>
<td>327</td>
</tr>
<tr>
<td>6.4 Further applications</td>
<td>344</td>
</tr>
<tr>
<td>6.4.1 ⊳ Uniform random trees: local limit</td>
<td>344</td>
</tr>
<tr>
<td>A Appendix</td>
<td>348</td>
</tr>
<tr>
<td>A.1 Some definitions and useful facts</td>
<td>348</td>
</tr>
<tr>
<td>A.1.1 Binomial coefficients</td>
<td>348</td>
</tr>
<tr>
<td>A.1.2 Conditional expectation: definition and properties</td>
<td>349</td>
</tr>
<tr>
<td>A.1.3 A Taylor expansion</td>
<td>350</td>
</tr>
<tr>
<td>A.1.4 Spectral representation of reversible matrices</td>
<td>350</td>
</tr>
<tr>
<td>A.1.5 A fact about trees</td>
<td>350</td>
</tr>
<tr>
<td>A.1.6 A Poincaré inequality</td>
<td>351</td>
</tr>
</tbody>
</table>
Preface

These lecture notes form the basis for a one-semester introduction to modern discrete probability with an emphasis on essential techniques used throughout the field. The material covered here is considered in much greater depth in the following excellent texts. (Consult the bibliography for complete citations.)


[RAS] Firas Rassoul-Agha and Timo Seppäläinen. *A course on large deviations with an introduction to Gibbs measures.*


In fact, these notes are meant as a summary of some the basic topics covered in these more specialized references. My hope is that, by the end of the course, students will have picked up enough background to learn the advanced material on
their own with greater ease. Many more sources were used in putting these notes together. They are acknowledged in the “Bibliographic remarks” at the end of each chapter. These notes were also strongly influenced by graduate courses of David Aldous, Elchanan Mossel, Yuval Peres, and Alistair Sinclair at UC Berkeley. Some of the material covered here can also be found in [Gri10a], with a different emphasis.

I assume that students have taken at least one semester of graduate probability at the level of [Dur10]. I am also particularly fond of [Wil91]. Some familiarity with countable Markov chain theory will be necessary, as covered for instance in [Dur10, Chapter 6]. An advanced undergraduate treatment such as [Dur12] will largely suffice however. See also [LPW06, Chapter 1].
Chapter 1

Introduction

In this chapter, we describe a few essential discrete probability models. We also formulate various questions about these models and give references to later sections where these questions are partly answered. The list of models described here, which is far from exhaustive, contains mostly models defined on graphs.

1.1 Background

We start with a brief review of graph theory and Markov chains.

1.1.1 Review of graph theory

Basic definitions  An undirected graph (or graph for short) is a pair $G = (V, E)$ where $V$ is the set of vertices (or nodes or sites) and

$$E \subseteq \{\{u, v\} : u, v \in V\},$$

is the set of edges (or bonds). See Figure 1.1 for an example. We occasionally write $V(G)$ and $E(G)$ for the vertices and edges of the graph $G$. The set of vertices $V$ is either finite or countably infinite. Edges of the form $\{u\}$ are called self-loops. In general, we do not allow $E$ to be a multiset. When $E$ is a multiset, $G$ is called a multigraph.

A vertex $v \in V$ is incident with an edge $e \in E$ if $v \in e$. The incident vertices of an edge are sometimes called endvertices. Two vertices $u, v \in V$ are adjacent (or neighbors), denoted by $u \sim v$, if $\{u, v\} \in E$. The set of adjacent vertices of $v$,
denoted by \( N(v) \), is called the *neighborhood* of \( v \) and its size, i.e., \( \delta(v) := |N(v)| \), is the *degree* of \( v \). A vertex \( v \) with \( \delta(v) = 0 \) is called *isolated*. A graph is called *\( d \)-regular* if all its degrees are \( d \). A countable graph is *locally finite* if all its vertices have a finite degree.

**Example 1.1** (Petersen graph). All vertices in the Petersen graph in Figure 1.1 have degree 3, i.e., it is 3-regular. In particular there is no isolated vertex. ◀

A convenient way of specifying a graph is the following matrix representation. Assume the graph \( G = (V,E) \) has \( n = |V| \) vertices. The *adjacency matrix* \( A \) of \( G \) is the \( n \times n \) matrix defined as

\[
A_{xy} = \begin{cases} 
1 & \text{if } \{x,y\} \in E \\
0 & \text{o.w.}
\end{cases}
\]

**Example 1.2** (Triangle). The adjacency matrix of a *triangle*, i.e. 3 vertices with all possible non-loop edges, is

\[
A = \begin{bmatrix} 
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 0 \\
\end{bmatrix}.
\]

**Paths, cycles, and trees** A *subgraph* of \( G = (V,E) \) is a graph \( G' = (V',E') \)
with $V' \subseteq V$ and $E' \subseteq E$. Implied in this definition is the fact that the edges in $E'$ are incident only to $V'$. The subgraph $G'$ is said to be induced if

$$E' = \{ \{x, y\} : x, y \in V', \{x, y\} \in E\},$$

i.e., it contains all edges of $G$ between the vertices of $V'$. In that case the notation $G' := G[V']$ is used. A subgraph is said to be spanning if $V' = V$. A subgraph containing all possible non-loop edges between its vertices is called a complete subgraph or clique.

**Example 1.3** (Petersen graph (continued)). The Petersen graph contains no triangle, induced or not.

A path in $G$ is a sequence of (not necessarily distinct) vertices $x_0 \sim x_1 \sim \cdots \sim x_k$. The number of edges, $k$, is called the length of the path. If the endvertices $x_0, x_k$ coincide, i.e., $x_0 = x_k$, we call the path a cycle. If the vertices are all distinct (except possibly for the endvertices), we say that the path (or cycle) is self-avoiding. A self-avoiding path or cycle can be seen as a (not necessarily induced) subgraph of $G$. We write $u \leftrightarrow v$ if there is a path between $u$ and $v$. Clearly $\leftrightarrow$ is an equivalence relation. The equivalence classes are called connected components. The length of the shortest self-avoiding path connecting two distinct vertices $u, v$ is called the graph distance between $u$ and $v$, denoted by $\rho(u, v)$. It can be checked that the graph distance is a metric.

A graph is connected if any two vertices are linked by a path, i.e., if $u \leftrightarrow v$ for all $u, v \in V$. Or put differently, if there is only one connected component.

**Example 1.4** (Petersen graph (continued)). The Petersen graph is connected.

A forest is a graph with no self-avoiding cycle. A tree is a connected forest. Vertices of degree 1 are called leaves. A spanning tree of $G$ is a subgraph which is a tree and is also spanning. It is straightforward to check that a tree on $n$ vertices has $n - 1$ edges. (Proof: induction.) A tree is called rooted if it has a single distinguished vertex called the root.

**Some standard graphs** Here are a few examples of finite graphs.

- **Complete graph** $K_n$: This graph is made of $n$ vertices with all non-loop edges.

- **Cycle** $C_n$: The vertex set is $\{0, 1, \ldots, n - 1\}$ and two vertices $i \neq j$ are adjacent if and only if $|i - j| = 1$ or $n - 1$.  

\[1\]
- **Hypercube** $\mathbb{Z}_2^n$: The vertex set is $\{0,1\}^n$ and two vertices $x \neq y$ are adjacent if and only if $\|x - y\|_1 = 1$.

- **Rooted $b$-ary tree** $\hat{T}_\ell^b$: This graph is a tree with $\ell$ levels. The unique vertex on level 0 is called the root. For $j = 1, \ldots, \ell - 1$, level $j$ has $b^j$ vertices, each of which has exactly one neighbor on level $j - 1$ (its parent) and $b$ neighbors on level $j + 1$ (its children). The $b^\ell$ vertices on level $\ell$ are leaves.

Here are a few examples of infinite graphs.

- **Infinite $d$-regular tree** $\mathbb{T}_d$: This is an infinite tree where each vertex has exactly $d$ neighbors.

- **Lattice** $\mathbb{L}^d$: The vertex set is $\mathbb{Z}^d$ and two vertices $x \neq y$ are adjacent if and only if $\|x - y\|_1 = 1$.

We also consider a few important classes of graphs.

A **bipartite graph** $G = (L, R, E)$ is a graph whose vertex set is composed of the union of two sets $L \cup R$ and whose edge set $E$ is a subset of $\{(\ell, r) : \ell \in L, r \in R\}$. That is, there are no edges between vertices in $L$ and likewise for $R$.

**Example 1.5** (Some bipartite graphs). The cycle $C_{2n}$ is a bipartite graph. So is the complete bipartite graph $K_{n,m}$ with vertex set $\{\ell_1, \ldots, \ell_n\} \cup \{r_1, \ldots, r_m\}$ and edge set $\{(\ell_i, r_j) : i \in [n], j \in [m]\}$.

In a bipartite graph $G = (L, R, E)$, a **perfect matching** is a collection of edges $M \subseteq E$ such that each vertex in $L \cup R$ is incident to exactly one edge $M$.

An **automorphism** of a graph $G = (V, E)$ is a bijection $\phi$ of $V$ to itself that preserves the edges, i.e., such that $\{x, y\} \in E$ if and only if $\{\phi(x), \phi(y)\} \in E$. A graph $G = (V, E)$ is **vertex-transitive** if for any $u, v \in V$ there is an automorphism mapping $u$ to $v$.

**Example 1.6** (Petersen graph (continued)). For any $\ell \in \mathbb{Z}$, a $(2\pi \ell/5)$-rotation of the planar representation of the Petersen graph in Figure 1.1 corresponds to an automorphism.

**Example 1.7** (Trees). The graph $\mathbb{T}_d$ is vertex-transitive. The graph $\hat{T}_\ell^b$ has many automorphisms, but it is not vertex-transitive.

**Flows** Let $G = (V, E)$ be a connected graph with two distinguished, distinct vertex sets, a **source-set** (or **source** for short) $A \subseteq V$ and a **sink-set** (or **sink** for source, sink, flow, network)

*Usually in graph theory a path is called a “walk” but that term has a different meaning in probability theory.*
short) $Z$. Let $c : E \to \mathbb{R}_+$ be a capacity function. A flow on the network $(G, c)$ from source $A$ to sink $Z$ is a function $f : V \times V \to \mathbb{R}$ such that:

F1 (Antisymmetry) $f(x, y) = -f(y, x), \forall x, y \in V$.

F2 (Capacity constraint) $|f(x, y)| \leq c(e), \forall e = \{x, y\} \in E$, and $f(x, y) = 0$ otherwise.

F3 (Flow-conservation constraint)
$$
\sum_{y : y \sim x} f(x, y) = 0, \quad \forall x \in V \setminus (A \cup Z).
$$

For $U, W \subseteq V$ and $F \subseteq E$, let $f(U, W) := \sum_{u \in U, w \in W} f(u, w)$ and $c(F) := \sum_{e \in F} c(e)$. The strength of $f$ is $|f| := f(A, A^c)$. Let $F \subseteq E$. We call $F$ a cutset separating $A$ and $Z$ if all paths connecting $A$ and $Z$ include an edge in $F$. Let $A_F$ be the set of vertices not separated from $A$ by $F$, and similarly for $Z_F$.

**Lemma 1.8 (Max flow ≤ min cut).** For any cutset $F$ separating $A$ from $Z$, $|f| \leq c(F)$.

**Proof.** Note that
$$
|f(A, A^c)| = (F3)\ f(A, A^c) + \sum_{u \in A_F \setminus A} f(u, V) \quad (F1)\ f(A_F, A_F^c) \quad (F2)\ \leq c(F). \quad (1.1)
$$

Remarkably, this bound is tight.

**Theorem 1.9 (Max-flow min-cut theorem).** Let $(G, c)$ be a finite connected network with source $A$ and sink $Z$. Then the following holds
$$
\max\{|f| : \text{flow } f \text{ on } (G, c)\} = \min\{c(F) : \text{cutset } F \text{ separating } A \text{ and } Z\}.
$$

**Proof.** Note that, by compactness, the supremum on the l.h.s. is achieved. Let $f$ be an optimal flow. The idea of the proof is to find a matching cutset. An augmentable path is a self-avoiding path $x_0 \sim \cdots \sim x_k$ with $x_0 \in A$, $x_i \notin A \cup Z$ for all $i \neq 0$ or $k$, and $f(x_{i-1}, x_i) < c(\{x_{i-1}, x_i\})$ for all $i$. By maximality of $f$ there cannot be such a path with $x_k \in Z$, otherwise we could push more flow through that path and increase the strength of $f$—a contradiction. Let $B \subseteq V \setminus (A \cup Z)$ be the set of all final vertices in some augmentable path and let $F$ be the edge set between $B$ and $B^c$. Note that, by definition of $B$, all vertices in $B$ can be reached from $A$ without crossing $F$ and that $f(x, y) = c(e)$ for all $e = \{x, y\} \in F$ with $x \in B$ and $y \in B^c$. Furthermore $F$ is a cutset separating $A$ from $Z$; trivially $A \subseteq B$; $Z \subseteq B^c$ by maximality of $f$; and any path from $A$ to $Z$ must exit $B$ and enter $B^c$ through an edge in $F$. Thus we have equality in (1.1) with $A_F = B$. ☑
Colorings and independent sets  A coloring of a graph \( G = (V, E) \) is an assignment of colors to each vertex in \( G \). In a coloring, two vertices may share the same color. A coloring is proper if for every edge \( e \) in \( G \) the endvertices of \( e \) have distinct colors. The smallest number of colors in a proper coloring of a graph \( G \) is called the chromatic number \( \chi(G) \) of \( G \).

An independent set of \( G \) is a subset of vertices \( W \subseteq V \) such that all pairs of vertices in \( W \) are non-adjacent.

Directed graphs  A directed graph (or digraph for short) is a pair \( G = (V, E) \) where \( V \) is a set of vertices (or nodes or sites) and \( E \subseteq V^2 \) is a set of directed edges. A directed path is a sequence of vertices \( x_0, \ldots, x_k \) with \( (x_{i-1}, x_i) \in E \) for all \( i = 1, \ldots, k \). We write \( u \rightarrow v \) if there is such a path with \( x_0 = u \) and \( x_k = v \). We say that \( u, v \in V \) communicate, denoted by \( u \leftrightarrow v \), if \( u \rightarrow v \) and \( v \rightarrow u \). The \( \leftrightarrow \) relation is clearly an equivalence relation. The equivalence classes of \( \leftrightarrow \) are called the (strongly) connected components of \( G \).

1.1.2 Review of Markov chain theory

Construction  Formally, a Markov chain is a stochastic process satisfying the Markov property: given the present, the future is independent of the past. On a countable state space, such a process is characterized by its initial distribution and transition matrix.

For our purposes, it will suffice to “define” a Markov chain through a particular construction. Let \( V \) be a finite or countable space. Recall that a stochastic matrix on \( V \) is a nonnegative matrix \( P = (P(i,j))_{i,j \in V} \) satisfying
\[
\sum_{j \in V} P(i,j) = 1, \quad \forall i \in V.
\]

Let \( \mu \) be a probability measure on \( V \) and let \( P \) be a stochastic matrix on \( V \). One way to construct a Markov chain \((X_t)\) on \( V \) with transition matrix \( P \) and initial distribution \( \mu \) is the following:

- Let \( X_0 \sim \mu \) and let \((Y(i,n))_{i \in V, n \geq 1}\) be a mutually independent array of random variables with \( Y(i,n) \sim P(i, \cdot) \).

- Set inductively \( X_n := Y(X_{n-1}, n), n \geq 1 \).

So in particular:
\[
\mathbb{P}[X_0 = x_0, \ldots, X_t = x_t] = \mu(x_0)P(x_0, x_1) \cdots P(x_{t-1}, x_t).
\]

We use the notation \( \mathbb{P}_x, \mathbb{E}_x \) for the probability distribution and expectation under the chain started at \( x \). Similarly for \( \mathbb{P}_\mu, \mathbb{E}_\mu \) where \( \mu \) is a probability measure.
Example 1.10 (Simple random walk on a graph). Let \( G = (V, E) \) be a finite or infinite, locally finite graph. Simple random walk on \( G \) is the Markov chain on \( V \), started at an arbitrary vertex, which at each time picks a uniformly chosen neighbor of the current state.

Markov property Let \((X_t)\) be a Markov chain with transition matrix \( P \) and initial distribution \( \mu \). Let \( \mathcal{F}_t = \sigma(X_0, \ldots, X_t) \). As mentioned above, a fundamental property of Markov chains known as the Markov property is that, given the present, the future is independent of the past. In its simplest form, that can be interpreted as \( \mathbb{P}[X_{t+1} = y \mid \mathcal{F}_t] = \mathbb{P}_{X_t}[X_{t+1} = y] = P(X_t, y) \). More generally (see e.g. [Dur10, Theorem 6.3.1]):

**Theorem 1.11** (Markov property). Let \( f : V^\infty \to \mathbb{R} \) be bounded, measurable and let \( F(x) := \mathbb{E}_x[f((X_t)_{t \geq 0})] \), then

\[
\mathbb{E}[f((X_{s+t})_{t \geq 0}) \mid \mathcal{F}_s] = F(X_s) \quad \text{a.s.}
\]

**Remark 1.12.** We will come back to the strong Markov property in Chapter 3.

Let \((X_t)\) be a Markov chain with transition matrix \( P \). We define \( P^t(x, y) := \mathbb{P}_x[X_t = y] \). An important consequence of the Markov property is the following.

**Theorem 1.13** (Chapman-Kolmogorov).

\[
P^t(x, z) = \sum_{y \in V} P^s(x, y)P^{t-s}(y, z), \quad s \in \{0, 1, \ldots, t\}.
\]

**Proof:** This follows from the Markov property. Note that \( \mathbb{P}_x[X_t = z \mid \mathcal{F}_s] = F(X_s) \) with \( F(y) := \mathbb{P}_y[X_{t-s} = z] \) and take \( \mathbb{E}_x \) on each side. \( \blacksquare \)

Example 1.14 (Simple random walk on \( \mathbb{Z} \)). Let \((X_t)\) be simple random walk on \( \mathbb{Z} \) where \( i \sim j \) if \( |i - j| = 1 \). Then \( P^1(0, x) = 1/2 \) if \( |x| = 1 \). And \( P^2(0, x) = 1/4 \) if \( |x| = 2 \) and \( P^2(0, 0) = 1/2 \). \( \blacksquare \)

If we write \( \mu_s \) for the law of \( X_s \) as a row vector, then

\[
\mu_s = \mu_0 P^s
\]

where here \( P^s \) is the matrix product of \( P \) by itself \( s \) times.
Some more definitions  The transition graph of a chain is the directed graph on $V$ whose edges are the transitions with nonzero probabilities. A chain is irreducible if $V$ is the unique connected component of its transition graph, i.e., if all pairs of states communicate.

Example 1.15 (Simple random walk on a graph (continued)). Simple random walk on $G$ is irreducible if and only if $G$ is connected. ◀

A chain is said to be aperiodic if for all $x \in V$

$$\gcd \{ t : P^t(x, x) > 0 \} = 1.$$  

Example 1.16 (Lazy walk on a graph). A lazy, simple random walk on $G$ is a Markov chain such that, at each time, it stays put with probability 1/2 or chooses a uniformly random neighbor of the current state otherwise. Such a walk is aperiodic. ◀

Let $(X_t)$ be a Markov chain with transition matrix $P$. A stationary measure $\pi$ is a measure such that

$$\sum_{x \in V} \pi(x)P(x, y) = \pi(y), \quad \forall y \in V,$$

or in matrix form $\pi = \pi P$. We say that $\pi$ is a stationary distribution if in addition $\pi$ is a probability measure.

Example 1.17 (Simple random walk on $\mathbb{L}^d$). The measure $\pi \equiv 1$ is stationary for simple random walk on $\mathbb{L}^d$. ◀

Finite, irreducible always have a unique stationary distribution (see e.g. [LPW06, Corollary 1.17]).

Theorem 1.18 (Existence and uniqueness: finite case). If $P$ is irreducible and has a finite state space, then it has a unique stationary distribution.

A transition matrix $P$ is reversible w.r.t. a measure $\eta$ if

$$\eta(x)P(x, y) = \eta(y)P(y, x)$$

for all $x, y \in V$. By summing over $y$, such a measure is necessarily stationary. By induction, if $(X_t)$ is reversible w.r.t. a stationary distribution $\pi$

$$\mathbb{P}_\pi[X_0 = x_0, \ldots, X_t = x_t] = \pi(x_0)P(x_0, x_1) \cdots P(x_{t-1}, x_t) = \pi(x_t)$$

which explains the name.
Example 1.19 (Simple random walk on $\mathbb{L}^d$ (continued)). The measure $\eta \equiv 1$ is reversible for simple random walk on $\mathbb{L}^d$.

Example 1.20 (Simple random walk on a graph (continued)). Let $(X_t)$ be simple random walk on a connected graph $G = (V,E)$. Then $(X_t)$ is reversible w.r.t. $\eta(v) := \delta(v)$. Indeed, for all $\{u,v\} \in E$,

$$\delta(u)P(u,v) = \delta(u)\frac{1}{\delta(u)} = 1 = \delta(v)\frac{1}{\delta(v)} = \delta(v)P(v,u).$$

Example 1.21 (Metropolis chain). The Metropolis algorithm modifies a given irreducible, symmetric (i.e., the transition matrix is symmetric) chain $Q$ to produce a new chain $P$ with the same transition graph and a prescribed positive stationary distribution $\pi$. The definition is of the new chain is:

$$P(x,y) := \begin{cases} Q(x,y)\left[\frac{\pi(y)}{\pi(x)} \wedge 1\right], & \text{if } x \neq y, \\ 1 - \sum_{z \neq x} Q(x,z)\left[\frac{\pi(z)}{\pi(x)} \wedge 1\right], & \text{o.w.} \end{cases}$$

Note that $P(x,z) \leq Q(x,z)$ for all $z \neq x$ so

$$\sum_{z \neq x} P(x,z) \leq 1,$$

and hence $P$ is well-defined as a transition matrix. We claim further that $P$ is reversible w.r.t. $\pi$. Suppose $x \neq y$ and $\pi(x) \geq \pi(y)$. Then, by the definition of $P$, we have

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

$$= Q(x,y)\pi(y)$$

$$= Q(y,x)\pi(y)$$

$$= P(y,x)\pi(y),$$

where we used the symmetry of $Q$.

Convergence and mixing time A key property of Markov chains is that, under suitable assumptions, they converge to a stationary regime (see e.g. [Dur10, Theorem 6.6.4]).
**Theorem 1.22** (Convergence to stationarity). Suppose $P$ is irreducible, aperiodic and has stationary distribution $\pi$. Then, for all $x, y$,

$$P^t(x, y) \to \pi(y),$$

as $t \to +\infty$.

We will be interested in quantifying the speed of this convergence. For probability measures $\mu, \nu$ on $V$, let their total variation distance be

$$\|\mu - \nu\|_{TV} := \sup_{A \subseteq V} |\mu(A) - \nu(A)|.$$  

**Lemma 1.23** (Alternative definition of total variation distance). It holds that

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in V} |\mu(x) - \nu(x)|.$$  

**Proof.** Let $B := \{x : \mu(x) \geq \nu(x)\}$. Then, for any $A \subseteq V$,

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B),$$

and similarly $\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c)$. The two bounds are equal so $|\mu(A) - \nu(A)| \leq \mu(B) - \nu(B)$, which is achieved at $A = B$. Also

$$\mu(B) - \nu(B) = \frac{1}{2} [\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)] = \frac{1}{2} \sum_{x \in V} |\mu(x) - \nu(x)|.$$

To quantify convergence to stationarity, we define

$$d(t) := \sup_{x \in V} \|P^t(x, \cdot) - \pi(\cdot)\|_{TV}.$$  

**Lemma 1.24.** $d(t)$ is non-increasing in $t$.

**Proof.** Note that

$$d(t + 1) = \sup_{x \in V} \sup_{A \subseteq V} |P^{t+1}(x, A) - \pi(A)|$$

$$= \sup_{x \in V} \sup_{A \subseteq V} \left| \sum_z P(x, z)(P^t(z, A) - \pi(A)) \right|$$

$$\leq \sup_{x \in V} \sum_z P(x, z) \sup_{A \subseteq V} |P^t(z, A) - \pi(A)|$$

$$\leq \sup_{x \in V} \sup_{A \subseteq V} |P^t(z, A) - \pi(A)|$$

$$= d(t),$$

where on the second line we used that $P$ is a stochastic matrix.
The following concept will play a key role here.

**Definition 1.25** (Mixing time). *For a fixed $\varepsilon > 0$, the mixing time is defined as*

$$t_{\text{mix}}(\varepsilon) := \inf\{t \geq 0 : d(t) \leq \varepsilon\}.$$

1.2 Percolation

There are several types of percolation processes. We mostly focus on bond percolation.

**Definition 1.26** (Bond percolation). *Let $G = (V, E)$ be a finite or infinite graph. The bond percolation process on $G$ with density $p \in [0, 1]$, whose measure is denoted by $\mathbb{P}_p$, is defined as follows: each edge of $G$ is independently set to open with probability $p$, otherwise it is set to closed. Write $x \leftrightarrow y$ if $x, y \in V$ are connected by a path all of whose edges are open. The open cluster of $x$ is* $C_x := \{y \in V : x \leftrightarrow y\}.$

We will mostly consider bond percolation on $\mathbb{L}^d$ or $\mathbb{T}_d$.

Typical questions regarding bond percolation (and pointers to some answers) include:

- For which values of $p$ is there an infinite open cluster? How many infinite clusters are there? What is the probability that $y$ is in the open cluster of $x$?
  
  - Lattice: Sections 2.2.4, 4.3.6
  - Trees: Sections 2.3.3, 3.1.4, 6.3.1.

1.3 Random graphs

There are many different models of random graphs. We focus on the celebrated Erdős-Rényi model. We will also encounter preferential attachment graphs.

**Erdős-Rényi graphs** The Erdős-Rényi random graph is defined as follows.

**Definition 1.27** (Erdős-Rényi graphs). *Let $V = [n]$ and $p \in [0, 1]$. The Erdős-Rényi graph $G = (V, E)$ on $n$ vertices with density $p$ is defined as follows: for each pair $x \neq y$ in $V$, the edge $\{x, y\}$ is in $E$ with probability $p$ independently of all other edges. We write $G \sim \mathbb{G}_{n,p}$ and we denote the corresponding measure by $\mathbb{P}_{n,p}$.*
Typical questions regarding Erdős-Rényi graphs (and pointers to some answers) include:

- What is the probability of observing a particular subgraph, e.g., a triangle?
  - Sections 2.3.2, 4.3.5.
- Is $G$ connected? If not, how large are the components?
  - Sections 2.3.2, 6.3.3.
- How are the degrees distributed?
  - Section 4.2.2.
- What is the typical chromatic number?
  - Section 3.2.3.

Preferential attachment graphs

In the definition of preferential attachment graphs, we consider multigraphs, i.e., we allow parallel edges.

**Definition 1.28** (Preferential attachment graph). The preferential attachment process produces a sequence of graphs $(G_t)_{t \geq 1}$ as follows. We start at time 1 with two vertices, denoted $v_0$ and $v_1$, connected by an edge. At time $t$, we add vertex $v_t$ with a single edge connecting it to an old vertex, which is picked proportionally to its degree. We write $(G_t)_{t \geq 1} \sim \text{PA}_1$. For $m > 1$, we run the previous process for time $nm$ and then collapse vertices $km, km-1, \ldots, (k-1)m + 1$ to make vertex $k$.

Typical questions regarding preferential attachment graphs (and pointers to some answers) include:

- How are the degrees distributed?
  - Section 3.2.4.

1.4 Markov random fields

We start with Gibbs random fields.

**Definition 1.29** (Gibbs random field). Let $S$ be a finite set and let $G = (V, E)$ be a finite graph. Denote by $K$ the set of all cliques of $G$. A positive probability measure $\mu$ on $\mathcal{X} := S^V$ is called a Gibbs random field if there exist clique potentials $\phi_K$:
$S^K \to \mathbb{R}$, $K \in \mathcal{K}$, such that

$$
\mu(x) = \frac{1}{Z} \exp \left( \sum_{K \in \mathcal{K}} \phi_K(x_K) \right),
$$

where $x_K$ is $x$ restricted to the vertices of $K$ and $Z$ is a normalizing constant.

The following example introduces the primary Gibbs random field we will consider.

**Example 1.30 (Ising model).** For $\beta > 0$, the (ferromagnetic) Ising model with inverse temperature $\beta$ is the Gibbs random field with $S := \{-1, +1\}$, $\phi_{\{i,j\}}(\sigma_{\{i,j\}}) = \beta \sigma_i \sigma_j$ and $\phi_K \equiv 0$ if $|K| \neq 2$. The function $H(\sigma) := -\sum_{\{i,j\} \in E} \sigma_i \sigma_j$ is known as the Hamiltonian. The normalizing constant $Z := Z(\beta)$ is called the partition function. The states $(\sigma_i)_{i \in V}$ are referred to as spins.

Typical questions regarding Gibbs random field (and pointers to some answers) include:

- How fast is correlation decaying?

**Proof of Hammersley-Clifford** To be written. See [Gri10a, Theorem 7.12].

### 1.5 Random walks on graphs

Recall the following definition.

**Definition 1.31 (Simple random walk on a graph).** Let $G = (V, E)$ be a finite or countable, locally finite graph. Simple random walk on $G$ is the Markov chain on $V$, started at an arbitrary vertex, which at each time picks a uniformly chosen neighbor of the current state.

We generalize the definition slightly by adding weights to the edges.

**Definition 1.32 (Random walk on a network).** Let $G = (V, E)$ be a finite or countable, locally finite graph. Let $c : E \to \mathbb{R}_+$ be a positive edge weight function on $G$. We call $\mathcal{N} = (G, c)$ a network. We assume that for all $u \in V$

$$
c_u := \sum_{v : e = \{u, v\} \in E} c(e) < +\infty.
$$

13
Random walk on network $N$ is the Markov chain on $V$, started at an arbitrary vertex, which at each time picks a neighbor of the current state proportionally to the weight of the corresponding edge, i.e., the transition matrix is given by

$$P(u, v) = \begin{cases} 
c(u,v) / c_u & \text{if } \{u, v\} \in E \\
0 & \text{o.w.}
\end{cases}$$

Any countable, reversible Markov chain can be seen as a random walk on a network (not necessarily locally finite) by setting $c(e) := \pi(x)P(x, y) = \pi(y)P(y, x)$ for all $e = \{x, y\} \in E$.

Typical questions regarding random walks on networks (and pointers to some answers) include:

- **How long does it take to visit all vertices at least once or a particular subset of vertices for the first time?**
  - This question is related to the notions of commute time and hitting time, which are defined in Section 3.1.1.
  - General results: Sections 3.1.2, 3.3.1, 3.3.2, 3.3.3.
  - Special chains: birth-and-death chains (Section 3.3.2); trees (Section 3.3.2); complete graph (Section 3.3.3).
  - Application to branching processes: Section 6.1.

- **How often does the walk return to its starting point?**
  - This question is related to the notion of recurrence, which is defined in Section 3.1.1.
  - General results: Sections 3.3.2, 3.3.3.
  - Special chains: lattice (Section 3.3.3); trees (Section 3.3.3).

- **How fast does it approach stationarity?**
  - This question is related to the notion of mixing time, which is defined in Section 1.1.2.
  - General results: lower bounds (Sections 2.4.7); upper bounds (Sections 4.4.1, 4.4.3).
  - Special chains: cycle (Sections 4.4.2); trees (Sections 4.4.2); hypercube (Sections 4.4.2).
**Glauber dynamics**  We will also consider a particular class of Markov chains over Markov random fields, the Glauber dynamics of the Ising model.

**Definition 1.33** (Glauber dynamics for the Ising model). Let \( \mu_\beta \) be the Ising model with inverse temperature \( \beta > 0 \) on a graph \( G = (V, E) \). The (single-site) Glauber dynamics is the Markov chain on \( X := \{-1, +1\}^V \) which at each time:

- selects a site \( i \in V \) uniformly at random, and
- updates the spin at \( i \) according to \( \mu_\beta \) conditioned on agreeing with the current state at all sites in \( V \setminus \{i\} \).

Specifically, for \( \gamma \in \{-1, +1\} \), \( i \in \Lambda \), and \( \sigma \in X \), let \( \sigma^i,\gamma \) be the configuration \( \sigma \) with the spin at \( i \) being set to \( \gamma \). Let \( n = |V| \) and \( S_i(\sigma) := \sum_{j \sim i} \sigma_j \). Because the Ising measure factorizes, the nonzero entries of the transition matrix are

\[
Q_\beta(\sigma, \sigma^i,\gamma) := \frac{1}{n} \cdot \frac{e^{\gamma \beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}}.
\]

**Theorem 1.34.** The Glauber dynamics is reversible w.r.t. \( \mu_\beta \).

*Proof.* This chain is clearly irreducible. For all \( \sigma \in X \) and \( i \in V \), let

\[
S_{\neq i}(\sigma) := \mathcal{H}(\sigma^{i,+}) + S_i(\sigma) = \mathcal{H}(\sigma^{i,-}) - S_i(\sigma).
\]

We have

\[
\mu_\beta(\sigma^{i,-}) Q_\beta(\sigma^{i,-}, \sigma^{i,+}) = \frac{e^{-\beta S_{\neq i}(\sigma)} e^{-\beta S_i(\sigma)}}{Z(\beta)} \cdot \frac{e^{\beta S_i(\sigma)}}{n [e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]}.
\]

\[
= \frac{1}{n Z(\beta)} \frac{e^{-\beta S_{\neq i}(\sigma)} e^{\beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}}
\]

\[
= \frac{e^{-\beta S_{\neq i}(\sigma)} e^{\beta S_i(\sigma)}}{Z(\beta)} \cdot \frac{1}{n [e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]}
\]

\[
= \mu_\beta(\sigma^{i,+}) Q_\beta(\sigma^{i,+}, \sigma^{i,-}).
\]

That concludes the proof. \( \blacksquare \)

Typical questions regarding the Glauber dynamics (and pointers to some answers) include:

- How quickly does the chain approach \( \mu_\beta \) as a function of \( \beta \)? This question is related to the notion of mixing time, as discussed above in the context of random walks on networks. The answer of course depends on the graph.

- General results: Section 4.4.4.

- Special graphs: empty graph (Sections 4.4.2).
Bibliographic remarks

For an introduction to graph theory, see for example [Die10] or [Bol98]. Markov chain theory is covered in details in [Dur10, Chapter 6]. For a more gentle introduction, see e.g. [Dur12, Chapter 1] or [Res92, Chapter 2]. See also [LPW06, Chapters 1-3].
Chapter 2

Moments and tails

Moments capture useful information about the tail of a random variable. In this chapter we recall a few inequalities quantifying this intuition. Although they are often straightforward to derive, such inequalities are surprisingly powerful. We illustrate their use on a range of applications. In particular we discuss three of the most fundamental tools in discrete probability: the first moment method, the second moment method, and the Chernoff-Cramér method.

2.1 Background

We start with a few basic definitions and standard inequalities.

2.1.1 Definitions

Moments

As a quick reminder, let $X$ be a random variable with $\mathbb{E}|X|^k < +\infty$ for some non-negative integer $k$. In that case we write $X \in L^k$. Recall that the quantities $\mathbb{E}[X^k]$ and $\mathbb{E}[(X - \mathbb{E}X)^k]$, which are well-defined when $X \in L^k$, are called respectively the $k$-th moment and $k$-th central moment of $X$. The first moment and the second central moment are the mean and variance, the square root of which is the standard deviation. A random variable is said to be centered if its mean is 0. Recall that for a non-negative random variable $X$, the $k$-th moment can be expressed as

$$
\mathbb{E}[X^k] = \int_0^{+\infty} x^k \mathbb{P}[X > x] \, dx.
$$

(2.1)
The moment-generating function of $X$ is the function

$$M_X(s) = \mathbb{E}[e^{sX}],$$

defined for all $s \in \mathbb{R}$ where it is finite, which includes at least $s = 0$. If $M_X(s)$ is defined on $(-s_0, s_0)$ for some $s_0 > 0$ then $X$ has finite moments of all orders and the following expansion holds

$$M_X(s) = \sum_{k \geq 0} \frac{s^k}{k!} \mathbb{E}[X^k], \quad |s| < s_0.$$

One more piece of notation: if $A$ is an event and $X \in L^1$, then we use the shorthand

$$\mathbb{E}[X; A] = \mathbb{E}[X 1_A].$$

**Tails**  We refer to a probability of the form $\mathbb{P}[X \geq x]$ as an upper tail (or right tail) probability. Typically $x$ is greater than the mean or median of $X$. Similarly we refer to $\mathbb{P}[X \leq x]$ as a lower tail (or left tail) probability. Our general goal in this chapter is to bound tail probabilities using moments and moment-generating functions.

Tail probabilities arise naturally in many contexts, as events of interest can often be framed in terms of a random variable being unusually large or small. Such probabilities are often hard to compute directly however. As we will see in this chapter, moments offer an effective means to control tail probabilities using moments and moment-generating functions.

2.1.2 Basic inequalities

**Markov’s inequality**  Our first bound on the tail of a random variable is Markov’s inequality. In words, for a non-negative random variable: the heavier the tail, the larger the expectation. This simple inequality is in fact a key ingredient in more sophisticated tail bounds as we will see.

**Theorem 2.1** (Markov’s inequality). Let $X$ be a non-negative random variable. Then, for all $b > 0$,

$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E}X}{b}. \quad (2.2)$$

**Proof.**

$$\mathbb{E}X \geq \mathbb{E}[X; X \geq b] \geq \mathbb{E}[b; X \geq b] = b \mathbb{P}[X \geq b].$$

18
Figure 2.1: Proof of Markov’s inequality: taking expectations of the two functions depicted above yields the inequality.

See Figure 2.1 for a proof by picture. Note that this inequality is non-trivial only when \( b > \mathbb{E}X \).

**Chebyshev’s inequality**  An application of Markov’s inequality to the random variable \(|X - \mathbb{E}X|^2\) gives a classical tail inequality featuring the second moment of a random variable.

**Theorem 2.2** (Chebyshev’s inequality). Let \( X \) be a random variable with \( \mathbb{E}X^2 < +\infty \). Then, for all \( \beta > 0 \),

\[
P( |X - \mathbb{E}X| > \beta ) \leq \frac{\text{Var}[X]}{\beta^2}.
\]

(2.3)

**Proof.** This follows immediately by applying (2.2) to \(|X - \mathbb{E}X|^2\) with \( b = \beta^2 \).

Of course this bound is non-trivial only when \( \beta \) is larger than the standard deviation.

**Example 2.3.** Let \( X \) be a Gaussian random variable with mean 0 and variance \( \sigma^2 \).

A direct computation shows that \( \mathbb{E}|X| = \sigma \sqrt{\frac{2}{\pi}} \). Hence Markov’s inequality gives

\[
\mathbb{P}( |X| \geq b ) \leq \frac{\mathbb{E}|X|}{b} = \frac{\sigma}{b} \sqrt{\frac{2}{\pi}}.
\]

while Chebyshev’s inequality gives

\[
\mathbb{P}( |X| \geq b ) \leq \left( \frac{\sigma}{b} \right)^2.
\]
Figure 2.2: Comparison of Markov’s and Chebyshev’s inequalities: the squared deviation from the mean (in green) gives a better approximation of the indicator function (in grey) close to the mean (here 0) than the absolute deviation (in orange).

For $b$ large enough, Chebyshev’s inequality produces a stronger bound. See Figure 2.2 for some insight.

Example 2.4 (Coupon collector’s problem). Let $(X_i)$ be i.i.d. uniform random variables over $[n]$. Let $T_{n,k}$ be the first time that $k$ elements of $[n]$ have been picked, i.e.,

$$T_{n,k} = \inf \{ i : |\{X_1, \ldots, X_i\}| = k \},$$

with $T_{n,0} := 0$. We prove that the time it takes to pick all elements at least once—or “collect each coupon”—has the following tail. For any $\varepsilon > 0$, we have as $n \to +\infty$:

Claim 2.5.

$$\Pr \left[ \left| T_{n,n} - n \sum_{j=1}^{n} j^{-1} \right| \geq \varepsilon n \log n \right] \to 0.$$

To prove this claim we note that the time elapsed between $T_{n,k-1}$ and $T_{n,k}$, which we denote by $\tau_{n,k} := T_{n,k} - T_{n,k-1}$, is geometric with success probability $1 - \frac{k-1}{n}$. And all $\tau_{n,k}$s are independent. So, by standard results on geometric variables.
(e.g. [Dur10, Example 1.6.5]), the expectation and variance of $T_{n,n}$ are

$$E[T_{n,n}] = \sum_{i=1}^{n} \left(1 - \frac{i - 1}{n}\right)^{-1} = n \sum_{j=1}^{n} j^{-1} \sim n \log n, \quad (2.4)$$

and

$$\text{Var}[T_{n,n}] \leq \sum_{i=1}^{n} \left(1 - \frac{i - 1}{n}\right)^{-2} = n^2 \sum_{j=1}^{n} j^{-2} \leq n^2 \sum_{j=1}^{\infty} j^{-2} = \Theta(n^2). \quad (2.5)$$

So by Chebyshev’s inequality (Theorem 2.2)

$$\mathbb{P}\{|T_{n,n} - E[T_{n,n}]| \geq \varepsilon n \log n\} \leq \frac{\text{Var}[T_{n,n}]}{(\varepsilon n \log n)^2} \leq \frac{n^2 \sum_{j=1}^{\infty} j^{-2}}{(\varepsilon n \log n)^2} \rightarrow 0,$$

by (2.4) and (2.5).

2.2 First moment method

We begin with techniques based on the first moment. First recall that the expectation of a random variable has an elementary, yet handy property: linearity. That is, if random variables $X_1, \ldots, X_n$ defined on a joint probability space have finite first moments, then

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n], \quad (2.6)$$

without any further assumption. In particular linearity holds whether or not the $X_i$s are independent.

2.2.1 The probabilistic method

A key technique of probabilistic combinatorics is the so-called probabilistic method. The idea is that one can establish the existence of an object satisfying a certain property—without having to construct one explicitly. Instead one argues that a randomly chosen object exhibits the given property with positive probability. The following “obvious” observation, sometimes referred to as the first moment principle, plays a key role in this context.
Theorem 2.6 (First moment principle). Let $X$ be a random variable with finite expectation. Then, for any $\mu \in \mathbb{R}$,

$$
\mathbb{E}X \leq \mu \implies \mathbb{P}[X \leq \mu] > 0.
$$

Proof. We argue by contradiction, assume $\mathbb{E}X \leq \mu$ and $\mathbb{P}[X \leq \mu] = 0$. Write

$$
\{X \leq \mu\} = \bigcap_{n \geq 1} \{X < \mu + 1/n\}. \quad \text{That implies by monotonicity that, for any } \epsilon \in (0, 1), \mathbb{P}[X < \mu + 1/n] < \epsilon \text{ for } n \text{ large enough. Hence}
$$

$$
\mu \geq \mathbb{E}X = \mathbb{E}[X; X < \mu + 1/n] + \mathbb{E}[X; X \geq \mu + 1/n] \\
\geq \mu \mathbb{P}[X < \mu + 1/n] + (\mu + 1/n)(1 - \mathbb{P}[X < \mu + 1/n]) \\
> \mu,
$$

a contradiction. \qed

The power of this principle is easier to appreciate on an example.

Example 2.7 (Balancing vectors). Let $v_1, \ldots, v_n$ be arbitrary unit vectors in $\mathbb{R}^n$. How small can we make the norm of the combination

$$
x_1v_1 + \cdots + x_nv_n
$$

by appropriately choosing $x_1, \ldots, x_n \in \{-1, +1\}$? We claim that it can be as small as $\sqrt{n}$, for any collection of $v_i$s. At first sight, this may appear to be a complicated geometry problem. But the proof is trivial once one thinks of choosing the $x_i$s at random. Let $X_1, \ldots, X_n$ be independent random variables uniformly distributed in $\{-1, +1\}$. Then

$$
\mathbb{E}\|X_1v_1 + \cdots + X_nv_n\|^2 = \mathbb{E}\left[\sum_{i,j} X_iX_jv_i \cdot v_j\right] \\
= \sum_{i,j} \mathbb{E}[X_iX_j v_i \cdot v_j] \quad (2.7) \\
= \sum_{i,j} v_i \cdot v_j \mathbb{E}[X_iX_j] \\
= \sum_{i} ||v_i||^2 \\
= n,
$$

22
where we used the linearity of expectation in (2.7). But note that a discrete random variable \( Z = \|X_1v_1 + \cdots + X_nv_n\|^2 \) with expectation \( \mathbb{E}Z = n \) must take a value \( \leq n \) with positive probability by the first moment principle (Theorem 2.6). In other words, there must be a choice of \( X_i \)'s such that \( Z \leq n \). That proves the claim. ▶

Here is a slightly more subtle example of the probabilistic method, where one has to modify the original random choice.

**Example 2.8 (Independent sets).** Let \( G = (V,E) \) be a \( d \)-regular graph with \( n \) vertices and \( m = \frac{nd}{2} \) edges, where \( d \geq 1 \). Our goal is to derive a lower bound on the size, \( \alpha(G) \), of the largest independent set in \( G \). Recall that an independent set is a set of vertices in a graph, no two of which are adjacent. Again, at first sight, this may seem like a rather complicated graph-theoretic problem. But an appropriate random choice gives a non-trivial bound. Specifically, we claim that \( \alpha(G) \geq \frac{n}{2d} \).

The proof proceeds in two steps:

1. We first prove the existence of a subset \( S \) of vertices with relatively few edges.
2. We remove vertices from \( S \) to obtain an independent set.

Let \( 0 < p < 1 \) to be chosen below. To form the set \( S \), pick each vertex in \( V \) independently with probability \( p \). Letting \( X \) be the number of vertices in \( S \), we have by the linearity of expectation that

\[
\mathbb{E}X = \mathbb{E}\left[ \sum_{v \in V} 1_{v \in S} \right] = np,
\]

where we used that \( \mathbb{E}[1_{v \in S}] = p \). Letting \( Y \) be the number of edges between vertices in \( S \), we have by the linearity of expectation that

\[
\mathbb{E}Y = \mathbb{E}\left[ \sum_{\{i,j\} \in E} 1_{i \in S} 1_{j \in S} \right] = \frac{nd}{2} p^2,
\]

where we also used that \( \mathbb{E}[1_{i \in S} 1_{j \in S}] = p^2 \) by independence. Hence, subtracting,

\[
\mathbb{E}[X - Y] = np - \frac{nd}{2} p^2,
\]

which, as a function of \( p \), is maximized at \( p = 1/d \) where it takes the value \( n/2d \). As a result, by the first moment principle (Theorem 2.6) there must exist a set \( S \) of vertices in \( G \) such that

\[
|S| - |\{\{i,j\} \in E : i, j \in S\}| \geq n/2d.
\] (2.8)
For each edge $e$ connecting two vertices in $S$, remove one of the end-vertices of $e$. Then, by (2.8), the resulting subset of vertices has at least $n/2d$ vertices, with no edge between them. This proves the claim.

Note that a graph $G$ made of $n/(d+1)$ cliques of size $d+1$ (with no edge between the cliques) has $\alpha(G) = n/(d+1)$, showing that our bound is tight up to a constant. This is known as a Turán graph.

Remark 2.9. The previous result can be strengthened to

$$\alpha(G) \geq \sum_{v \in V} \frac{1}{d_v + 1},$$

for a general graph $G = (V, E)$, where $d_v$ is the degree of $v$. This bound is achieved for Turán graphs. See, e.g., [AS11, The probabilistic lens: Turán’s theorem].

The previous example also illustrates the important indicator trick, i.e., writing a random variable as a sum of indicators, which is often used in combination with the linearity of expectation.

2.2.2 Union bound

Markov’s inequality (Theorem 2.1) can be interpreted as a quantitative version of the first moment principle (Theorem 2.6). In this context, it is often stated in the following special form.

**Theorem 2.10** (First moment method). *If $X$ is a non-negative, integer-valued random variable, then*

$$\Pr[X > 0] \leq \mathbb{E}X.$$  \hspace{1cm} (2.9)

**Proof.** Take $b = 1$ in Markov’s inequality (Theorem 2.1). □

Theorem 2.6 implies that, if a non-negative integer-valued random variable $X$ has expectation smaller than 1, then its value is 0 with positive probability. Theorem 2.10 adds: if $X$ has “small” expectation, then its value is 0 with “large” probability. This simple fact is typically used in the following manner: one wants to show that a certain “bad event” does not occur with probability approaching 1; the random variable $X$ then counts the number of such “bad events.” In that case, $X$ is a sum of indicators and Theorem 2.10 reduces to the standard union bound, also known as Boole’s inequality.

**Corollary 2.11.** Let $B_n = A_{n,1} \cup \cdots \cup A_{n,m_n}$, where $A_{n,1}, \ldots, A_{n,m_n}$ is a collection of events for each $n$. Then, letting

$$\mu_n := \sum_{i=1}^{m_n} \Pr[A_{n,i}],$$

24
we have
\[ P[B_n] \leq \mu_n. \]

In particular, if \( \mu_n \to 0 \) as \( n \to +\infty \), then \( P[B_n] \to 0 \).

**Proof.** This is of course a fundamental property of probability measures. (Or take \( X := X_n = \sum_{i=1}^{m_n} 1_{A_n,i} \) in Theorem 2.10.)

A useful generalization of the union bound is given in Exercise 2.1.

Applications of Theorem 2.10 are often referred to as the first moment method.

We give a few more examples.

**Example 2.12** (Random k-SAT threshold). For \( r \in \mathbb{R}_+ \), let \( \Phi_{n,r} : \{0,1\}^n \to \{0,1\} \) be a random \( k \)-CNF formula on \( n \) Boolean variables \( z_1, \ldots, z_n \) with \( rn \) clauses. (For simplicity, assume \( rn \) is an integer.) That is, \( \Phi_{n,r} \) is an AND of \( rn \) ORs, each obtained by picking independently \( k \) literals uniformly at random (with replacement). Recall that a literal is a variable \( z_i \) or its negation \( \overline{z}_i \). The formula \( \Phi_{n,r} \) is said to be satisfiable if there exists an assignment \( z \) such that \( \Phi_{n,r}(z) = 1 \). Clearly the higher the value of \( r \), the less likely it is for \( \Phi_{n,r} \) to be satisfiable. In fact it is conjectured that there exists an \( r^*_k \in \mathbb{R}_+ \) such that

\[
\lim_{n \to \infty} P[\Phi_{n,r} \text{ is satisfiable}] = \begin{cases} 0, & \text{if } r > r^*_k, \\ 1, & \text{if } r < r^*_k. \end{cases}
\]

Studying such threshold phenomena is a major theme of modern discrete probability. Using the first moment method, we give an upper bound on this conjectured threshold. Specifically, we claim that

\[ r > 2^k \ln 2 \implies \limsup_{n \to \infty} P[\Phi_{n,r} \text{ is satisfiable}] = 0. \]

How to start the proof should be obvious: let \( X_n \) be the number of satisfying assignments of \( \Phi_{n,r} \). Applying the first moment method, since

\[ P[\Phi_{n,r} \text{ is satisfiable}] = P[X_n > 0], \]

it suffices to show that \( \mathbb{E}X_n \to 0 \). To compute \( \mathbb{E}X_n \), we use the indicator trick

\[ X_n = \sum_{z \in \{0,1\}^n} 1_{\{z \text{ satisfies } \Phi_{n,r}\}}. \]

There are \( 2^n \) possible assignments, each of which satisfies \( \Phi_{n,r} \) with probability \( (1 - 2^{-k})^{rn} \). Indeed note that the \( rn \) clauses are independent and each clause
literal picked is satisfied with probability \(1/2\). Therefore, by the assumption on \(r\), for some \(\varepsilon > 0\)
\[
\mathbb{E}X_n = 2^n(1 - 2^{-k})^rn \\
< 2^n(1 - 2^{-k})^{(2k \ln2)(1+\varepsilon)n} \\
< 2^n e^{-(\ln 2)(1+\varepsilon)n} \\
= 2^{-\varepsilon n} \\
\to 0.
\]

\[\blacktriangleright\]

**Remark 2.13.** For \(k \geq 3\), it has been shown that if \(r < 2^k \ln 2 - k\)
\[
\lim_{n \to \infty} \mathbb{P}[\Phi_{n,r} \text{ is satisfiable}] = 1.
\]
See [ANP05]. For the \(k = 2\) case, the conjecture above has been established and \(r_2^* = 1\) [CR92].

### 2.2.3 Random permutations: longest increasing subsequence

In this section, we bound the expected length of a longest increasing subsequence in a random permutation. Let \(\sigma_n\) be a uniformly random permutation of \([n] := \{1, \ldots, n\}\) and let \(L_n\) be the length of a longest increasing subsequence of \(\sigma_n\).

**Claim 2.14.**
\[
\mathbb{E}L_n = \Theta(\sqrt{n}).
\]

**Proof.** We first prove that
\[
\lim_{n \to \infty} \sup_{n} \frac{\mathbb{E}L_n}{\sqrt{n}} \leq e,
\]
which implies half of the claim. Bounding the expectation of \(L_n\) is not straightforward as it is the expectation of a maximum. A natural way to proceed is to find a value \(\ell\) for which \(\mathbb{P}[L_n \geq \ell]\) is “small.” More formally, we bound the expectation as follows
\[
\mathbb{E}L_n \leq \ell \mathbb{P}[L_n < \ell] + n \mathbb{P}[L_n \geq \ell] \leq \ell + n \mathbb{P}[L_n \geq \ell], \tag{2.10}
\]
for an \(\ell\) chosen below. To bound the probability on the r.h.s., we appeal to the first moment method by letting \(X_n\) be the number of increasing subsequences of length \(\ell\). We also use the indicator trick, i.e., we think of \(X_n\) as a sum of indicators over subsequences (not necessarily increasing) of length \(\ell\). There are \(\binom{n}{\ell}\) such subsequences, each of which is increasing with probability \(1/\ell!\). Note that these
subsequences are not independent. Nevertheless, by the linearity of expectation and the first moment method,

\[ P[L_n \geq \ell] = P[X_n > 0] \leq \mathbb{E} X_n = \frac{1}{\ell!} \left( \frac{n}{\ell} \right)^\ell \leq \frac{n^\ell}{(\ell!)^2} \leq \frac{n^\ell}{e^{2(\ell/e)^2}} \leq \left( \frac{e\sqrt{n}}{\ell} \right)^{2\ell}, \]

where we used a standard bound on factorials recalled in Section A.1.1. Note that, in order for this bound to go to 0, we need \( \ell > e\sqrt{n} \). The first claim follows by taking \( \ell = (1 + \delta)e\sqrt{n} \) in (2.10), for an arbitrarily small \( \delta > 0 \).

For the other half of the claim, we show that \( \mathbb{E} L_n \sqrt{n} \geq 1 \).

This part does not rely on the first moment method (and may be skipped). We seek a lower bound on the expected length of a longest increasing subsequence. The proof uses the following two ideas. First observe that there is a natural symmetry between the lengths of the longest increasing and decreasing subsequences—they are identically distributed. Moreover if a permutation has a “short” longest increasing subsequence, then intuitively it must have a “long” decreasing subsequence, and vice versa. Combining these two observations gives a lower bound on the expectation of \( L_n \). Formally, let \( D_n \) be the length of a longest decreasing subsequence. By symmetry and the arithmetic mean-geometric mean inequality, note that

\[ \mathbb{E} L_n = \mathbb{E} \left[ \frac{L_n + D_n}{2} \right] \geq \mathbb{E} \sqrt{L_n D_n}. \]

We show that \( L_n D_n \geq n \), which proves the claim. We use a clever combinatorial argument. Let \( L_n^{(k)} \) be the length of a longest increasing subsequence ending at position \( k \), and similarly for \( D_n^{(k)} \). It suffices to show that the pairs \( (L_n^{(k)}, D_n^{(k)}) \), \( 1 \leq k \leq n \) are distinct. Indeed, noting that \( L_n^{(k)} \leq L_n \) and \( D_n^{(k)} \leq D_n \), the number of pairs in \([L_n] \times [D_n]\) is at most \( L_n D_n \) which must then be at least \( n \). Let \( 1 \leq j < k \leq n \). If \( \sigma_n(k) > \sigma_n(j) \) then we see that \( L_n^{(k)} > L_n^{(j)} \) by appending \( \sigma_n(k) \) to the subsequence ending at position \( j \) achieving \( L_n^{(j)} \). The opposite holds for the decreasing case, which implies that \( (L_n^{(j)}, D_n^{(j)}) \) and \( (L_n^{(k)}, D_n^{(k)}) \) must be distinct. This combinatorial argument is known as the Erdős-Szekeres theorem. That concludes the proof of the second claim.

**Remark 2.15.** It has been shown that in fact

\[ \mathbb{E} L_n = 2\sqrt{n} + cn^{1/6} + o(n^{1/6}), \]

where \( c = -1.77... \) [BDJ99].
2.2.4 ★ Percolation on \( \mathbb{Z}^2 \): non-trivial threshold

We use the first moment method to prove the existence a non-trivial threshold in bond percolation on lattices.

Threshold in bond percolation Consider bond percolation on the two-dimensional lattice \( \mathbb{L}^2 \) with density \( p \) and let \( \mathbb{P}_p \) denote the corresponding measure. Writing \( x \leftrightarrow y \) if \( x, y \in \mathbb{L}^2 \) are connected by an open path, recall that the open cluster of \( x \) is
\[ C_x := \{ y \in \mathbb{Z}^2 : x \leftrightarrow y \}. \]
The percolation function is defined as
\[ \theta(p) := \mathbb{P}_p[|C_0| = +\infty], \]
i.e., \( \theta(p) \) is the probability that the origin is connected by open paths to infinitely many vertices. It is intuitively clear that the function \( \theta(p) \) is non-decreasing. Indeed consider the following alternative representation of the percolation process: to each edge \( e \), assign a uniform \([0, 1]\) random variable \( U_e \) and declare the edge open if \( U_e \leq p \). Using the same \( U_e \)s for densities \( p_1 < p_2 \), it follows immediately from the monotonicity of the construction that \( \theta(p_1) \leq \theta(p_2) \). (We will have much more to say about this type of “coupling” argument in Chapter 4.) Moreover note that \( \theta(0) = 0 \) and \( \theta(1) = 1 \). The critical value is defined as
\[ p_c(\mathbb{L}^2) = \sup\{ p \geq 0 : \theta(p) = 0 \}, \]
the point at which the probability that the origin is contained in an infinite open cluster becomes positive. Note that by a union bound over all vertices, when \( \theta(p) = 0 \), we have that \( \mathbb{P}_p[\exists x, |C_x| = +\infty] = 0 \). Conversely, because \( \{\exists x, |C_x| = +\infty\} \) is a tail event, by Kolmogorov’s 0-1 law (e.g. [Dur10, Theorem 2.5.1]) it holds that \( \mathbb{P}_p[\exists x, |C_x| = +\infty] = 1 \) when \( \theta(p) > 0 \).

Using the first moment method we show that the critical value is non-trivial, i.e., it is strictly between 0 and 1. This is another example of a threshold phenomenon.

Claim 2.16.
\[ p_c(\mathbb{L}^2) \in (0, 1). \]

Proof. We first show that, for any \( p < 1/3 \), \( \theta(p) = 0 \). In order to apply the first moment method, roughly speaking, we need to reduce the problem to counting the number of instances of an appropriately chosen sub-structure. The key observation is the following:
An infinite $C_0$ must contain an open self-avoiding path starting at 0 of infinite length and, as a result, of all lengths.

Hence, we let $X_n$ be the number of open self-avoiding paths of length $n$ starting at 0. Then, by monotonicity,

$$\mathbb{P}[|C_0| = +\infty] \leq \mathbb{P}({\cap_n \{ X_n > 0 \}}) = \lim_n \mathbb{P}[X_n > 0] \leq \lim sup_n \mathbb{E}[X_n], \quad (2.11)$$

where the last inequality follows from Theorem 2.10. We bound the number of self-avoiding paths by noting that they cannot backtrack. That gives 4 choices at the first step, and at most 3 choices at each subsequent step. Hence, we get the following bound

$$\mathbb{E}X_n \leq 4(3^n - 1)p^n. \quad (2.12)$$

The r.h.s. in (2.12) goes to 0 when $p < 1/3$. When combined with (2.11), that proves the first part of the claim. (Why did we use self-avoiding—rather than unconstrained—paths?)

For the other direction, we show that $\theta(p) > 0$ for $p$ close enough to 1. This time, we count “dual cycles.” This type of proof is known as a contour argument, or Peierls’ argument, and is based on the following construction. Consider the dual lattice $\hat{\mathbb{L}}^2$ whose vertices are $\mathbb{Z}^2 + (1/2, 1/2)$ and whose edges connect vertices $u, v$ with $\|u - v\|_1 = 1$. See Figure 2.3. Note that each edge in the primal lattice $\mathbb{L}^2$ has a unique corresponding edge in the dual lattice which crosses it perpendicularly. We make the same assignment, open or closed, for corresponding primal and dual edges. The following graph-theoretic lemma, whose proof is sketched below, forms the basis of contour arguments.

**Lemma 2.17 (Contour lemma).** If $|C_0| < +\infty$, then there is a closed self-avoiding cycle around the origin in the dual lattice $\hat{\mathbb{L}}^2$.

To prove that $\theta(p) > 0$ for $p$ close enough to 1, the idea is to use the first moment method with $X_n$ equal to the number of closed self-avoiding dual cycles of length $n$ surrounding the origin. We bound from above the number of dual cycles of length $n$ around the origin by the number of choices for the starting edge across the upper
Figure 2.3: Primal (black) and dual (blue) lattices.

$y$-axis and for each $n - 1$ subsequent non-backtracking choices. Namely,

$$
\mathbb{P}[|C_0| < +\infty] \leq \mathbb{P}[\exists n \geq 4, X_n > 0] \\
\leq \sum_{n \geq 4} \mathbb{P}[X_n > 0] \\
\leq \sum_{n \geq 4} \mathbb{E}X_n \\
\leq \sum_{n \geq 4} \frac{n}{2} 3^{n-1}(1-p)^n \\
= \frac{3^3(1-p)^4}{2} \sum_{m \geq 1} (m + 3)(3(1-p))^{m-1} \\
= \frac{3^3(1-p)^4}{2} \left( \frac{1}{(1-3(1-p))^2} + 3 \frac{1}{1-3(1-p)} \right),
$$

when $p > 2/3$ (where the first term on the last line comes from differentiating the geometric series). This expression can be taken smaller than 1 if we let $p \to 1$. We have shown that $\theta(p) > 0$ for $p$ close enough to 1, and that concludes the proof. (Exercise 2.2 sketches a proof that $\theta(p) > 0$ for $p > 2/3$.)

It is straightforward to extend the claim to $\mathbb{L}^d$. Exercise 2.3 asks for the details.

**Proof of the contour lemma** We conclude this section by sketching the proof of the contour lemma, which relies on topological arguments.
Proof of Lemma 2.17. Assume $|C_0| < +\infty$. Imagine identifying each vertex in $\mathbb{L}^2$ with a square of side 1 centered around it so that the sides line up with dual edges. Paint green the squares of vertices in $C_0$. Paint red the squares of vertices in $C_0^C$ which share a side with a green square. Leave the other squares white. Let $u_0$ be the highest vertex in $C_0$ along the $y$-axis and let $v_0$ be the dual vertex corresponding to the upper left corner of the square of $u_0$. Because $u_0$ is highest, it must be that the square above it is red. Walk along the dual edge $\{v_0, v_1\}$ separating the squares of $u_0$ and $u_0 + (0,1)$ from $v_0$ to $v_1$. Notice that this edge satisfies what we call the red-green property: a red square sits on your left and a green square is on your right. Proceed further by iteratively walking along an incident dual edge with the following rule. Choose an edge satisfying the red-green property, with the edges to your left, straight ahead, and to your right in decreasing order of priority (i.e., choose the highest priority edge that satisfies the red-green property). Stop when a previously visited dual vertex is reached. The claim is that this procedure constructs the desired cycle. Let $v_0, v_1, v_2, \ldots$ be the dual vertices visited. By construction $\{v_{i-1}, v_i\}$ is a dual edge for all $i$.

- (A dual cycle is produced) We first argue that this procedure cannot get stuck. Let $\{v_{i-1}, v_i\}$ be the edge just crossed and assume that it has the red-green property. If there is a green square to the left ahead, then the edge to the left, which has highest priority, has the red-green property. If the left square ahead is not green, but the right one is, then the left square must in fact be red by construction. In that case, the edge straight ahead has the red-green property. Finally, if neither square ahead is green, then the right square must in fact be red because the square behind to the right is green by assumption. That implies that the edge to the right has the red-green property. Hence we have shown that the procedure does not get stuck. Moreover, because by assumption the number of green squares is finite, this procedure must eventually terminate when a previously visited dual vertex is reached, forming a cycle.

- (The origin lies within the cycle) The inside of a cycle in the plane is well-defined by the Jordan curve theorem. So the dual cycle produced above has its adjacent green squares either on the inside (negative orientation) or on the outside (positive orientation). In the former case, the origin must lie inside the cycle as otherwise the vertices corresponding to the green squares on the inside would not be in $C_0$. So it remains to consider the latter case where, for similar reasons, the origin is outside the cycle.

Let $v_j$ be the repeated dual vertex. Assume first that $v_j \neq v_0$ and let $v_{j-1}$ and $v_{j+1}$ be the dual vertices preceding and following $v_j$ during the first visit to
Let \( v_k \) be the dual vertex preceding \( v_j \) on the second visit. After traversing \( \{v_{j-1}, v_j\} \), \( v_k \) cannot be to the left or to the right because in those cases the red-green property of the two corresponding edges are not compatible. So \( v_k \) is straight ahead and, by the priority rules, \( v_{j+1} \) must be to the left. But in that case, for the origin to lie outside the cycle and for the cycle to avoid the path \( v_0, \ldots, v_{j-1} \), we must traverse the cycle with a negative orientation, i.e., the green squares adjacent to the cycle must be on the inside, a contradiction.

So, finally, assume \( v_0 \) is the repeated vertex. If the cycle is traversed with a positive orientation and the origin is on the outside, it must be that the cycle crosses the \( y \)-axis at least once above \( u_0 + (0, 1) \), a contradiction.

Hence we have shown that the origin is inside the cycle.

That concludes the proof. ■

**Remark 2.18.** It turns out that \( p_c(L^2) = 1/2 \). We will prove \( p_c(L^2) \geq 1/2 \), known as Harris’ theorem, in Section 4.3.6. The other direction is due to Kesten [Kes80].

### 2.3 Second moment method

The first moment method gives an upper bound on the probability that a non-negative, integer-valued random variable is positive—provided its expectation is small enough. In this section we seek a lower bound on that probability. We first note that a large expectation does not suffice in general. Say \( X_n \) is \( n^2 \) with probability \( 1/n \), and 0 otherwise. Then \( \mathbb{E}X_n = n \to +\infty \), yet \( \mathbb{P}[X_n > 0] \to 0 \). That is, although the expectation diverges, the probability that \( X_n \) is positive can be arbitrarily small.

So we turn to the second moment. Intuitively the basis for the so-called second moment method is that, if the expectation of \( X_n \) is large and its variance is relatively small, then we can bound the probability that \( X_n \) is close to 0. As we will see in applications, the first and second moment methods often work hand in hand.

#### 2.3.1 Paley-Zygmund inequality

As an immediate corollary of Chebyshev’s inequality (Theorem 2.2), we get a first version of the second moment method: if the standard deviation of \( X \) is less than its expectation, then the probability that \( X \) is 0 is bounded away from 1. See Figure 2.4. Formally, let \( X \) be a non-negative, integer-valued random variable (not
identically zero). Then

\[ P[X > 0] \geq 1 - \frac{\text{Var}[X]}{(\E[X])^2}. \]  

(2.13)

Indeed, by (2.3),

\[ P[X = 0] \leq P[|X - \E[X]| \geq \E[X]] \leq \frac{\text{Var}[X]}{(\E[X])^2}. \]

The following tail inequality, a simple application of Cauchy-Schwarz, leads to an improved version of the second moment method.

**Theorem 2.19** (Paley-Zygmund inequality). *Let X be a non-negative random variable. For all 0 < \( \theta < 1 \),

\[ P[X \geq \theta \E[X]] \geq (1 - \theta)^2 \frac{(\E[X])^2}{\E[X^2]}. \]  

(2.14)

**Proof.** We have

\[ \E[X] = \E[X1_{\{X < \theta \E[X]\}}} + \E[X1_{\{X \geq \theta \E[X]\}}] \]

\[ \leq \theta \E[X] + \sqrt{\E[X^2]} P[X \geq \theta \E[X]], \]

where we used Cauchy-Schwarz. Rearranging gives the result.
As an immediate application:

**Theorem 2.20** (Second moment method). Let $X$ be a non-negative random variable (not identically zero). Then

$$
P[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}.
$$

(2.15)

**Proof.** Take $\theta \downarrow 0$ in (2.14).

Since

$$
\frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} = 1 - \frac{\text{Var}[X]}{(\mathbb{E}X)^2 + \text{Var}[X]},
$$

we see that (2.15) is stronger than (2.13). We typically apply the second moment method to a sequence of random variables $(X_n)$. The previous theorem gives a uniform lower bound on the probability that $\{X_n > 0\}$ when $\mathbb{E}[X_n^2] \leq C(\mathbb{E}[X_n])^2$ for some $C > 0$.

Just like the first moment method, the second moment method is often applied to a sum of indicators (but see Section 2.3.3 for a weighted case).

**Corollary 2.21.** Let $B_n = A_{n,1} \cup \cdots \cup A_{n,m_n}$, where $A_{n,1}, \ldots, A_{n,m_n}$ is a collection of events for each $n$. Write $i \sim j$ if $i \neq j$ and $A_{n,i}$ and $A_{n,j}$ are not independent. Then, letting

$$
\mu_n := \sum_{i=1}^{m_n} \mathbb{P}[A_{n,i}], \quad \gamma_n := \sum_{i \sim j} \mathbb{P}[A_{n,i} \cap A_{n,j}],
$$

where the second sum is over ordered pairs, we have $\lim_n \mathbb{P}[B_n] > 0$ whenever $\mu_n \to +\infty$ and $\gamma_n \leq C\mu_n^2$ for some $C > 0$. If moreover $\gamma_n = o(\mu_n^2)$ then $\lim_n \mathbb{P}[B_n] = 1$.

**Proof.** Take $X := X_n = \sum_{i=1}^{m_n} 1_{A_{n,i}}$ in the second moment method (Theorem 2.20). Note that

$$
\text{Var}[X_n] = \sum_i \text{Var}[1_{A_{n,i}}] + \sum_{i \neq j} \text{Cov}[1_{A_{n,i}}, 1_{A_{n,j}}],
$$

where

$$
\text{Var}[1_{A_{n,i}}] = \mathbb{E}[(1_{A_{n,i}})^2] - (\mathbb{E}[1_{A_{n,i}}])^2 \leq \mathbb{P}[A_{n,i}],
$$

and, if $A_{n,i}$ and $A_{n,j}$ are independent,

$$
\text{Cov}[1_{A_{n,i}}, 1_{A_{n,j}}] = 0,
$$

34
whereas, if $i \sim_{n} j$,

$$
\text{Cov}[1_{A_{n,i}}, 1_{A_{n,j}}] = E[1_{A_{n,i}} 1_{A_{n,j}}] - E[1_{A_{n,i}}]E[1_{A_{n,j}}] \leq P[A_{n,i} \cap A_{n,j}].
$$

Hence

$$
\frac{\text{Var}[X_n]}{(E[X_n])^2} \leq \frac{\mu_n + \gamma_n}{\mu_n^2} = \frac{1}{\mu_n} + \frac{\gamma_n}{\mu_n^2}.
$$

Noting

$$
\frac{(E[X_n])^2}{E[X_n^2]} = \frac{(E[X_n])^2}{(E[X_n])^2 + \text{Var}[X_n]} = \frac{1}{1 + \text{Var}[X_n]/(E[X_n])^2},
$$

and applying Theorem 2.20 gives the result.

We give applications of Theorem 2.20 and Corollary 2.21 in Sections 2.3.2 and 2.3.3.

2.3.2 \textbf{Erdős-Rényi: subgraph containment and connectivity}

We have seen examples of threshold phenomena in constraint satisfaction problems and percolation. Such thresholds are also common in random graphs. We consider here the Erdős-Rényi random graph. Formally, a threshold function for a graph property $P$ is a function $r(n)$ such that

$$
\lim_n P_{n,p_n}[G_n \text{ has property } P] = \begin{cases} 
0, & \text{if } p_n \ll r(n) \\
1, & \text{if } p_n \gg r(n),
\end{cases}
$$

where, under $P_{n,p_n}$, $G_n \sim G_{n,p_n}$ is an Erdős-Rényi graph with $n$ vertices and density $p_n$. In this section, we illustrate this type of phenomenon on two properties: the containment of small subgraphs and connectivity.

Small subgraphs: the containment problem

We first consider the clique number, then we turn to the more general subgraph containment problem.

**Cliquess** Let $\omega(G)$ be the clique number of a graph $G$, i.e., the size of its largest clique.

**Claim 2.22.** The property $\omega(G) \geq 4$ has threshold function $n^{-2/3}$. 

35
Proof. Let $X_n$ be the number of 4-cliques in the Erdös-Rényi graph $G_n \sim G_{n,p}$. Then, noting that there are $\binom{4}{2} = 6$ edges in a 4-clique,

$$\mathbb{E}_{n,p} [X_n] = \binom{n}{4} p_n^6 = \Theta(n^4 p_n^6),$$

which goes to 0 when $p_n \ll n^{-2/3}$. Hence the first moment method (Theorem 2.10) gives one direction.

For the other direction, we apply the second moment method for sums of indicators, Corollary 2.21. For an enumeration $S_1, \ldots, S_m$ of the 4-tuples of vertices in $G_n$, let $A_1, \ldots, A_m$ be the events that the corresponding 4-cliques are present. By the calculation above we have $\mu_m = \Theta(n^4 p_n^6)$ which goes to $+\infty$ when $p_n \gg n^{-2/3}$. Also $\mu_m^2 = \Theta(n^8 p_n^{12})$ so it suffices to show that $\gamma_m = o(n^8 p_n^{12})$.

Note that two 4-cliques with disjoint edge sets (but possibly sharing one vertex) are independent. Suppose $S_i$ and $S_j$ share 3 vertices. Then

$$\mathbb{P}_{n,p} [A_i | A_j] = p_n^3,$$

as the event $A_j$ implies that all edges between three of the vertices in $S_i$ are present, and there are 3 edges between the remaining vertex and the rest of $S_i$. Similarly if $|S_i \cap S_j| = 2$, $\mathbb{P}_{n,p} [A_i | A_j] = p_n^5$. Putting these together we get

$$\gamma_m = \sum_{i \sim j} \mathbb{P}_{n,p} [A_j] \mathbb{P}_{n,p} [A_i | A_j]$$

$$= \binom{n}{4} p_n^6 \left[ \binom{4}{3} (n-4) p_n^3 + \binom{4}{2} \binom{n-4}{2} p_n^5 \right]$$

$$= O(n^5 p_n^9) + O(n^6 p_n^{11})$$

$$= O \left( \frac{n^8 p_n^{12}}{n^3 p_n^3} \right) + O \left( \frac{n^8 p_n^{12}}{n^2 p_n} \right)$$

$$= o(n^8 p_n^{12})$$

$$= o(\mu_m^2),$$

where we used that $p_n \gg n^{-2/3}$ (so that for example $n^3 p_n^3 \gg 1$). Corollary 2.21 gives the result.

Roughly speaking, the first and second moments suffice to pinpoint the threshold in this case because the indicators in $X_n$ are “mostly” pairwise independent and, as a result, the sum is concentrated around its mean.
General subgraphs  The methods of Claim 2.22 can be applied to more general subgraphs. However the situation is somewhat more complicated than it is for cliques. For a graph $H_0$, let $v_{H_0}$ and $e_{H_0}$ be the number of vertices and edges of $H_0$ respectively. Let $X_n$ be the number of copies of $H_0$ in $G_n \sim \mathbb{G}_{n,p_n}$. By the first moment method,

$$\mathbb{P}[X_n > 0] \leq \mathbb{E}[X_n] = \Theta(n^{v_{H_0}} p_n^{e_{H_0}}) \rightarrow 0,$$

when $p_n \ll n^{-v_{H_0}/e_{H_0}}$. (The constant factor, which does not play a role in the asymptotics, accounts in particular for the number of automorphisms of $H_0$.) From the proof of Claim 2.22, one might guess that the threshold function is $n^{-v_{H_0}/e_{H_0}}$.

That is not the case in general. To see what can go wrong, consider the graph of Figure 2.5 whose edge density is $e_{H_0}/v_{H_0} = 6/5$. When $p_n \gg n^{-5/6}$, the expected number of copies of $H_0$ tends to $+\infty$. But observe that the subgraph $H$ of $H_0$ has the higher density $5/4$ and, hence, when $n^{-5/6} \ll p_n \ll n^{-4/5}$ the expected number of copies of $H$ tends to 0. By the first moment method, the probability that a copy of $H_0$—and therefore $H$—is present in that regime is asymptotically negligible despite its diverging expectation. This leads to the following definition

$$r_{H_0} := \max \left\{ \frac{e_H}{v_H} : H \subseteq H_0, v_H > 0 \right\}.$$ 

Assume $H_0$ has at least one edge.

Claim 2.23. “Having a copy of $H_0$” has threshold $n^{-1/r_{H_0}}$.

Proof: We proceed as in Claim 2.22. Let $H_0^*$ be a subgraph of $H_0$ achieving $r_{H_0}$. When $p_n \ll n^{-1/r_{H_0}}$, the probability that a copy of $H_0^*$ is in $G_n$ tends to 0 by the argument above. Therefore the same conclusion holds for $H_0$ itself.

Assume $p_n \gg n^{-1/r_{H_0}}$. Let $I_1, \ldots, I_m$ be an enumeration of the copies of $H_0$ in a complete graph on the vertices of $G_n$. Let $A_i$ be the event that $I_i \subseteq G_n$. Using the notation of Corollary 2.21,

$$\mu_m = \Theta(n^{v_{H_0}} p_n^{e_{H_0}}) = \Omega(\Phi_{H_0}),$$

where

$$\Phi_{H_0} := \min_{H \subseteq H_0, e_H > 0} n^{v_H} p_n^{e_H}.$$ 

Note that $\Phi_{H_0} \rightarrow +\infty$. The events $A_i$ and $A_j$ are independent if $I_i$ and $I_j$ share no edge. Otherwise we write $i \sim j$. Note that there are $\Theta(n^{v_{H_0}} n^{2(v_{H_0} - v_H)})$ pairs...
Figure 2.5: Graph $H_0$ and subgraph $H$.

$I_i, I_j$ whose intersection is isomorphic to $H$. The probability that both elements of such a pair are present in $G_n$ is $\Theta(p_n^{2e_{H_0}} p_n^{e_{H_0} - e_H})$. Hence

\[
\gamma_m = \sum_{i \sim j} P[A_i \cap A_j] = \sum_{H \subseteq H_0, e_H > 0} \Theta\left(n^{2v_{H_0} - v_H} p_n^{2e_{H_0} - e_H}\right) = \frac{\Theta(\mu_m^2)}{\Theta(\Phi_{H_0})} = o(\mu_m^2).
\]

The result follows from Corollary 2.21.

Going back to the example of Figure 2.5, the proof above confirms that when $n^{-5/6} \ll p_n \ll n^{-4/5}$ the second moment method fails for $H_0$ since $\Phi_{H_0} \to 0$. In that regime, although there is in expectation a large number of copies of $H_0$, those copies are highly correlated as they are produced from a (vanishingly) small number of copies of $H$—explaining the failure of the second moment method.
Connectivity threshold

In this section, we use the second moment method to show that the threshold function for connectivity in the Erdős-Rényi random graph is $\frac{\log n}{n}$. In fact we prove this result by deriving the threshold function for the presence of isolated vertices. The connection between the two is obvious in one direction. Isolated vertices imply a disconnected graph. What is less obvious is that it also works the other way in the following sense: the two thresholds actually coincide.

Isolated vertices

We begin with isolated vertices.

**Claim 2.24.** “Not having an isolated vertex” has threshold function $\frac{\log n}{n}$.

**Proof.** Let $X_n$ be the number of isolated vertices in the Erdős-Rényi graph $G_{n,p} \sim G_{n,p}$. Using $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$,

$$\mathbb{E}_{n,p_n}[X_n] = n(1 - p_n)^{n-1} \leq e^{\log n - (n-1)p_n} \to 0,$$

when $p_n \gg \frac{\log n}{n}$. So the first moment method gives one direction: $\mathbb{P}_{n,p_n}[X_n > 0] \to 0$.

For the other direction, we use the second moment method. Let $A_j$ be the event that vertex $j$ is isolated. By the computation above, using $1 - x \geq e^{-x-x^2}$ for $x \in [0, 1/2]$ (Exercise: Check),

$$\mu_n = \sum_i \mathbb{P}_{n,p_n}[A_i] = n(1 - p_n)^{n-1} \geq e^{\log n - np_n - np_n^2}, \quad (2.16)$$

which goes to $+\infty$ when $p_n \ll \frac{\log n}{n}$. Note that $i \sim j$ for all $i \neq j$ and

$$\mathbb{P}_{n,p_n}[A_i \cap A_j] = (1 - p_n)^{2(n-2)+1},$$

so that

$$\gamma_n = \sum_{i \neq j} \mathbb{P}_{n,p_n}[A_i \cap A_j] = n(n - 1)(1 - p_n)^{2n-3}.$$

Because $\gamma_n$ is not a $o(\mu_n^2)$, we cannot apply Corollary 2.21. Instead we use Theorem 2.20 directly. We have

$$\frac{\mathbb{E}_{n,p_n}[X_n^2]}{(\mathbb{E}_{n,p_n}[X_n])^2} = \frac{\mu_n + \gamma_n}{\mu_n^2} \leq \frac{n(1 - p_n)^{n-1} + n^2(1 - p_n)^{2n-3}}{n^2(1 - p_n)^{2n-2}} \leq \frac{1}{n(1 - p_n)^{n-1}} + \frac{1}{1 - p_n}, \quad (2.17)$$

which is $1 + o(1)$ when $p_n \ll \frac{\log n}{n}$.
Connectivity We use Claim 2.24 to study the threshold for connectivity.

Claim 2.25. Connectivity has threshold function \( \frac{\log n}{n} \).

Proof. We start with the easy direction. If \( p_n \ll \frac{\log n}{n} \), Claim 2.24 implies that the graph has isolated vertices, and therefore is disconnected, with probability going to 1 as \( n \to +\infty \).

Assume that \( p_n \gg \frac{\log n}{n} \). Let \( D_n \) be the event that \( G_n \) is disconnected. To bound \( P_{n,p_n}[D_n] \), for \( k \in \{1, \ldots, n/2\} \) we let \( Y_k \) be the number of subsets of \( k \) vertices that are disconnected from all other vertices in the graph. Then, by the first moment method,

\[
P_{n,p_n}[D_n] \leq P_{n,p_n}\left[ \sum_{k=1}^{n/2} Y_k > 0 \right] \leq \sum_{k=1}^{n/2} E_{n,p_n}[Y_k].
\]

The expectation of \( Y_k \) is straightforward to estimate. Using that \( k \leq n/2 \) and \( \binom{n}{k} \leq n^k \),

\[
E_{n,p_n}[Y_k] = \binom{n}{k} (1 - p_n)^{k(n-k)} \leq \left(n(1 - p_n)^{n/2}\right)^k.
\]

The expression in parentheses is \( o(1) \) when \( p_n \gg \frac{\log n}{n} \). Summing over \( k \),

\[
P_{n,p_n}[D_n] \leq \sum_{k=1}^{+\infty} \left(n(1 - p_n)^{n/2}\right)^k = O(n(1 - p_n)^{n/2}) = o(1),
\]

where we used that the geometric series (started at \( k = 1 \)) is dominated asymptotically by its first term.

A closer look We have shown that connectivity and the absence of isolated vertices have the same threshold function. In fact, in a sense, isolated vertices are the “last obstacle” to connectivity. A slight modification of the proof above leads to the following more precise result. For \( k \in \{1, \ldots, n/2\} \), let \( Z_k \) be the number of connected components of size \( k \) in \( G_n \). In particular, \( Z_1 \) is the number of isolated vertices. We consider the critical window \( p_n = \frac{c_n}{n} \) where \( c_n := \log n + s \) for some fixed \( s \in \mathbb{R} \). We show that, in that regime, the graph is composed of a large connected component together with some isolated vertices.

Claim 2.26.

\[
P_{n,p_n}[Z_1 > 0] \geq \frac{1}{1 + e^s} + o(1) \quad \text{and} \quad P_{n,p_n}\left[ \sum_{k=2}^{n/2} Z_k > 0 \right] = o(1).
\]
Proof. We first consider isolated vertices. From (2.16) and (2.17),
\[
P_{n,p} \left[ Z_1 > 0 \right] \geq \left( e^{- \log n + np_n + np_n^2 + \frac{1}{1 - p_n}} \right)^{-1} = \frac{1}{1 + e^s + o(1)},
\]
as \( n \to +\infty \).

To bound the number of components of size \( k > 1 \), we note first the random variable \( Y_k \) used in the previous claim (which imposes no condition on the edges between the vertices) is too loose to provide a suitable bound. Instead, to bound the probability that a set of \( k \) vertices forms a connected component, we observe that a connected component is characterized by two properties: it is disconnected from the rest of the graph; and it contains a spanning tree. Formally, for \( k = 2, \ldots, n/2 \), we let \( Z'_k \) be the the number of (not necessarily induced) maximal trees of size \( k \) or, put differently, the number of spanning trees of connected components of size \( k \). Then, by the first moment method, the probability that a connected component of size \( > 1 \) is present in \( G_n \) is bounded by
\[
P_{n,p} \left[ \sum_{k=2}^{n/2} Z_k > 0 \right] \leq \sum_{k=2}^{n/2} P_{n,p} \left[ Z'_k > 0 \right] \leq \sum_{k=2}^{n/2} E_{n,p_n} \left[ Z'_k \right].
\] (2.18)

To estimate the expectation of \( Z'_k \), we use Cayley’s theorem (e.g. [LP, Corollary 4.5]) which implies that there are \( k^{k-2} \) labelled trees on a set of \( k \) vertices. Recall further that a tree on \( k \) vertices has \( k-1 \) edges. Hence,
\[
E_{n,p_n} \left[ Z'_k \right] = \left( \begin{array}{c} n \\ k \end{array} \right) k^{k-2} \left( \frac{p_n}{n} \right)^{k-1} (1 - p_n)^{k(n-k)},
\]
where (a) is the number of trees of size \( k \), (b) is the probability that such a tree is present in the graph, and (c) is the probability that this tree is disconnected from every other vertex in the graph. Using that \( k! \geq (k/e)^k \) (see Section A.1.1),
\[
E_{n,p_n} \left[ Z'_k \right] \leq \frac{n^k}{k!} k^{k-2} \left( \frac{p_n}{n} \right)^{k-1} (1 - p_n)^{k(n-k)} \leq n \left( e c_n e^{-\left(1 - \frac{k}{2}\right)c_n} \right)^k.
\]
For \( k \leq n/2 \), the expression in parentheses is \( o(1) \). In fact, for \( k \geq 2 \), \( E_{n,p_n} \left[ Z'_k \right] = o(1) \). Furthermore, summing over \( k > 2 \),
\[
\sum_{k=3}^{n/2} E_{n,p_n} \left[ Z'_k \right] \leq \sum_{k=3}^{+\infty} n \left( e c_n e^{-\frac{1}{2}c_n} \right)^k = O(n^{-1/2} \log^3 n) = o(1).
\]
Plugging this back into (2.18) concludes the proof.

The limit of \( P_{n,p_n} \left[ Z_1 > 0 \right] \) can be computed explicitly using the method of moments. See Exercise 2.14.
2.3.3 Percolation on trees: critical value and branching number

Consider bond percolation on the infinite $d$-regular tree $T_d$. Root the tree arbitrarily at a vertex $0$ and let $C_0$ be the open connected component containing $0$. In this section we illustrate the use of the first and second moment methods on the identification of the critical value

$$p_c(T_d) = \sup\{p \in [0, 1] : \theta(p) = 0\},$$

where recall that the percolation function is $\theta(p) = \mathbb{P}_p[|C_0| = +\infty]$. We then consider general trees, introduce the branching number, and present a weighted version of the second moment method.

**Regular tree** Our main result for $T_d$ is the following.

**Theorem 2.27.**

$$p_c(T_d) = \frac{1}{d - 1}.$$

**Proof.** Let $\partial_n$ be the $n$-th level of $T_d$, i.e., the set of vertices at graph distance $n$ from $0$. Let $X_n$ be the number of vertices in $\partial_n \cap C_0$. In order for the component of the root to be infinite, there must be at least one vertex on the $n$-th level connected to the root by an open path. By the first moment method,

$$\theta(p) \leq \mathbb{P}_p[X_n > 0] \leq \mathbb{E}_p X_n = d(d - 1)^{n-1} p^n \to 0,$$

when $p < \frac{1}{d - 1}$. Here we used that there is a unique path between $0$ and any vertex in the tree to deduce that $\mathbb{P}_p[x \in C_0] = p^n$ for $x \in \partial_n$. Equation (2.19) implies $p_c(T_d) \geq \frac{1}{d - 1}$.

The second moment method gives a lower bound on $\mathbb{P}_p[X_n > 0]$. To simplify the notation, it is convenient to introduce the “branching ratio” $b := d - 1$. We say that $x$ is a descendant of $z$ if the path between $0$ and $x$ goes through $z$. Each $z \neq 0$ has $d - 1$ descendant subtrees, i.e., subtrees of $T_d$ rooted at $z$ made of all descendants of $z$. Let $x \wedge y$ be the most recent common ancestor of $x$ and $y$, i.e., the furthest vertex from $0$ that lies on both the path from $0$ to $x$ and the path from $0$ to $y$; see Figure 2.6. Letting $\mu := \mathbb{E}_p[X_n]$
Figure 2.6: Most recent common ancestor of $x$ and $y$.

\[
\mathbb{E}_p[X_n^2] = \sum_{x,y \in \partial_n} \mathbb{P}_p[x, y \in C_0]
\]

\[
= \sum_{x \in \partial_n} \mathbb{P}_p[x \in C_0] + \sum_{m=0}^{n-1} \sum_{x,y \in \partial_n} 1_{\{x \land y \in \partial_m\}} p^m p^{2(n-m)}
\]

\[
= \mu_n + (b + 1)b^{n-1} \sum_{m=0}^{n-1} (b-1)b^{n-m-1}p^{2n-m}
\]

\[
\leq \mu_n + (b + 1)(b - 1)b^{2n-2}p^{2n} \sum_{m=0}^{+\infty} (bp)^{-m}
\]

\[
= \mu_n + \mu_n^2 \cdot \frac{b - 1}{b + 1} \cdot \frac{1}{1 - (bp)^{-1}},
\]

where, on the third line, we used that all vertices on the $n$-th level are equivalent and that, for a fixed $x$, the set $\{y : x \land y \in \partial_m\}$ is composed of those vertices in $\partial_n$ that are descendants of $x \land y$ but not in the descendant subtree of $x \land y$ containing...
When $p > \frac{1}{d-1} = \frac{1}{b}$, dividing by $(\mathbb{E}_p X_n)^2 = \mu_n^2 \to +\infty$, we get

$$\frac{\mathbb{E}_p[X_n^2]}{(\mathbb{E}_p X_n)^2} \leq \frac{1}{\mu_n} + \frac{b-1}{b+1} \cdot \frac{1}{-1}(bp)^{-1}$$

$\leq 1 + \frac{b-1}{b+1} \cdot \frac{1}{-1}(bp)^{-1}$

$=: C_{b,p}$

for $n$ large enough. By the second moment method and monotonicity,

$$\theta(p) = \mathbb{P}_p[\forall n, X_n > 0] = \lim_{n} \mathbb{P}_p[X_n > 0] \geq C_{b,p}^{-1} > 0,$$

which concludes the proof. (Note that the version of the second moment method in Equation (2.13) does not work here. Subtract 1 in (2.20) and take $p$ close to $1/b$.)

The argument above relies crucially on the fact that, in a tree, any two vertices are connected by a unique path. For instance, estimating $\mathbb{P}_p[x \in C_0]$ is much harder on a lattice. Note furthermore that, intuitively, the reason why the first moment captures the critical threshold exactly in this case is that bond percolation on $T_d$ is a branching process, where $X_n$ represents the population size at generation $n$. The qualitative behavior of a branching process is governed by its expectation: when the mean number of offsprings $bp$ exceeds 1, the process grows exponentially on average and “explodes” with positive probability. We will come back to this point of view in Chapter 6 where branching processes are used to give a more refined analysis of bond percolation on $T_d$.

**General trees** Let $\mathcal{T}$ be a locally finite tree (i.e., all its degrees are finite) with root 0. For an edge $e$, let $v_e$ be the endvertex of $e$ furthest from the root. We denote by $|e|$ the graph distance between 0 and $v_e$. A cutset separating 0 and $+\infty$ is a set of edges $\Pi$ such that all infinite self-avoiding paths starting at 0 go through $\Pi$. For a cutset $\Pi$, we let $\Pi_v := \{v_e : e \in \Pi\}$. Repeating the argument in (2.19), for any cutset $\Pi$, by the first moment method

$$\theta(p) \leq \mathbb{P}_p[C_0 \cap \Pi_v \neq 0] \leq \sum_{u \in \Pi} \mathbb{P}_p[u \in C_0] = \sum_{e \in \Pi} p^{|e|}.$$  

This bound naturally leads to the following definition.

**Definition 2.28 (Branching number).** The branching number of $\mathcal{T}$ is given by

$$\text{br}(\mathcal{T}) = \sup \left\{ \lambda \geq 1 : \inf_{\text{cutset} \Pi} \sum_{e \in \Pi} \lambda^{-|e|} > 0 \right\}.$$  

44
Remark 2.29. For locally finite trees, it suffices to consider finite cutsets. See the proof of [LP, Theorem 3.1].

Equation (2.21) implies that $p_c(\mathcal{T}) \geq \frac{1}{\text{br}(\mathcal{T})}$. Remarkably, this bound is tight. The proof is based on a weighted second moment method.

Theorem 2.30. For any rooted, locally finite tree $\mathcal{T}$,

$$p_c(\mathcal{T}) = \frac{1}{\text{br}(\mathcal{T})}.$$ 

Proof. As we argued above, $p_c(\mathcal{T}) \geq \frac{1}{\text{br}(\mathcal{T})}$ follows from (2.21) and (2.22).

Let $p > \frac{1}{\text{br}(\mathcal{T})}$, $p^{-1} < \lambda < \text{br}(\mathcal{T})$, and $\varepsilon > 0$ such that

$$\sum_{e \in \Pi} \lambda^{-|e|} \geq \varepsilon$$

for all cutsets $\Pi$. The existence of such an $\varepsilon$ is guaranteed by the definition of the branching number. As in the proof of Theorem 2.27, we use that $\theta(p)$ is the limit as $n \to +\infty$ of the probability that $C_0$ reaches the $n$-th level. However, this time, we use a weighted count on the $n$-th level. Let $\mathcal{T}_n$ be the first $n$ levels of $\mathcal{T}$ and, as before, let $\partial_n$ be the vertices on the $n$-th level. For a probability measure $\nu$ on $\partial_n$, we define the weighted count

$$X_n = \sum_{z \in \partial_n} \nu(z) \mathbf{1}_{\{z \in C_0\}} \frac{1}{\mathbb{E}_p[z \in C_0]}.$$ 

The purpose of the denominator is normalization:

$$\mathbb{E}_p X_n = \sum_{z \in \partial_n} \nu(z) = 1.$$ 

(Note that multiplying all weights by a constant does not affect the event $\{X_n > 0\}$ and that the constant cancels out in the ratio of $(\mathbb{E}_p X_n)^2$ and $\mathbb{E}_p X_n^2$ in the second moment method.) Because of (2.23), a natural choice of $\nu$ follows from the max-flow min-cut theorem which guarantees the existence of a unit flow $\phi$ from 0 to $\partial_n$ satisfying the capacity constraint $\phi(e) \leq \varepsilon^{-1} \lambda^{-|e|}$, for all edges $e$ in $\mathcal{T}_n$. Define $\nu(z)$ to be the flow entering $z \in \mathcal{T}_n$ under $\phi$. In particular, because $\phi$ is a unit flow, $\nu$ restricted to $\partial_n$ defines a probability measure. It remains to bound the second
moment of $X_n$ under this choice. We have

\[
E_p X_n^2 = \sum_{x,y \in \partial_n} \nu(x) \nu(y) \frac{\mathbb{P}_p[x, y \in C_0]}{\mathbb{P}_p[x \in C_0] \mathbb{P}_p[y \in C_0]}
\]

\[
= \sum_{m=0}^{n} \sum_{x,y \in \partial_n} 1_{\{x \wedge y \in \partial_m\}} \nu(x) \nu(y) \frac{p^m p^{2(n-m)}}{p^{2n}}
\]

\[
= \sum_{m=0}^{n} p^{-m} \sum_{z \in \partial_m} \left( \sum_{x,y \in \partial_n} 1_{\{x \wedge y = z\}} \nu(x) \nu(y) \right).
\]

In the expression in parentheses, for each $x$ the sum over $y$ is at most $\nu(x) \nu(z)$ by the definition of a flow. Hence

\[
E_p X_n^2 \leq \sum_{m=0}^{n} p^{-m} \sum_{z \in \partial_m} (\nu(z))^2
\]

\[
\leq \sum_{m=0}^{n} p^{-m} \sum_{z \in \partial_m} (\varepsilon^{-1} \lambda^{-m}) \nu(z)
\]

\[
\leq \varepsilon^{-1} \sum_{m=0}^{+\infty} (p \lambda)^{-m}
\]

\[
= \frac{\varepsilon^{-1}}{1 - (p \lambda)^{-1}} =: C_{\varepsilon, \lambda, p} < +\infty,
\]

where the second line follows from the capacity constraint, and we used $p \lambda > 1$ on the last line. From the second moment method (recall that $E_p X_n = 1$), it follows that

\[
\theta(p) \geq C_{\varepsilon, \lambda, p}^{-1} > 0,
\]

and $p_c(\mathcal{F}) \leq \frac{1}{\text{br}(\mathcal{F})}$. 

### 2.4 Chernoff-Cramér method

Chebyshev’s inequality (Theorem 2.2) gives a bound on the concentration around the mean of a square integrable random variable that is, in general, best possible. Indeed take $X$ to be $\mu + b \sigma$ or $\mu - b \sigma$ with probability $(2b^2)^{-1}$ respectively, and $\mu$ otherwise. Then $E X = \mu$, $\text{Var} X = \sigma^2$, and for $\beta = b \sigma$,

\[
\mathbb{P}[|X - E X| \geq \beta] = \mathbb{P}[|X - E X| = \beta] = \frac{1}{b^2} = \frac{\text{Var} X}{\beta^2}.
\]
However, in many cases, much stronger bounds can be derived. For instance, if $X \sim \mathcal{N}(0, 1)$, by the following lemma

$$
\mathbb{P}[|X - \mathbb{E}X| \geq \beta] \sim \sqrt{\frac{2}{\pi}} \beta^{-1} \exp(-\beta^2/2) \ll \frac{1}{\beta^2},
$$

(2.24)
as $\beta \to +\infty$. Indeed:

**Lemma 2.31.** For $x > 0$,

$$(x^{-1} - x^{-3}) e^{-x^2/2} \leq \int_x^{+\infty} e^{-y^2/2} dy \leq x^{-1} e^{-x^2/2}.$$

**Proof.** By the change of variable $y = x + z$ and using $e^{-z^2/2} \leq 1$

$$
\int_x^{+\infty} e^{-y^2/2} dy \leq e^{-x^2/2} \int_0^{+\infty} e^{-xz} dz = e^{-x^2/2} x^{-1}.
$$

For the other direction, by differentiation

$$
\int_x^{+\infty} (1 - 3y^{-4}) e^{-y^2/2} dy = (x^{-1} - x^{-3}) e^{-x^2/2}.
$$

In this section we discuss the Chernoff-Cramér method, which produces exponential tail inequalities, provided the moment-generating function is finite in a neighborhood of 0.

### 2.4.1 Tail bounds via the moment-generating function

Under a finite variance, squaring within Markov’s inequality (Theorem 2.1) produces Chebyshev’s inequality (Theorem 2.2). This “boosting” can be pushed further when stronger integrability conditions hold. We refer to (2.25) in the next lemma as the Chernoff-Cramér bound.

**Lemma 2.32** (Chernoff-Cramér bound). Assume $X$ is a centered random variable such that $M_X(s) < +\infty$ for $s \in (-s_0, s_0)$ for some $s_0 > 0$. For any $\beta > 0$ and $s > 0$,

$$
\mathbb{P}[X \geq \beta] \leq \exp \left[-\{s\beta - \Psi_X(s)\}\right],
$$

(2.25)

where

$$
\Psi_X(s) = \log M_X(s),
$$
is the cumulant-generating function of $X$. 

47
Proof. Exponentiating within Markov’s inequality gives
\[ P[X \geq \beta] = P[e^{sX} \geq e^{s\beta}] \leq \frac{M_X(s)}{e^{s\beta}} = \exp\left[-\{s\beta - \Psi_X(s)\}\right]. \]

Returning to the Gaussian case, let \( X \sim N(0, \nu) \) where \( \nu > 0 \) is the variance and note that
\[
M_X(s) = \int_{-\infty}^{+\infty} e^{sx} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{x^2}{2\nu}} \, dx = \int_{-\infty}^{+\infty} e^{\frac{s^2\nu}{2}} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{(x-s\nu)^2}{2\nu}} \, dx = \exp\left(\frac{s^2\nu}{2}\right).
\]
By straightforward calculus, the optimal choice of \( s \) in (2.25) gives the exponent
\[
\sup_{s > 0} (s\beta - \frac{s^2\nu}{2}) = \frac{\beta^2}{2\nu},
\]
achieved at \( s_\beta = \beta/\nu \). For \( \beta > 0 \), this leads to the bound
\[
P[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right), \tag{2.27}
\]
which is much sharper than Chebyshev’s inequality for large \( \beta \)—compare to (2.24).

As another toy example, we consider simple random walk on \( \mathbb{Z} \).

Lemma 2.33 (Chernoff bound for simple random walk on \( \mathbb{Z} \)). Let \( Z_1, \ldots, Z_n \) be independent \( \{-1, 1\}-valued \) random variables with \( P[Z_i = 1] = P[Z_i = -1] = 1/2 \). Let \( S_n = \sum_{i \leq n} Z_i \). Then, for any \( \beta > 0 \),
\[
P[S_n \geq \beta] \leq e^{-\beta^2/2n}. \tag{2.28}
\]
Proof. The moment-generating function of \( Z_1 \) can be bounded as follows
\[
M_{Z_1}(s) = \frac{e^s + e^{-s}}{2} = \sum_{j \geq 0} \frac{s^{2j}}{(2j)!} \leq \sum_{j \geq 0} \frac{(s^2/2)^j}{j!} = e^{s^2/2}. \tag{2.29}
\]
Taking \( s = \beta/n \) in the Chernoff-Cramér bound (2.25), we get
\[
P[S_n \geq \beta] \leq \exp(-s\beta + n\Psi_{Z_1}(s)) \leq \exp(-s\beta + ns^2/2) = e^{-\beta^2/2n},
\]
which concludes the proof.
Observe the similarity between (2.28) and the Gaussian bound (2.27), if one takes \( \nu \) to be the variance of \( S_n \)

\[
\nu := \text{Var}[S_n] = n \text{Var}[Z_1] = n \mathbb{E}[Z_1^2] = n,
\]

where we used that \( Z_1 \) is centered. The central limit theorem says that simple random walk is well approximated by a Gaussian in the bulk of the distribution; the bound above extends the approximation in the large deviation regime. The bounding technique used in the proof of Lemma 2.33 will be substantially extended in Section 2.4.3.

**Example 2.34 (Set balancing).** Let \( v_1, \ldots, v_m \) be arbitrary non-zero vectors in \( \{0, 1\}^n \). Think of \( v_i = (v_{i,1}, \ldots, v_{i,n}) \) as representing subsets of \( [n] = \{1, \ldots, n\} \): \( v_{i,j} = 1 \) indicates that \( j \) is in subset \( i \). Suppose we want to partition \( [n] \) into two groups such that the subsets corresponding to the \( v_i \)'s are as balanced as possible, that is, are as close as possible to having the same number of elements from each group. More formally, we seek a vector \( x = (x_1, \ldots, x_n) \in \{-1, +1\}^n \) such that \( B^* = \max_{i=1, \ldots, m} |x \cdot v_i| \) is as small as possible. Once again we select each \( x_i \) independently, uniformly at random in \( \{-1, +1\} \). Fix \( \varepsilon > 0 \). We claim that

\[
\mathbb{P} \left[ B^* \geq \sqrt{2n(\log m + \log(2\varepsilon^{-1}))} \right] \leq \varepsilon. \tag{2.30}
\]

Indeed, by (2.28) (considering only the non-zero entries of \( v_i \)),

\[
\mathbb{P} \left[ |x \cdot v_i| \geq \sqrt{2n(\log m + \log(2\varepsilon^{-1}))} \right]
\leq 2 \exp \left( -\frac{2n(\log m + \log(2\varepsilon^{-1}))}{2\|v_i\|_1} \right)
\leq \frac{\varepsilon}{m},
\]

where we used \( \|v_i\|_1 \leq n \). Taking a union bound over the \( m \) vectors gives the result. In (2.30), the \( \sqrt{n} \) term on the right-hand side of the inequality is to be expected since it is the standard deviation of \( |x \cdot v_i| \) in the worst case. The power of the exponential tail bound (2.28) appears in the logarithmic terms, which would have been much larger if one had used Chebyshev’s inequality (Theorem 2.2) instead. (Try it!)

The Chernoff-Cramér bound is particularly useful for sums of independent random variables as the moment-generating function then factorizes. Let \( \Psi_X^*(s) = \sup_{s \in \mathbb{R}_+} (s \beta - \Psi_X(s)) \), be the Fenchel-Legendre dual of the cumulant-generating function of \( X \).
Theorem 2.35 (Chernoff-Cramér method). Let \( S_n = \sum_{i \leq n} X_i \), where the \( X_i \)'s are i.i.d. centered random variables. Assume \( M_{X_1}(s) < +\infty \) on \( s \in (-s_0, s_0) \) for some \( s_0 > 0 \). For any \( \beta > 0 \),
\[
\mathbb{P}[S_n \geq \beta] \leq \exp \left( -n \Psi^*_X\left( \frac{\beta}{n} \right) \right). \tag{2.31}
\]
In particular, in the large deviations regime, i.e., when \( \beta = bn \) for some \( b > 0 \), we have
\[
- \limsup_n \frac{1}{n} \log \mathbb{P}[S_n \geq bn] \geq \Psi^*_X\left( b \right). \tag{2.32}
\]

Proof. Observe that
\[
\Psi^*_X\left( \beta \right) = \sup_{s > 0} (s\beta - n\Psi_X(s)) = \sup_{s > 0} n \left( s \left( \frac{\beta}{n} \right) - \Psi_X(s) \right) = n\Psi^*_X\left( \frac{\beta}{n} \right),
\]
and optimize over \( s \) in (2.25).

2.4.2 Some examples

We use the Chernoff-Cramér method to derive a few standard bounds.

Poisson variables Let \( Z \sim \text{Poi}(\lambda) \) be Poisson with mean \( \lambda \) and recall that, letting \( X = Z - \lambda \),
\[
\Psi_X(s) = \log \left( \sum_{\ell \geq 0} e^{-\lambda} \frac{\lambda^\ell}{\ell!} e^{s(\ell - \lambda)} \right) = \log \left( e^{-(1-s)\lambda} \sum_{\ell \geq 0} \frac{(e^s\lambda)^\ell}{\ell!} \right) = \log (e^{-(1-s)\lambda} e^{e^s\lambda}) = \lambda(e^s - s - 1),
\]
so that straightforward calculus gives for \( \beta > 0 \)
\[
\Psi^*_X(\beta) = \sup_{s > 0} (s\beta - \lambda(e^s - s - 1)) = \lambda \left[ (1 + \frac{\beta}{\lambda}) \log \left( 1 + \frac{\beta}{\lambda} \right) - \frac{\beta}{\lambda} \right] =: \lambda h\left( \frac{\beta}{\lambda} \right),
\]

50
achieved at $s_\beta = \log \left(1 + \frac{\beta}{\lambda}\right)$, where $h$ is defined as the expression in square brackets. Plugging $\Psi_X(\beta)$ into Theorem 2.35 leads for $\beta > 0$ to the bound

$$\mathbb{P}[Z \geq \lambda + \beta] \leq \exp \left(-\lambda h \left(\frac{\beta}{\lambda}\right)\right).$$  \hspace{1cm} (2.33)

A similar calculation for $-(Z - \lambda)$ gives for $\beta < 0$

$$\mathbb{P}[Z \leq \lambda + \beta] \leq \exp \left(-\lambda h \left(\frac{\beta}{\lambda}\right)\right).$$  \hspace{1cm} (2.34)

If $S_n$ is a sum of $n$ i.i.d. Po(\lambda) variables, then by (2.32) for $a > \lambda$

$$-\lim_{n} \frac{1}{n} \log \mathbb{P}[S_n \geq an] \geq \lambda h \left(\frac{a - \lambda}{\lambda}\right) = a \log \left(\frac{a}{\lambda}\right) - a + \lambda =: I^\text{Poi}_\lambda(a),$$  \hspace{1cm} (2.35)

and similarly for $a < \lambda$

$$-\lim_{n} \frac{1}{n} \log \mathbb{P}[S_n \leq an] \geq I^\text{Poi}_\lambda(a).$$  \hspace{1cm} (2.36)

In fact, these bounds follow immediately from (2.33) and (2.34) by noting that $S_n \sim \text{Poi}(n\lambda)$.

**Binomial variables and Chernoff bounds** Let $Z$ be a binomial random variable with parameters $n$ and $p$. Recall that $Z$ is a sum of i.i.d. indicators $Y_1, \ldots, Y_n$ and, letting $X_i = Y_i - p$ and $S_n = Z - np$, 

$$\Psi_{X_1}(s) = \log (pe^s + (1 - p)) - ps.$$  \hspace{1cm}

For $b \in (0, 1 - p)$, letting $a = b + p$, direct calculation gives

$$\Psi_{X_1}(b) = \sup_{s>0} (sb - (\log [pe^s + (1 - p)] - ps))$$

$$= (1 - a) \log \frac{1 - a}{1 - p} + a \log \frac{a}{p} =: D(a\|p),$$  \hspace{1cm} (2.37)

achieved at $s_b = \log \frac{(1-p)a}{p(1-a)}$. The function $D(a\|p)$ in (2.37) is the so-called **Kullback-Leibler divergence** or **relative entropy** of Bernoulli variables with parameters $a$ and $p$ respectively. By (2.31) for $\beta > 0$

$$\mathbb{P}[Z \geq np + \beta] \leq \exp \left(-n D(p + \beta/n\|p)\right).$$

Applying the same argument to $Z' = n - Z$ gives a bound in the other direction.
Remark 2.36. In the large deviations regime, it can be shown that the previous bound is tight in the sense that
\[-\frac{1}{n} \log P[Z \geq np + bn] \to D(p + b\|p) =: I^{\text{Bin}}_{n,p}(b),\]
as \(n \to +\infty\). The theory of large deviations, which deals with asymptotics of probabilities of rare events, provides general results along those lines. See e.g. [Dur10, Section 2.6]. Upper bounds will be enough for our purposes. A similar remark applies to (??) and (2.35).

The following related bounds, proved in Exercise 2.6, are often useful.

Theorem 2.37 (Chernoff bounds for Poisson trials). Let \(Y_1, \ldots, Y_n\) be independent \(\{0, 1\}\)-valued random variables with \(P[Y_i = 1] = p_i\) and \(\mu = \sum_i p_i\). These are called Poisson trials. Let \(Z = \sum_i Y_i\). Then:

1. Above the mean
   (a) For any \(\delta > 0\),
   \[P[Z \geq (1 + \delta)\mu] \leq \left(\frac{e^\delta}{(1 + \delta)^{(1+\delta)}}\right)^\mu.\]
   (b) For any \(0 < \delta \leq 1\),
   \[P[Z \geq (1 + \delta)\mu] \leq e^{-\mu\delta^2/3}.\]

2. Below the mean
   (a) For any \(0 < \delta < 1\),
   \[P[Z \leq (1 - \delta)\mu] \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{(1-\delta)}}\right)^\mu.\]
   (b) For any \(0 < \delta < 1\),
   \[P[Z \leq (1 - \delta)\mu] \leq e^{-\mu\delta^2/2}.\]

There are innumerable applications of the Chernoff-Cramér method. We discuss a few in the next subsections.

2.4.3 Sub-Gaussian random variables

The bounds in Section 2.4.2 were obtained by computing the moment-generating function explicitly (possibly with some approximations). This is seldom possible. In this section, we give some important examples of concentration inequalities derived from the Chernoff-Cramér method for broad classes of random variables under natural conditions on their distributions.
Sub-Gaussian random variables  Here is our key definition.

**Definition 2.38** (Sub-Gaussian random variables). We say that a random variable $X$ with mean $\mu$ is sub-Gaussian with variance factor $\nu$ if

$$\Psi_{X-\mu}(s) \leq \frac{s^2\nu}{2}, \quad \forall s \in \mathbb{R},$$

(2.38)

for some $\nu > 0$. We use the notation $X \in sG(\nu)$.

Note that the r.h.s. in (2.38) is the cumulant-generating function of a $\mathcal{N}(0, \nu)$. By the Chernoff-Cramér method and (2.26) it follows immediately that

$$\mathbb{P}[X - \mu \leq -\beta] \lor \mathbb{P}[X - \mu \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right),$$

(2.39)

where we used that $X \in sG(\nu)$ implies $-X \in sG(\nu)$. As a quick example, note that this is the approach that we took in Lemma 2.33, that is, we showed that a uniform random variable in $\{-1, 1\}$ is sub-Gaussian with variance factor 1.

When considering (weighted) sums of independent sub-Gaussian random variables, we get the following.

**Theorem 2.39** (General Hoeffding inequality). Suppose $X_1, \ldots, X_n$ are independent random variables where, for each $i$, $X_i \in sG(\nu_i)$ with $0 < \nu_i < +\infty$. For $(w_1, \ldots, w_n) \in \mathbb{R}^n$, let $S_n = \sum_{i \leq n} w_i X_i$. Then

$$S_n \in sG\left(\sum_{i=1}^n w_i^2 \nu_i\right).$$

In particular, for all $\beta > 0$,

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \beta] \leq \exp\left(-\frac{\beta^2}{2 \sum_{i=1}^n w_i^2 \nu_i}\right).$$

Proof. By independence,

$$\Psi_{S_n}(s) = \sum_{i \leq n} \Psi_{w_i X_i}(s) = \sum_{i \leq n} \Psi_{X_i}(sw_i) \leq \sum_{i \leq n} \frac{(sw_i)^2 \nu_i}{2} = \frac{s^2 \sum_{i \leq n} w_i^2 \nu_i}{2}.$$
**Bounded random variables**  For bounded random variables, the previous inequality reduces to a standard bound.

**Theorem 2.40** (Hoeffding’s inequality). Let $X_1, \ldots, X_n$ be independent random variables where, for each $i$, $X_i$ takes values in $[a_i, b_i]$ with $-\infty < a_i \leq b_i < +\infty$. Let $S_n = \sum_{i \leq n} X_i$. For all $\beta > 0$,

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \beta] \leq \exp \left( -\frac{2\beta^2}{\sum_{i \leq n} (b_i - a_i)^2} \right).$$

By Theorem 2.39, it suffices to show that $X_i - \mathbb{E}X_i \in \mathcal{G}(\nu_i)$ with $\nu_i = \frac{1}{4}(b_i - a_i)^2$. We first give a quick proof of a weaker version that uses a trick called *symmetrization*. Suppose the $X_i$s are centered and satisfy $|X_i| \leq c_i$ for some $c_i > 0$. Let $X'_i$ be an independent copy of $X_i$ and let $Z_i$ be an independent uniform random variable in $\{-1, 1\}$. For any $s$, by Jensen’s inequality and (2.29),

$$\mathbb{E} \left[ e^{sX_i} \right] = \mathbb{E} \left[ e^{s\mathbb{E}[X_i - X'_i | X_i]} \right]$$

$$\leq \mathbb{E} \left[ e^{s(X_i - X'_i)} \right]$$

$$= \mathbb{E} \left[ e^{s(X_i - X'_i)} \right]$$

$$= \mathbb{E} \left[ e^{sZ_i(X_i - X'_i)} \right]$$

$$= \mathbb{E} \left[ e^{sZ_i(X_i - X'_i)} \right]$$

$$\leq \mathbb{E} \left[ e^{sZ_i(X_i - X'_i)^2/2} \right]$$

$$\leq e^{-4c_i^2s^2/2},$$

where on the fifth line we used the fact that $X_i - X'_i$ is identically distributed to $-(X_i - X'_i)$ and that $Z_i$ is independent of both $X_i$ and $X'_i$ (together with Lemma A.6). That is, $X_i$ is sub-Gaussian with variance factor $4c_i^2$. By Theorem 2.39, $S_n$ is sub-Gaussian with variance factor $\sum_{i \leq n} 4c_i^2$ and

$$\mathbb{P}[S_n \geq t] \leq \exp \left( -\frac{t^2}{8\sum_{i \leq n} c_i^2} \right).$$
Proof of Theorem 2.40. As pointed out above, it suffices to show that \( X_i - \mathbb{E}X_i \) is sub-Gaussian with variance factor \( \frac{1}{4}(b_i - a_i)^2 \). This is the content of Hoeffding’s lemma (which we will use again in Chapter 3). First an observation:

Lemma 2.41 (Variance of bounded random variables). For any random variable \( Z \) taking values in \([a, b]\) with \(-\infty < a \leq b < +\infty\), we have

\[
\text{Var}[Z] \leq \frac{1}{4}(b - a)^2.
\]

Proof. Indeed

\[
\left| Z - \frac{a + b}{2} \right| \leq \frac{b - a}{2},
\]

and

\[
\text{Var}[Z] = \text{Var}\left[ Z - \frac{a + b}{2} \right] \leq \mathbb{E}\left[ \left( Z - \frac{a + b}{2} \right)^2 \right] \leq \left( \frac{b - a}{2} \right)^2.
\]

Lemma 2.42 (Hoeffding’s lemma). Let \( X \) be a random variable taking values in \([a, b]\) for \(-\infty < a \leq b < +\infty\). Then \( X \in s\mathcal{G}(\frac{1}{4}(b - a)^2) \).

Proof. Note first that \( X - \mathbb{E}X \in [a - \mathbb{E}X, b - \mathbb{E}X] \) and \( \frac{1}{4}\left((b - \mathbb{E}X) - (a - \mathbb{E}X))^2 = \frac{1}{4}(b - a)^2 \right). \) So w.l.o.g. we assume \( \mathbb{E}X = 0 \). Because \( X \) is bounded, \( M_X(s) \) is finite for all \( s \in \mathbb{R} \). From standard results on moment-generating functions (e.g., [Bil12, Section 21]; see also [Dur10, Theorem A.5.1]), for any \( k \in \mathbb{Z} \),

\[
M_X^{(k)}(s) = \mathbb{E}\left[ X^k e^{sX} \right].
\]

Hence

\[
\Psi_X(0) = \log M_X(0) = 0, \quad \Psi_X'(0) = \frac{M_X'(0)}{M_X(0)} = \mathbb{E}X = 0,
\]

and by a Taylor expansion

\[
\Psi_X(s) = \Psi_X(0) + s\Psi_X'(0) + \frac{s^2}{2}\Psi_X''(s^*) = \frac{s^2}{2}\Psi_X''(s^*),
\]

for some \( s^* \in [0, s] \). Therefore it suffices to show that for all \( s \)

\[
\Psi_X''(s) \leq \frac{1}{4}(b - a)^2. \tag{2.40}
\]
Note that
\[
\Psi_X''(s) = \frac{M_X''(s)}{M_X(s)} - \left( \frac{M_X'(s)}{M_X(s)} \right)^2
\]
\[
= \frac{1}{M_X(s)} \mathbb{E} \left[ X^2 e^{sx} \right] - \left( \frac{1}{M_X(s)} \mathbb{E} \left[ X e^{sx} \right] \right)^2
\]
\[
= \mathbb{E} \left[ X^2 \frac{e^{sx}}{M_X(s)} \right] - \left( \mathbb{E} \left[ X \frac{e^{sx}}{M_X(s)} \right] \right)^2.
\]

The trick to conclude is to notice that \(e^{sx}/M_X(s)\) defines a density (i.e., a Radon-Nikodym derivative) on \([a, b]\) with respect to the law of \(X\). The variance under this density—the last line above—is less than \(\frac{1}{4}(b - a)^2\) by Lemma 2.41. This establishes (2.40) and concludes the proof.

Remark 2.43. The change of measure above is known as tilting and is a standard trick in large deviation theory. See, e.g., [Dur10, Section 2.6].

Example: A knapsack problem  To be written. See [FR98, Section 5.3].

2.4.4 Sub-exponential random variables

Unfortunately, not every random variable of interest is sub-Gaussian. A simple example is the square of a Gaussian variable. Indeed, suppose \(X \sim N(0, 1)\). Then \(W = X^2\) is \(\chi^2\)-distributed and its moment-generating function can be computed explicitly. Using the change of variable \(u = x \sqrt{1 - 2s}\), for \(s < 1/2\),

\[
M_W(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{sx} e^{-x^2/2} \, dx
\]
\[
= \frac{1}{\sqrt{1 - 2s}} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-u^2/2} \, du
\]
\[
= \frac{1}{(1 - 2s)^{1/2}}.
\]  (2.41)

When \(s \geq 1/2\), however, \(M_W(s) = +\infty\). In particular, \(W\) cannot be sub-Gaussian for any variance factor \(\nu > 0\). (Note that centering \(W\) produces an additional factor of \(e^{-s}\) in the moment-generating factor which does not prevent it from diverging.)
Further confirming this observation, arguing as in (2.24), the upper tail of $W$ decays as

$$\mathbb{P}[W \geq \beta] = \mathbb{P}[X \geq \sqrt{\beta}]$$

$$\sim \sqrt{\frac{1}{2\pi}} \frac{1}{\sqrt{\beta}} \exp\left(-\frac{1}{2}\beta^{2}\right)$$

$$\sim \sqrt{\frac{1}{2\pi \beta}} \exp(-\beta/2),$$

as $\beta \to +\infty$. That is, it decays exponentially with $\beta$, but much slower than the Gaussian tail.

**Sub-exponential random variables** We now define a broad class of distributions which have such exponential tail decay.

**Definition 2.44** (Sub-exponential random variable). We say that a random variable $X$ with mean $\mu$ is sub-exponential with parameters $(\nu, \alpha)$ if

$$\Psi_{X-\mu}(s) \leq \frac{s^2 \nu}{2}, \quad \forall |s| \leq \frac{1}{\alpha},$$

(2.42)

for some $\nu, \alpha > 0$. We write $X \in sE(\nu, \alpha)$.

Observe that the key difference between (2.38) and (2.42) is the interval of $s$ over which it holds. As we will see below, the parameter $\alpha$ dictates the exponential decay rate of the tail. The specific form of the bound in (2.42) is natural once one notices that, as $|s| \to 0$, a centered random variable with variance $\nu$ should roughly satisfy

$$\log \mathbb{E}[e^{sX}] \approx \log \left(1 + s\mathbb{E}[X] + \frac{s^2 \mathbb{E}[X^2]}{2}\right) \approx \log \left(1 + \frac{s^2 \nu}{2}\right) \approx \frac{s^2 \nu}{2}. $$

57
Returning to the $\chi^2$ distribution, note that from (2.41) we have for $|s| < 1/4$

\[
\Psi_{W^{-1}}(s) = -s - \frac{1}{2} \log(1 - 2s) = -s - \frac{1}{2} \left[ -\sum_{i=1}^{+\infty} \frac{(2s)^i}{i} \right] = \frac{s^2}{2} \left[ 4 \sum_{i=2}^{+\infty} \frac{(2s)^{i-2}}{i} \right] \leq \frac{s^2}{2} \left[ 2 \sum_{i=2}^{+\infty} |1/2|^{i-2} \right] \leq \frac{s^2}{2} \times 4.
\]

Hence $W \in sE(4, 1/4)$.

Using the Chernoff-Cramér bound (Lemma 2.32), we obtain the following tail bound for sub-exponential variables.

**Theorem 2.45** (Sub-exponential tail bound). Suppose the random variable $X$ with mean $\mu$ is sub-exponential with parameters $(\nu, \alpha)$. Then for all $\beta \in \mathbb{R}_+$

\[
\mathbb{P}[X - \mu \geq \beta] = \begin{cases} 
\exp(-\frac{\beta^2}{2\nu}), & \text{if } 0 \leq \beta \leq \nu/\alpha, \\
\exp(-\frac{\beta}{2\alpha}), & \text{if } \beta > \nu/\alpha.
\end{cases}
\]

(2.43)

In words, the tail decays exponentially at large deviations but behaves as in the sub-Gaussian case for smaller deviations. We will see below that this awkward double-tail allows to extrapolate naturally between different regimes. First we prove the claim.

**Proof of Theorem 2.45.** We start by applying the Chernoff-Cramér bound (Lemma 2.32). For any $\beta > 0$ and $|s| \leq 1/\alpha$

\[
\mathbb{P}[X - \mu \geq \beta] \leq \exp\left(-s\beta + \Psi_X(s)\right) \leq \exp\left(-s\beta + s^2\nu/2\right).
\]

At this point, the proof diverges from the sub-Gaussian case because the optimal choice of $s$ depends on $\beta$ because of the additional constraint $|s| \leq 1/\alpha$. When $s^* = \beta/\nu$ satisfies $s^* \leq 1/\alpha$, the quadratic function of $s$ in the exponent is minimized at $s^*$, giving the bound

\[
\mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right),
\]

58
for $0 \leq \beta \leq \nu/\alpha$.

On the other hand, when $\beta > \nu/\alpha$, the exponent is strictly decreasing over the interval $s \leq 1/\alpha$. Hence the optimal choice is $s^* = 1/\alpha$, which produces the bound

$$\mathbb{P}[X \geq \beta] \leq \exp \left( -\frac{\beta}{\alpha} + \frac{\nu}{2\alpha^2} \right)$$
$$\leq \exp \left( -\frac{\beta}{\alpha} + \frac{\beta}{2\alpha} \right)$$
$$= \exp \left( -\frac{\beta}{2\alpha} \right),$$

where we used that $\nu < \beta\alpha$ on the second line.

For (weighted) sums of independent sub-exponential random variables, we get the following.

**Theorem 2.46 (General Bernstein inequality).** Suppose $X_1, \ldots, X_n$ are independent random variables where, for each $i$, $X_i \in s\mathcal{E}(\nu_i, \alpha_i)$ with $0 < \nu_i, \alpha_i < +\infty$. For $(w_1, \ldots, w_n) \in \mathbb{R}^n$, let $S_n = \sum_{i\leq n} w_i X_i$. Then

$$S_n \in s\mathcal{E} \left( \sum_{i=1}^n w_i^2 \nu_i, \max_i |w_i| \alpha_i \right).$$

In particular, for all $\beta > 0$,

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \beta] \leq \begin{cases} 
\exp \left( -\frac{\beta^2}{2\sum_{i=1}^n w_i^2 \nu_i} \right), & \text{if } 0 \leq \beta \leq \frac{\sum_{i=1}^n w_i^2 \nu_i}{\max_i |w_i| \alpha_i}, \\
\exp \left( -\frac{\beta}{2\max_i |w_i| \alpha_i} \right), & \text{if } \beta > \frac{\sum_{i=1}^n w_i^2 \nu_i}{\max_i |w_i| \alpha_i}.
\end{cases}$$

**Proof.** By independence,

$$\Psi_{S_n}(s) = \sum_{i\leq n} \Psi_{w_i X_i}(s) = \sum_{i\leq n} \Psi_{X_i}(s w_i) \leq \sum_{i\leq n} \frac{(s w_i)^2 \nu_i}{2} = \frac{s^2 \sum_{i\leq n} w_i^2 \nu_i}{2},$$

provided $|s w_i| < 1/\alpha_i$ for all $i$, that is,

$$|s| < \frac{1}{\max_i |w_i| \alpha_i}.$$
**Bounded random variables** We apply the previous result to bounded random variables.

**Theorem 2.47 (Bernstein’s inequality).** Let $X_1, \ldots, X_n$ be independent random variables where, for each $i$, $X_i$ has mean $\mu_i$, variance $\nu_i$ and satisfies $|X_i - \mu_i| \leq c$ for some $0 < c < +\infty$. Let $S_n = \sum_{i \leq n} X_i$. For all $\beta > 0$,

$$
P \left[ S_n - E[S_n] \geq \beta \right] \leq \begin{cases} 
\exp \left( -\frac{\beta^2}{4 \sum_{i=1}^{n} \nu_i} \right), & \text{if } 0 \leq \beta \leq \frac{\sum_{i=1}^{n} \nu_i}{c}, \\
\exp \left( -\frac{\beta}{4c} \right), & \text{if } \beta > \frac{\sum_{i=1}^{n} \nu_i}{c}.
\end{cases}
$$

**Proof.** We claim that $X_i \in sE(2\nu_i, 2c)$. To establish the claim, we derive a bound on all moments of $X_i$. Note that for all integers $k \geq 2$

$$
E[|X_i - \mu_i|^k] \leq c^{k-2}E[|X_i - \mu_i|^2] = c^{k-2}\nu_i.
$$

Hence, first applying the dominated convergence theorem (see e.g. [Bil12, Section 21] for details), we have for $|s| < \frac{1}{2c}$

$$
E[e^{s(X_i - \mu_i)}] = \sum_{k=0}^{+\infty} \frac{s^k}{k!} E[(X_i - \mu_i)^k] 
$$

$$
\leq 1 + s E[(X_i - \mu_i)] + \sum_{k=2}^{+\infty} \frac{s^k}{k!} c^{k-2} \nu_i 
$$

$$
\leq 1 + \frac{s^2\nu_i}{2} + \frac{s^2\nu_i}{3!} \sum_{k=3}^{+\infty} (cs)^{k-2} 
$$

$$
= 1 + \frac{s^2\nu_i}{2} \left\{ 1 + \frac{1}{3} - \frac{cs}{3} \right\} 
$$

$$
\leq 1 + \frac{s^2\nu_i}{2} \left\{ 1 + \frac{1}{3} - \frac{1/2}{1/2} \right\} 
$$

$$
\leq 1 + \frac{s^2\nu_i}{2} 
$$

$$
\leq \exp \left( \frac{s^2}{2} \nu_i \right).
$$

That proves the claim. Using the General Bernstein inequality (Theorem 2.46) gives the result. 

It may seem counter-intuitive to derive a tail bound based on the sub-exponential property of bounded random variables when we have already done so using their
sub-Gaussian behavior. After all, the latter is on the surface a strengthening of the former. However, note that we have obtained a better bound in Theorem 2.47 than we did in Theorem 2.40 when \( \beta \) is not too large. That improvement stems from the use of the variance for moderate deviations, at the expense of the sub-Gaussian tail for large deviations. We give an example below.

**Example 2.48 (Erdős-Rényi: maximum degree)**. Let \( G_n = (V_n, E_n) \sim G_{n,p_n} \) be an Erdős-Rényi graph with \( n \) vertices and density \( p_n \). Recall that two vertices \( u, v \in V_n \) are adjacent if \( \{u, v\} \in E_n \) and that the set of adjacent vertices of \( v \), denoted by \( N(v) \), is called the neighborhood of \( v \). The degree of \( v \) is the size of its neighborhood, i.e. \( \delta(v) = |N(v)| \). Here we study the maximum degree of \( G_n \)

\[
D_n = \max_{v \in V_n} \delta(v).
\]

We focus on the regime \( np_n = \omega(\log n) \). Note that, for any vertex \( v \in V_n \), its degree is \( \text{Bin}(n-1, p_n) \) by independence of the edges. In particular its expected degree is \( (n-1)p_n \). To prove a high-probability upper bound on the maximum \( D_n \), we need to control the deviation of each vertex’s degree from its expectation. Observe that the degrees are not independent. Instead we apply a union bound over all vertices, after using a tail bound.

**Claim 2.49**. For any \( \varepsilon > 0 \), as \( n \to +\infty \),

\[
\mathbb{P} \left[ |D_n - np_n| \geq 2\sqrt{(1 + \varepsilon)np_n \log n} \right] \to 0.
\]

**Proof**. For a fixed vertex \( v \), think of \( \delta(v) = S_{n-1} \sim \text{Bin}(n-1, p_n) \) as a sum of \( n-1 \) independent \{0, 1\}-valued random variables, one for each possible edge. That is, \( S_{n-1} = \sum_{i=1}^{n-1} X_i \) where \( X_i \) are bounded random variables. The mean of \( X_i \) is \( p_n \) and its variance is \( p_n(1-p_n) \). So in Bernstein’s inequality (Theorem 2.47), we can take \( \mu_i = p_n, \nu_i = p_n(1-p_n) \) and \( c := 1 \) for all \( i \). We get

\[
\mathbb{P} \left[ S_{n-1} \geq (n-1)p_n + \beta \right] \leq \begin{cases} 
\exp \left( -\frac{\beta^2}{4\nu} \right), & \text{if } 0 \leq \beta \leq \nu, \\
\exp \left( -\frac{\beta}{4} \right), & \text{if } \beta > \nu.
\end{cases}
\]

where \( \nu = (n-1)p_n(1-p_n) = \omega(\log n) \) by assumption. The value of \( \beta \) is chosen to be the smallest that will produce a tail probability less than \( n^{-1-\varepsilon} \) for \( \varepsilon > 0 \), that is,

\[
\beta = \sqrt{4(n-1)p_n(1-p_n)} \times \sqrt{(1+\varepsilon)\log n} = o(\nu),
\]

which falls in the lower regime of the tail bound. In particular, \( \beta = o(np_n) \). Finally by a union bound over \( v \in V_n \)

\[
\mathbb{P} \left[ D_n \geq (n-1)p_n + \sqrt{4(1+\varepsilon)p_n(1-p_n)(n-1)\log n} \right] \leq n \times \frac{1}{n^{1+\varepsilon}} \to 0.
\]
The same holds in the other direction. That proves the claim.

Had we used Hoeffding’s inequality (Theorem 2.40) in the proof of Claim 2.49 we would have had to take \( \beta = \sqrt{(1 + \varepsilon) \log n} \). That would have produced a much weaker bound when \( p_n = o(1) \). Indeed the advantage of Bernstein’s inequality is that it makes explicit use of the variance, which when \( p_n = o(1) \) is much smaller than the worst case for bounded variables.

\[ \square \]

### 2.4.5 Epsilon-nets and chaining

Suppose we are interested in bounding the expectation or tail of the supremum of a stochastic process

\[ \sup_{t \in T} X_t, \]

where \( T \) is an arbitrary index set and the \( X_t \)'s are real-valued random variables. To avoid measurability issues, we assume throughout that \( T \) is countable.

So far we have developed tools that can handle cases where \( T \) is finite. When the supremum is over an infinite index set, however, new ideas are required. One way to proceed is to apply a tail inequality to a sufficiently dense finite subset of the index set and then extend the resulting bound by a Lipschitz continuity argument. We present this type of approach in this section, as well as a multi-scale version known as chaining.

First we summarize one important special case that will be useful below: \( T \) is finite and \( X_t \) is sub-Gaussian.

**Theorem 2.50** (Maximal inequalities: sub-Gaussian case). Let \( \{X_t\}_{t \in T} \) be a stochastic process with finite index set \( T \). Assume that there is \( \nu > 0 \) such that, for all \( t \), \( X_t \in sG(\nu) \) and that \( \mathbb{E}[X_t] = 0 \). Then

\[ \mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq \sqrt{2\nu \log |T|}, \]

and, for all \( \beta > 0 \),

\[ \mathbb{P} \left[ \sup_{t \in T} X_t \geq \sqrt{2\nu \log |T|} + \beta \right] \leq \exp \left( -\frac{\beta^2}{2\nu} \right). \]

**Proof.** For the expectation, we apply a variation on the Chernoff-Cramér method (Section 2.4). Naively, we could bound the supremum \( \sup_{t \in T} X_t \) by the sum \( \sum_{t \in T} |X_t| \), but that would lead to a bound growing linearly with the cardinality \( |T| \). Instead we first take an exponential, which tends to amplify the largest term.
and produces a much stronger bound. Specifically, by Jensen’s inequality, for any $s > 0$

$$
\mathbb{E} \left[ \sup_{t \in T} X_t \right] = \frac{1}{s} \mathbb{E} \left[ \sup_{t \in T} s X_t \right] \leq \frac{1}{s} \log \mathbb{E} \left[ \exp \left( \sup_{t \in T} s X_t \right) \right].
$$

Since $e^{a+b} \leq e^a + e^b$ by the non-negativity of the exponential, we can bound

$$
\mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq \frac{1}{s} \log \left( \sum_{t \in T} \mathbb{E} \left[ \exp (s X_t) \right] \right) = \frac{1}{s} \log \left( \sum_{t \in T} M_{X_t} (s) \right) \leq \frac{1}{s} \log \left( |T| e^{\frac{s^2 \nu}{2}} \right) = \frac{\log |T|}{s} + \frac{s \nu}{2}.
$$

The optimal choice of $s$ is obtained when the two terms in the sum above are equal, that is, $s = \sqrt{2 \nu^{-1} \log |T|}$, which gives finally

$$
\mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq \sqrt{2 \nu \log |T|},
$$

as claimed.

For the tail inequality, we use a union bound and (2.39)

$$
\mathbb{P} \left[ \sup_{t \in T} X_t \geq \sqrt{2 \nu \log |T|} + \beta \right] \leq \sum_{t \in T} \mathbb{P} \left[ X_t \geq \sqrt{2 \nu \log |T|} + \beta \right] \leq |T| \exp \left( -\frac{\left( \sqrt{2 \nu \log |T|} + \beta \right)^2}{2 \nu} \right) \leq \exp \left( -\frac{\beta^2}{2 \nu} \right),
$$

as claimed, where we used that $\beta > 0$ on the last line.

**Epsilon-nets and covering numbers**

Moving on to infinite index sets, we first define the notion of an $\varepsilon$-net. This notion requires that a pseudometric $\rho$ (i.e. $\rho : T \times T \to \mathbb{R}_+$ is symmetric and satisfies the triangle inequality) be defined over $T$. 

63
Definition 2.51 (\(\varepsilon\)-net). Let \(T\) be a subset of a pseudometric space \((M, \rho)\) and let \(\varepsilon > 0\). The collection of points \(N \subseteq M\) is called an \(\varepsilon\)-net of \(T\) if

\[
T \subseteq \bigcup_{t \in N} B_\rho(t, \varepsilon),
\]

where \(B_\rho(t, \varepsilon) = \{s \in T : \rho(s, t) \leq \varepsilon\}\), that is, each element of \(T\) is within distance \(\varepsilon\) of an element in \(N\). The smallest cardinality of an \(\varepsilon\)-net of \(T\) is called the covering number

\[
N(T, \rho, \varepsilon) = \inf\{|N| : N \text{ is an } \varepsilon\text{-net of } T\}.
\]

A natural way to construct an \(\varepsilon\)-net is the following algorithm. Start with \(N = \emptyset\) and successively add a point from \(T\) to \(N\) at distance at least \(\varepsilon\) from all other previous points until it is not possible to do so anymore. Provided \(T\) is compact, this procedure will terminate after a finite number of steps. This leads to the following dual perspective.

Definition 2.52 (\(\varepsilon\)-packing). Let \(T\) be a subset of a metric space \((M, \rho)\) and let \(\varepsilon > 0\). The collection of points \(N \subseteq T\) is called an \(\varepsilon\)-packing of \(T\) if

\[
B_\rho(t, \varepsilon) \cap B_\rho(t', \varepsilon) = \emptyset, \quad \forall t \neq t' \in N
\]

that is, every pair of element of \(N\) is at distance at least \(\varepsilon\). The largest cardinality of an \(\varepsilon\)-packing of \(T\) is called the packing number

\[
P(T, \rho, \varepsilon) = \inf\{|N| : N \text{ is an } \varepsilon\text{-packing of } T\}.
\]

Lemma 2.53 (Covering and packing numbers). For any \(T \subseteq M\) and all \(\varepsilon > 0\),

\[
N(T, \rho, \varepsilon) \leq P(T, \rho, \varepsilon).
\]

Proof. Observe that a maximal \(\varepsilon\)-packing \(N\) is an \(\varepsilon\)-net. Indeed, by maximality, any element of \(T \setminus N\) is at distance \(\varepsilon\) from an element of \(N\). \qed

Example 2.54 (Sphere in \(\mathbb{R}^k\)). Let \(S^{k-1}\) be the sphere of radius 1 centered around the origin in \(\mathbb{R}^k\) with the Euclidean metric. Let \(0 < \varepsilon < 1\).

Claim 2.55.

\[
N(S, \rho, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^k.
\]
Let $N$ be any maximal $\varepsilon$-packing of $S$. We show that $|N| \leq (3/\varepsilon)^k$, which implies the claim by Lemma 2.55. The balls of radius $\varepsilon/2$ around points in $N$, \(\mathbb{B}^k(x_i, \varepsilon/2) : x_i \in N\), satisfy two properties:

1. They are pairwise disjoint: if \(z \in \mathbb{B}^k(x_i, \varepsilon/2) \cap \mathbb{B}^k(x_j, \varepsilon/2)\), then \(\|x_i - x_j\|_2 \leq \|x_i - z\|_2 + \|x_j - z\|_2 \leq \varepsilon\), a contradiction.

2. They are included in the ball of radius $3/2$ around the origin: if \(z \in \mathbb{B}^k(x_i, \varepsilon/2)\), then \(\|z\|_2 \leq \|z - x_i\|_2 + \|x_i\| \leq \varepsilon/2 + 1 \leq 3/2\).

The volume of a ball of radius is \(\varepsilon/2\) is \(\pi^{k/2}(\varepsilon/2)^k\) and that of a ball of radius $3/2$ is \(\pi^{k/2}(3/2)^k\). Dividing one by the other proves the claim. \(\blacksquare\)

The basic approach to use an $\varepsilon$-net for controlling the supremum of a stochastic process is the following. We say that a stochastic process \(\{X_t\}_{t \in T}\) is Lipschitz for metric $\rho$ on $T$ if there is a random variable $0 < K < +\infty$ such that

\[|X_t - X_s| \leq K \rho(s, t), \quad \forall s, t \in T.\]

If, in addition, $X_t$ is sub-Gaussian for all $t$, then we can bound the expectation or tail probability of the supremum of \(\{X_t\}_{t \in T}\) if we can bound the expectation or tail probability of the Lipschitz constant $K$. Indeed, let $N$ be an $\varepsilon$-net of $T$ and, for each $t \in T$, let $\pi(t)$ be the closest element of $N$ to $t$. We will refer to $\pi$ as the projection map of $N$. We then have the inequality

\[\sup_{t \in T} X_t \leq \sup_{t \in T} (X_t - X_{\pi(t)}) + \sup_{t \in T} X_{\pi(t)} \leq K \varepsilon + \sup_{s \in N} X_s,\]

where we can use Theorem 2.50 to bound the last term. We give an example next. Another example can be found in Section 2.4.6.

**Spectral norm of a random matrix** For a $m \times n$ matrix $A \in \mathbb{R}^{m \times n}$, recall that the spectral norm is defined as

\[\|A\| := \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x \in S^{n-1}} \langle Ax, y \rangle,\]

(2.44)

where $S^{n-1}$ is the sphere of radius 1 around the origin in $\mathbb{R}^n$. (To see the equality above, note that Cauchy-Schwarz implies $\langle Ax, y \rangle \leq \|Ax\|_2 \|y\|_2$ and that one can take $y = Ax/\|Ax\|_2$ for any $x$ such that $Ax \neq 0$ in the rightmost expression.)
Theorem 2.56. Let $A \in \mathbb{R}^{m \times n}$ be a random matrix whose entries are centered, independent and sub-Gaussian with variance factor $\nu$. Then there exist a constant $0 < C < +\infty$ such that, for all $t > 0$,

$$\|A\| \leq C \sqrt{\nu} (\sqrt{m} + \sqrt{n} + t),$$

with probability at least $1 - e^{-t^2}$.

Proof. Fix $\varepsilon = 1/4$. By Claim 2.55, there is an $\varepsilon$-net $N$ (respectively $M$) of $\mathbb{S}^{n-1}$ (respectively $\mathbb{S}^{m-1}$) with $|N| \leq 12^n$ (respectively $|M| \leq 12^m$). We proceed in two steps:

1. We first apply the general Hoeffding inequality (Theorem 2.39) to control the deviations of the supremum in (2.44) restricted to $N$ and $M$.

2. We then extend the bound to the full supremum by Lipschitz continuity.

Formally, the result follows from the following two lemmas.

Lemma 2.57 (Spectral norm: $\varepsilon$-net). Let $N$ and $M$ be as above. For $C$ large enough, for all $t > 0$,

$$\mathbb{P} \left[ \max_{x \in N, y \in M} \langle Ax, y \rangle \geq \frac{1}{2} C \sqrt{\nu} (\sqrt{m} + \sqrt{n} + t) \right] \leq e^{-t^2}.$$ 

Lemma 2.58 (Spectral norm: Lipschitz constant). For any $\varepsilon$-nets $N$ and $M$ of $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$ respectively, the following inequalities hold

$$\sup_{x \in N} \langle Ax, y \rangle \leq \|A\| \leq \frac{1}{1 - 2\varepsilon} \sup_{x \in N} \sup_{y \in M} \langle Ax, y \rangle.$$ 

Proof of Lemma 2.57. Note that the quantity $\langle Ax, y \rangle$ is a linear combination of independent random variables:

$$\langle Ax, y \rangle = \sum_{i,j} x_i y_j A_{ij}.$$ 

By the general Hoeffding inequality (Theorem 2.39), $\langle Ax, y \rangle$ is sub-Gaussian with variance factor

$$\sum_{i,j} (x_i y_j)^2 \nu = \|x\|^2_2 \|y\|^2_2 \nu = \nu.$$
for all \( x \in N \) and \( y \in M \). In particular, for all \( \beta > 0 \),
\[
P \left[ \langle Ax, y \rangle \geq \beta \right] \leq \exp \left( -\frac{\beta^2}{2\nu} \right).
\]
Hence, by a union bound over \( N \) and \( M \),
\[
\mathbb{P} \left[ \max_{x \in N, y \in M} \langle Ax, y \rangle \geq \frac{1}{2} C \sqrt{\nu (\sqrt{m} + \sqrt{n} + t)} \right]
\leq \sum_{x \in N} \sum_{y \in M} \mathbb{P} \left[ \langle Ax, y \rangle \geq \frac{1}{2} C \sqrt{\nu (\sqrt{m} + \sqrt{n} + t)} \right]
\leq |N||M| \exp \left( -\frac{1}{2\nu} \left\{ \frac{1}{2} C \sqrt{\nu (\sqrt{m} + \sqrt{n} + t)} \right\}^2 \right)
\leq 12^{n+m} \exp \left( -\frac{C^2}{8} \left\{ m + n + t^2 \right\} \right)
\leq e^{-t^2},
\]
for \( C^2/8 = \ln 12 \geq 1 \), where in the third inequality we ignored all cross-products since they are non-negative.

Proof of Lemma 2.58. The first inequality is immediate. For the second inequality, we will use the following observation
\[
\langle Ax, y \rangle - \langle Ax_0, y_0 \rangle = \langle Ax, y - y_0 \rangle + \langle A(x - x_0), y_0 \rangle. \tag{2.45}
\]
Fix \( x \in S^{n-1} \) and \( y \in S^{m-1} \) such that \( \langle Ax, y \rangle = \|A\| \), and let \( x_0 \in N \) and \( y_0 \in M \) such that
\[
\|x - x_0\|_2 \leq \varepsilon \quad \text{and} \quad \|y - y_0\|_2 \leq \varepsilon.
\]
Then (2.45), Cauchy-Schwarz and the definition of the spectral norm imply
\[
\|A\| - \langle Ax_0, y_0 \rangle \leq \|A\|\|x\|_2\|y - y_0\|_2 + \|A\|\|x - x_0\|_2\|y_0\|_2 \leq 2\varepsilon \|A\|.
\]
Rearranging gives the claim.

Putting the two lemmas together concludes the proof of Theorem 2.56.
Chaining method

We go back to the inequality

\[ \sup_{t \in T} X_t \leq \sup_{t \in T} (X_t - X_{\pi(t)}) + \sup_{t \in T} X_{\pi(t)}. \]  \hspace{1cm} (2.46)

We consider cases where we may not have a good almost sure bound on the Lipschitz constant, but where we can control increments uniformly in the following probabilistic sense. We say that a stochastic process \( \{X_t\}_{t \in T} \) has sub-Gaussian increments on \((T, \rho)\) if there exists \( 0 \leq K < +\infty \) such that

\[ X_t - X_s \in s\mathcal{G}(K^2 \rho(s, t)^2), \quad \forall s, t \in T. \]

In (2.46), the first term on the r.h.s. remains a supremum over an infinite set. To control it, the chaining method repeats the argument at progressively smaller scales, leading to the following inequality. The diameter of \( T \), denoted by \( \text{diam}(T) \), is defined as

\[ \text{diam}(T) = \sup \{ \rho(s, t) : s, t, \in T \}. \]

We apply the discrete Dudley’s inequality in Section 2.4.8.

**Theorem 2.59** (Discrete Dudley’s Inequality). Let \( \{X_t\}_{t \in T} \) be a zero-mean stochastic process with sub-Gaussian increments on \((T, \rho)\) and assume \( \text{diam}(T) \leq 1 \). Then

\[ \mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq C \sum_{k=0}^{+\infty} 2^{-k} \sqrt{\log \mathcal{N}(T, \rho, 2^{-k})}. \]

for some constant \( 0 \leq C < +\infty \).

**Proof.** Recall that we assume there is a countable \( T_0 \subseteq T \) such that \( \sup_{t \in T} X_t = \sup_{t \in T_0} X_t \) almost surely. Let \( T_k \subseteq T_0, k \geq 1 \), be a sequence of finite sets such that \( T_k \uparrow T_0 \). By monotone convergence

\[ \mathbb{E} \left[ \sup_{t \in T_0} X_t \right] = \sup_{k \geq 1} \mathbb{E} \left[ \sup_{t \in T_k} X_t \right]. \]

Moreover, \( \mathcal{N}(T_k, \rho, \varepsilon) \leq \mathcal{N}(T, \rho, \varepsilon) \) for any \( \varepsilon > 0 \) since \( T_k \subseteq T \). Hence it suffices to handle the case \( |T| < +\infty \).
\(\varepsilon\)-nets at all scales For each \(k \geq 0\), let \(N_k\) be an \(2^{-k}\)-net with \(|N_k| = \mathcal{N}(T, \rho, 2^{-k})\) and projection map \(\pi_k\). Because \(\text{diam}(T) \leq 1\), \(N_0 = \{t_0\}\) where \(t_0 \in T\) can be taken arbitrarily. Moreover, because \(T\) is finite, there is \(1 \leq \kappa < +\infty\) such that \(N_k = T\) for all \(k \geq \kappa\). In particular, \(\pi_\kappa(t) = t\) for all \(t \in T\). Hence, by a telescoping argument,

\[
X_t = X_{t_0} + \sum_{k=0}^{\kappa-1} \left( X_{\pi_{k+1}(t)} - X_{\pi_k(t)} \right).
\]

Taking a supremum and an expectation gives

\[
\mathbb{E} \left[ \sup_{t \in T} X_t \right] = \sum_{k=0}^{\kappa-1} \mathbb{E} \left[ \sup_{t \in T} \left( X_{\pi_{k+1}(t)} - X_{\pi_k(t)} \right) \right],
\]

where we used \(\mathbb{E}[X_{t_0}] = 0\).

Sub-Gaussian bound We use Theorem 2.50 to bound the expectation in (2.47). The number of elements in the supremum is at most

\[
|\{(\pi_k(t), \pi_{k+1}(t)) : t \in T\}| \leq |N_k| \times |N_{k+1}|
\]

\[
= |N_k| \times |N_{k+1}|
\]

\[
\leq \mathcal{N}(T, \rho, 2^{-k-1})^2.
\]

For any \(t \in T\), by the triangle inequality

\[
\rho(\pi_k(t), \pi_{k+1}(t)) \leq \rho(\pi_k(t), t) + \rho(t, \pi_{k+1}(t)) \leq 2^{-k} + 2^{-k-1} \leq 2^{-k+1},
\]

so that

\[
X_{\pi_{k+1}(t)} - X_{\pi_k(t)} \in \mathcal{S}_{G}(K2^{-2k+2}),
\]

for some \(0 \leq K < +\infty\) by the sub-Gaussian increments assumption. We can therefore apply Theorem 2.50 to get

\[
\mathbb{E} \left[ \sup_{t \in T} \left( X_{\pi_{k+1}(t)} - X_{\pi_k(t)} \right) \right] \leq \sqrt{2K^22^{-2k+2} \log(\mathcal{N}(T, \rho, 2^{-k-1}))^2}
\]

\[
\leq C2^{-k-1} \sqrt{\log \mathcal{N}(T, \rho, 2^{-k-1})},
\]

for some constant \(0 \leq C < +\infty\). Plugging back into (2.47),

\[
\mathbb{E} \left[ \sup_{t \in T} X_t \right] \leq \sum_{k=0}^{\kappa-1} C2^{-k-1} \sqrt{\log \mathcal{N}(T, \rho, 2^{-k-1})},
\]

which implies the claim.\(\blacksquare\)
Using a similar argument, one can derive a tail inequality.

**Theorem 2.60 (Chaining Tail Inequality).** Let \( \{X_t\}_{t \in T} \) be a zero-mean stochastic process with sub-Gaussian increments on \((T, \rho)\) and assume \( \text{diam}(T) \leq 1 \). Then, for all \( t_0 \in T \) and \( \beta > 0 \),

\[
\mathbb{P} \left[ \sup_{t \in T} (X_t - X_{t_0}) \geq C \sum_{k=0}^{+\infty} 2^{-k} \sqrt{\log \mathcal{N}(T, \rho, 2^{-k})} + \beta \right] \leq C \exp \left( -\frac{\beta^2}{C} \right),
\]

for some constant \( 0 \leq C < +\infty \).

### 2.4.6 Data science: Johnson-Lindenstrauss and application to sparse recovery

In this section we discuss an application of the Chernoff-Cramér method to dimensionality reduction. We use once again an \( \varepsilon \)-net argument.

**Johnson-Lindenstrauss lemma** The Johnson-Lindenstrauss lemma states roughly that, for any collection of points in a high-dimensional Euclidean space, one can find an embedding of much lower dimension that preserves the metric relationships of the points, i.e., their norms and distances. Remarkably, no structure is assumed on the points and the result is independent of the original dimension. The method of proof simply involves applying a well-chosen random linear mapping.

Before stating and proving the lemma, we define the mapping employed. Let \( A \) be a \( d \times n \) matrix whose entries are independent \( \mathcal{N}(0, 1) \). Note that, for any fixed \( z \in \mathbb{R}^n \),

\[
\mathbb{E} \left\| A z \right\|^2 = \mathbb{E} \left[ \sum_{i=1}^{d} \left( \sum_{j=1}^{n} A_{ij} z_j \right)^2 \right] = d \text{Var} \left[ \sum_{j=1}^{n} A_{ij} z_j \right] = d \| z \|^2_2,
\]

where we used the independence of the \( A_{ij} \)’s and

\[
\mathbb{E} \left[ \sum_{j=1}^{n} A_{ij} z_j \right] = 0.
\]

Hence the linear mapping \( L = \frac{1}{\sqrt{d}} A \) is an isometry “on average.” We use the Chernoff-Cramér method to prove a high probability result.

**Theorem 2.61.** Let \( x^{(1)}, \ldots, x^{(m)} \) be arbitrary points in \( \mathbb{R}^n \). Fix \( \delta, \theta > 0 \) and \( d \geq \frac{8}{3} \theta^{-2}(\log m + \frac{1}{2} \log \delta^{-1}) \). Let \( A \) be a \( d \times n \) matrix whose entries are independent
and consider the linear mapping \( L = \frac{1}{\sqrt{d}} A \). Then the following hold with probability at least \( 1 - \delta \):

\[
(1 - \theta) \|x^{(i)}\|_2 \leq \|Lx^{(i)}\|_2 \leq (1 + \theta) \|x^{(i)}\|_2, \quad \forall i
\]

and

\[
(1 - \theta) \|x^{(i)} - x^{(j)}\|_2 \leq \|Lx^{(i)} - Lx^{(j)}\|_2 \leq (1 + \theta) \|x^{(i)} - x^{(j)}\|_2, \quad \forall i, j.
\]

**Proof.** Fix a \( z \in \mathbb{R}^n \) with \( \|z\|_2 = 1 \). To prove the theorem it suffices to show that

\[
\mathbb{P}[\|Lz\|_2 > 1 + \theta] \leq \exp \left( -\frac{3}{4} d \theta^2 \right),
\]

(2.50)

where we used (2.48) and (2.49) to compute the variance. Hence

\[
W = \|Az\|_2^2 = \sum_{k=1}^{d} (Az)_k^2,
\]

is a sum of squares of independent Gaussians (also known as \( \chi^2 \)-distribution with \( d \) degrees of freedom or \( \Gamma(d/2, 2) \)). Using independence and the change of variable \( u = x\sqrt{1 - 2s} \), for \( s < 1/2 \)

\[
M_W(s) = \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{sx^2} e^{-x^2/2} \, dx \right)^d
\]

\[
= \left( \frac{1}{\sqrt{1 - 2s}} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-u^2/2} \, du \right)^d
\]

\[
= \frac{1}{(1 - 2s)^{d/2}}.
\]
Applying the Chernoff-Cramér bound (2.25) with \( s = \frac{1}{2}(1 - d/\beta) \) gives

\[
P[W \geq \beta] \leq \frac{M_W(s)}{e^{s\beta}} = \frac{1}{e^{s\beta}(1 - 2s)d/2} = e^{(d-\beta)/2} \left( \frac{\beta}{d} \right)^{d/2}.
\]

Finally, take \( \beta = d(1 + \theta)^2 \). Rearranging we get

\[
P[\|Lz\|_2 \geq 1 + \theta] = P[\|Az\|_2^2 \geq d(1 + \theta)^2]
\]

\[
= P[W \geq \beta]
\]

\[
\leq e^{d(1-(1+\theta)^2)/2} \left\{ (1 + \theta)^2 \right\}^{d/2}
\]

\[
\leq \exp \left( -d(\theta + \theta^2/2 - \log(1 + \theta)) \right)
\]

\[
\leq \exp \left( -\frac{3}{4}d\theta^2 \right),
\]

where we used that \( \log(1 + x) \leq x - x^2/4 \) on \([0, 1]\). (To check the inequality, note that the two sides are equal at 0 and compare the derivatives on \([0, 1]\).)

Our choice of \( d \) gives that the r.h.s. in (2.50) is less than \( \delta/m^2 \).

Remark 2.63. The Johnson-Lindenstrauss lemma is essentially optimal [Alo03]. Note however that it relies crucially on the use of the Euclidean norm [BC03].

To give some further geometric insights into the Johnson-Lindenstrauss lemma, we make a series of observations:

1. The rows of \( \frac{1}{\sqrt{n}} A \) are “on average” orthonormal. Indeed, note that for \( i \neq j \)

\[
\mathbb{E} \left[ \frac{1}{n} \sum_{k=1}^{n} A_{ik}A_{jk} \right] = \mathbb{E}[A_{i1}] \mathbb{E}[A_{j1}] = 0,
\]

by independence and

\[
\mathbb{E} \left[ \frac{1}{n} \sum_{k=1}^{n} A_{ik}^2 \right] = \mathbb{E}[A_{i1}^2] = 1,
\]

since the \( A_{ik} \)s have 0 mean. When \( n \) is large, these two quantities are concentrated around their mean.

2. So, roughly speaking, \( \frac{1}{\sqrt{n}} Az \) corresponds to an orthogonal projection of \( z \) to a uniformly random subspace of dimension \( d \) and \( \frac{1}{\sqrt{n}} \|Az\|_2 \) gives the norm of that projection. Then (2.48) shows that \( \frac{1}{\sqrt{n}} \|Az\|_2 \) has expectation \( \sqrt{d/n} \).
3. A more geometric way to come to this last conclusion is to notice that projecting on a uniform random subspace of dimension $d$ is equivalent to performing a uniform rotation of $z$ first and then projecting the resulting vector on the first $d$ dimensions. In other words, $\frac{1}{\sqrt{n}} \|Az\|_2$ is approximately distributed as the norm of the first $d$ components of a uniform unit vector in $\mathbb{R}^n$.

To analyze this quantity, observe that a vector in $\mathbb{R}^n$ whose components are independent $\mathcal{N}(0,1)$, when divided by its norm, produces a uniform vector in $\mathbb{R}^n$. When $d$ is large, the norm of that vector is therefore a ratio whose numerator is concentrated around $\sqrt{d}$ and whose denominator is concentrated around $\sqrt{n}$.

4. Either way, $\|Lz\|_2 = \sqrt{\frac{n}{d}} \times \frac{1}{\sqrt{n}} \|Az\|_2$ can be expected to be close to 1.

The Johnson-Lindenstrauss lemma makes it possible to solve certain computational problems, e.g., finding the nearest point to a query, more efficiently by working in a smaller dimension. We discuss a different type of application next.

**Restricted isometries** In the compressed sensing problem, one seeks to recover a signal $x \in \mathbb{R}^n$ from a small number of measurements $(Lx)_i$, $i = 1, \ldots, d$. In complete generality, one needs $n$ such measurements to recover any $x \in \mathbb{R}^n$ as the sensing matrix $L$ must be invertible (or, more precisely, injective). However, by imposing extra structure on the signal, much better results can be obtained. We consider the case of sparse signals.

**Definition 2.64** (Sparse vectors). We say that a vector $z \in \mathbb{R}^n$ is $k$-sparse if it has at most $k$ non-zero entries. We let $\mathcal{S}_n^k$ be the set of $k$-sparse vectors in $\mathbb{R}^n$. Note that $\mathcal{S}_n^k$ is a union of $\binom{n}{k}$ linear subspaces, one for each support of the nonzero entries.

To solve the compressed sensing problem in this case, it suffices to find a sensing matrix $L$ satisfying that all subsets of $2k$ columns are linearly independent. Indeed, if $x, x' \in \mathcal{S}_n^k$, then $x - x'$ has at most $2k$ nonzero entries. Hence, in order to have $L(x - x') = 0$, it must be that $x - x' = 0$ under the previous condition on $L$. That implies the required injectivity. The implication goes in the other direction as well. Observe for instance that the matrix used in the Johnson-Lindenstrauss lemma satisfies this property as long as $d \geq 2k$: because of the continuous density of its entries, the probability that $2k$ of its columns are linearly dependent is 0 when $d \geq 2k$. For practical applications, however, other requirements must be met: computational efficiency and robustness to measurement noise as well as to the sparsity assumption. We discuss the first two issues (and see Remark 2.68 for the last one which can be dealt with along the same lines).

The following definition will play a key role.
Definition 2.65 (Restricted isometry property). A \( d \times n \) linear mapping \( L \) satisfies the \((k, \theta)\)-restricted isometry property (RIP) if for all \( z \in S^n_k \)

\[
(1 - \theta)\|z\|_2 \leq \|Lz\|_2 \leq (1 + \theta)\|z\|_2.
\]

We say that \( L \) is \((k, \theta)\)-RIP.

Given a \((k, \theta)\)-RIP matrix \( L \), can we recover \( z \in S^n_k \) from \( Lz \)? And how small can \( d \) be? The next two claims answer these questions.

Claim 2.66. Let \( L \) be \((10k, 1/3)\)-RIP. Then:

1. (Sparse case) For any \( x \in S^n_k \), the unique solution to the following minimization problem

\[
\min_{z \in \mathbb{R}^n} \|z\|_1 \text{ subject to } Lz = Lx,
\]

is \( z^* = x \).

2. (Almost sparse case) For any \( x \in \mathbb{R}^n \), the solution to (2.52) satisfies \( \|z^* - x\|_2 = O(\eta(x)/\sqrt{k}) \), where \( \eta(x) := \min_{x' \in S^n_k} \|x - x'\|_1 \).

Claim 2.67. Let \( A \) be a \( d \times n \) matrix whose entries are i.i.d. \( N(0, 1) \) and let \( L = \frac{1}{\sqrt{d}}A \). There is a constant \( 0 < C < +\infty \) such that if \( d \geq Ck \log n \) then \( L \) is \((k, 1/3)\)-RIP with probability at least \( 1 - 1/n \).

Roughly speaking, a restricted isometry preserves enough of the metric structure of \( S^n_k \) to be invertible on its image. The purpose of the \( \ell^1 \) minimization in (2.52) is to promote sparsity. See Figure 2.7 for an illustration. It may seem that a more natural approach would be to minimize the number of non-zero entries in \( z \). However the advantage of \( \ell^1 \) minimization is that it can be formulated as a linear program, i.e., the minimization of a linear objective subject to linear inequalities. This permits much faster computation of the solution using standard techniques. See Exercise 2.10.

In Claim 2.67, note that \( d \) is much smaller than \( n \) and not far from the \( 2k \) bound we derived above. Note also that Claim 2.67 does not follow immediately from the Johnson-Lindenstrauss lemma. Indeed that lemma shows that a matrix with i.i.d. Gaussian entries is an approximate isometry on a finite set of points. Here we need a linear mapping that is an approximate isometry for all vectors in \( S^n_k \). The proof of this stronger statement uses an \( \varepsilon \)-net argument.
Figure 2.7: Because $\ell^1$ balls (in orange) have corners, minimizing the $\ell^1$ norm over a linear subspace (in green) tends to produce sparse solutions.

**$\varepsilon$-net argument**  We start with the proof of Claim 2.67. For a subset of indices $J \subseteq [n]$ and a vector $y \in \mathbb{R}^n$, we let $y_J$ be the vector $y$ restricted to the entries in $J$, i.e., the sub-vector $(y_j)_{j \in J}$. Fix a subset of indices $I \subseteq [n]$ of size $k$. As we mentioned above, the Johnson-Lindenstrauss lemma only applies to a finite collection of vectors. However we need the RIP condition to hold for all $z \in \mathbb{R}^n$ with non-zero entries in $I$ (and all such $I$). The way out is to use an $\varepsilon$-net argument, as described in Section 2.4.5. Notice that, for $z \neq 0$, the function $\|Lz\|_2/\|z\|_2$

1. does not depend on the norm of $z$, so that we can restrict ourselves to the compact set $\partial B_I := \{z : z_{[n]} \setminus I = 0, \|z\|_2 = 1\}$, and

2. is continuous on $\partial B_I$, so that it suffices to construct a fine enough covering of $\partial B_I$ by a finite collection of balls and apply the Johnson-Lindenstrauss lemma to the centers of these balls.

**Proof of Claim 2.67.** Let $I \subseteq [n]$ be a subset of indices of size $k$. There are $\binom{n}{k} \leq n^k = \exp(k \log n)$ such subsets and we denote their collection by $\binom{[n]}{k}$. We let $N_I$ be an $\varepsilon$-net of $\partial B_I$. The proof strategy is to apply the Johnson-Lindenstrauss lemma to each $N_I, I \in \binom{[n]}{k}$, and use a continuity argument to extend the RIP property to all $k$-sparse vectors.
- **Bounding Lipschitz constant.** We first choose an appropriate value for $\varepsilon$. Let $A^*$ be the largest entry of $A$ in absolute value. For all $y \in N_I$ and all $z \in \partial B_I$ within distance $\varepsilon$ of $y$, by the triangle inequality, we have $\|Lz\|_2 \leq \|Ly\|_2 + \|L(z - y)\|_2$ and $\|Lz\|_2 \geq \|Ly\|_2 - \|L(z - y)\|_2$. Moreover

$$\|L(z - y)\|_2^2 = \sum_{i=1}^{d} \left( \sum_{j=1}^{n} L_{ij}(z_j - y_j) \right)^2 \leq \sum_{i=1}^{d} \left( \sum_{j=1}^{n} L_{ij}^2 \right) \left( \sum_{j=1}^{n} (z_j - y_j)^2 \right) \leq \|z - y\|_2^2 \cdot dn \left( \frac{1}{\sqrt{d}} A^* \right)^2 \leq (\varepsilon A^*)^2 n,$$

where we used Cauchy-Schwarz. It remains to bound $A^*$. For this we use the Chernoff-Cramér bound for Gaussians in (2.27) which implies

$$\mathbb{P} \left[ \exists i, j, |A_{ij}| \geq C \sqrt{\log n} \right] \leq n^2 e^{-\left( C \sqrt{\log n} \right)^2 / 2} \leq \frac{1}{2n}, \quad (2.53)$$

for a $C > 0$ large enough. Hence we take

$$\varepsilon = \frac{1}{C \sqrt{6n \log n}},$$

and, assuming that the event in (2.53) does not hold, we get

$$\|Lz\|_2 - \|Ly\|_2 \leq \frac{1}{6}. \quad (2.54)$$

- **Applying Johnson-Lindenstrauss to $\varepsilon$-net.** By Claim 2.55,

$$|\{N_I\}_I| \leq n^k \left( \frac{3}{\varepsilon} \right)^k \leq \exp(C' k \log n),$$

for some $C' > 0$. Apply the Johnson-Lindenstrauss lemma to $\{N_I\}_I$ with $\theta = 1/6$, $\delta = \frac{1}{2n}$, and

$$d = \frac{8}{3} \theta^{-2} \left( \log |\{N_I\}_I| + \frac{1}{2} \log(2n) \right) = \Theta(k \log n).$$

Then

$$\frac{5}{6} = \frac{5}{6} \|y\|_2 \leq \|Ly\|_2 \leq \frac{7}{6} \|y\|_2 = \frac{7}{6}, \quad \forall I, \forall y \in N_I. \quad (2.55)$$
Assuming (2.54) and (2.55) hold, an event of probability at least \(1 - 2(1/2n) = 1 - 1/n\), we finally get
\[
\frac{2}{3} \leq \|Lz\|_2 \leq \frac{4}{3}, \quad \forall I, \forall z \in \partial B_I.
\]
That concludes the proof. \(\blacksquare\)

**\(\ell^1\) minimization** It remains to prove Claim 2.66.

**Proof of Claim 2.66.** We only prove the sparse case \(x \in \mathcal{S}_k^n\). For the almost sparse case, see Exercise 2.11. Let \(z^*\) be a solution to (2.52) and note that such a solution exists because \(z = x\) satisfies the constraint in (2.52). W.l.o.g. assume that only the first \(k\) entries of \(x\) are non-zero, i.e., \(x_{[n]\setminus[k]} = 0\). Moreover order the remaining entries of \(x\) so that the residual \(r = z^* - x\) is so that the entries \(r_{[n]\setminus[k]}\) are non-increasing in absolute value. Our goal is to show that \(\|r\|_2 = 0\).

In order to leverage the RIP condition, we break up the vector \(r\) into \(9k\)-long sub-vectors. Let
\[
I_0 = [k], \quad I_i = \{9(i - 1) + 1)k + 1, \ldots, (9i + 1)k\}, \forall i \geq 1,
\]
and let \(I_{01} = I_0 \cup I_1, \bar{I}_i = \bigcup_{j>i} I_j\) and \(\bar{I}_{01} = \bar{I}_1\).

We first use the optimality of \(z^*\). Note that \(x_{I_0} = 0\) implies that
\[
\|z^*\|_1 = \|z^*_{I_0}\|_1 + \|z^*_{\bar{I}_0}\|_1 = \|z^*_{I_0}\|_1 + \|r_{I_0}\|_1,
\]
and
\[
\|x\|_1 = \|x_{I_0}\|_1 \leq \|z^*_{I_0}\|_1 + \|r_{I_0}\|_1,
\]
by the triangle inequality. Since \(\|z^*\|_1 \leq \|x\|_1\) we then have
\[
\|r_{I_0}\|_1 \leq \|r_{I_0}\|_1. \quad (2.56)
\]

On the other hand, the RIP condition gives a similar inequality in the other direction. Indeed notice that \(Lr = 0\) or \(Lr_{I_{01}} = -\sum_{i \geq 2} Lr_{I_i}\) by the constraint in (2.52). Then, by the RIP condition and the triangle inequality, we have that
\[
\frac{2}{3} \|r_{I_{01}}\|_2 \leq \|Lr_{I_{01}}\|_2 \leq \sum_{i \geq 2} \|Lr_{I_i}\|_2 \leq \frac{4}{3} \sum_{i \geq 2} \|r_{I_i}\|_2. \quad (2.57)
\]

We note that by the ordering of the entries of \(x\)
\[
\|r_{I_{i+1}}\|_2^2 \leq 9k \left(\frac{\|r_{I_i}\|_1}{9k}\right)^2, \quad (2.58)
\]
where we bounded $r_{I,i+1}$ entrywise by the expression in parenthesis. Combining (2.56) and (2.58), and using that $\|r_{I_0}\|_1 \leq \sqrt{k}\|r_{I_0}\|_2$ by Cauchy-Schwarz, we have

$$\sum_{i \geq 2} \|r_{I_i}\|_2 \leq \sum_{j \geq 1} \frac{\|r_{I_j}\|_1}{\sqrt{9k}} \leq \frac{\|r_{I_0}\|_1}{3\sqrt{k}} \leq \frac{\|r_{I_0}\|_2}{3} \leq \frac{\|r_{I_0}\|_2}{3}.$$  

Plugging this back into (2.57) gives

$$\|r_{I_{01}}\|_2 \leq 2\sum_{i \geq 2} \|r_{I_i}\|_2 \leq \frac{2}{3} \|r_{I_{01}}\|_2,$$

which implies $r_{I_{01}} = 0$. In particular $r_{I_0} = 0$ and, by (2.56), $r_{I_0} = 0$ as well. We have shown that $r = 0$ or $z^* = x$.

\[\Box\]

Remark 2.68. Claim 2.66 can be extended to noisy measurements (using a slight modification of (2.52)). See [CRT06b].

2.4.7 Markov chains: Varopoulos-Carne and application to mixing times

If $(S_t)$ is simple random walk on $\mathbb{Z}$, then Lemma 2.33 implies that, for any $x, y \in \mathbb{Z},$

$$P^t(x, y) \leq e^{-|x-y|^2/2t}, \quad (2.59)$$

where $P$ is the transition matrix of $(S_t)$. Interestingly a bound similar to (2.59) holds for any reversible Markov chain. And Lemma 2.33 plays an unexpected role in its proof. An application to mixing times is discussed below.

Varopoulos-Carne Our main bound is the following.

Theorem 2.69 (Varopoulos-Carne). Let $P$ be the transition matrix of an irreducible Markov chain $(X_t)$ on the countable state space $V$. Assume further that $P$ is reversible with respect to the stationary measure $\pi$ and that the corresponding network $\mathcal{N}$ is locally finite. Then the following hold

$$\forall x, y \in V, \forall t \in \mathbb{N}, \quad P^t(x, y) \leq 2\sqrt{\frac{\pi(y)}{\pi(x)}}e^{-\rho(x,y)^2/2t},$$

where $\rho(x,y)$ is the graph distance between $x$ and $y$ on $\mathcal{N}$.  

78
As a sanity check before proving the theorem, note that if the chain is aperiodic and \( \pi \) is a stationary distribution then

\[
P^t(x, y) \to \pi(y) \leq 2 \sqrt{\frac{\pi(y)}{\pi(x)}}, \quad \text{as } t \to +\infty,
\]

since \( \pi(x), \pi(y) \leq 1 \).

**Proof of Theorem 2.69.** The idea of the proof is to show that

\[
P^t(x, y) \leq 2 \sqrt{\frac{\pi(y)}{\pi(x)}} \mathbb{P}[S_t \geq \rho(x, y)],
\]

where \( S_t \) is simple random walk on \( \mathbb{Z} \) started at 0, and use Lemma 2.33.

By assumption only a finite number of states can be reached by time \( t \). Hence we can reduce the problem to a finite state space. More precisely, let \( \tilde{V} = \{ z \in V : \rho(x, z) \leq t \} \) and

\[
\tilde{P}(z, w) = \begin{cases} P(z, w), & \text{if } u \neq v \\ P(z, z) + P(z, V - \tilde{V}), & \text{otherwise.} \end{cases}
\]

By construction \( \tilde{P} \) is reversible with respect to \( \tilde{\pi} = \pi/\pi(\tilde{V}) \) on \( \tilde{V} \). Because in time \( t \) one never reaches a state \( z \) where \( P(z, V - \tilde{V}) > 0 \), by Chapman-Kolmogorov and using the fact that \( \tilde{\pi}(y)/\tilde{\pi}(x) = \pi(y)/\pi(x) \), it suffices to prove the result for \( \tilde{P} \). Hence we assume without loss of generality that \( V \) is finite with \( |V| = n \).

To relate \( (X_t) \) to simple random walk on \( \mathbb{Z} \), we use a special representation of \( P^t \) based on Chebyshev polynomials. For \( \xi = \cos \theta \in [-1, 1] \), \( T_k(\xi) = \cos k\theta \) is a **Chebyshev polynomial of the first kind**. Note that \( |T_k(\xi)| \leq 1 \) on \( [-1, 1] \). The classical trigonometric identity (to see this, write it in complex form)

\[
\cos((k + 1)\theta) + \cos((k - 1)\theta) = 2 \cos \theta \cos(k\theta),
\]

implies the recursion

\[
T_{k+1}(\xi) + T_{k-1}(\xi) = 2\xi T_k(\xi),
\]

which in turn implies that \( T_k \) is indeed a polynomial. It has degree \( k \) from induction and the fact that \( T_0(\xi) = 1 \) and \( T_1(\xi) = \xi \). The connection to simple random walk on \( \mathbb{Z} \) comes from the following somewhat miraculous representation (which does not rely on reversibility). Let \( T_k(P) \) denote the polynomial \( T_k \) evaluated at \( P \) as a matrix polynomial.
Lemma 2.70.

\[ P^t = \sum_{k=-t}^{t} \mathbb{P}[S_t = k] T_{|k|}(P). \]

Proof. This essentially follows from the binomial theorem. It suffices to prove

\[ \xi^t = \sum_{k=-t}^{t} \mathbb{P}[S_t = k] T_{|k|}(\xi), \]

as an identity of polynomials. Since \( \mathbb{P}[S_t = k] = 2^{-t}(\binom{t}{k}) \) for \(|k| \leq t\) of the same parity as \(t\) (and 0 otherwise), this follows immediately from the expansion

\[ \xi^t = \left( \frac{e^{i\theta} + e^{-i\theta}}{2} \right)^t = \sum_{\ell=0}^{t} 2^{-t}(\binom{t}{\ell}) (e^{i\theta})^\ell (e^{-i\theta})^{t-\ell} = \sum_{k=-t}^{t} \mathbb{P}[S_t = k] e^{ik\theta}, \]

where we used that the probability that

\[ S_t = -t + 2\ell = (+1)\ell + (-1)(t - \ell), \]

is the probability of making \(\ell\) steps to the right and \(t - \ell\) steps to the left. (Put differently, \((\cos \theta)^t\) is the characteristic function of \(S_t\).) Take real parts and use \(\cos(k\theta) = \cos(-k\theta)\).

We bound \(T_k(P)(x, y)\) as follows.

Lemma 2.71.

\[ T_k(P)(x, y) = 0, \quad \forall k < \rho(x, y). \]

and

\[ T_k(P)(x, y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}}, \quad \forall k \geq \rho(x, y). \]

Proof. Note that \(T_k(P)(x, y) = 0\) when \(k < \rho(x, y)\) because \(T_k(P)(x, y)\) is a function of the entries \(P^\ell(x, y)\) for \(\ell \leq k\).

Let \(f_1, \ldots, f_n\) be a right eigenvector decomposition of \(P\) orthonormal with respect to the inner product

\[ \langle f, g \rangle_\pi = \sum_{x \in V} \pi(x)f(x)g(x), \]

with eigenvalues \(\lambda_1, \ldots, \lambda_n \in [-1, 1]\). Such a decomposition exists by the reversibility of \(P\). See Lemma A.7. Then \(f_1, \ldots, f_n\) is also an eigenvector decomposition of the polynomial \(T_k(P)\) with eigenvalues \(T_k(\lambda_1), \ldots, T_k(\lambda_n) \in [-1, 1]\).
by the definition of Chebyshev polynomials. By decomposing any $f = \sum_{i=1}^{n} \alpha_i f_i$ according to this eigenbasis, that implies that

$$\|T_k(P)f\|_\pi^2 = \sum_{i=1}^{n} \alpha_i^2 T_k(\lambda_i)^2 \langle f_i, f_i \rangle_\pi \leq \sum_{i=1}^{n} \alpha_i^2 \langle f_i, f_i \rangle_\pi = \|f\|_\pi^2,$$

(2.60)

where we used the norm $\|f\|_\pi$ associated with the inner product $\langle \cdot, \cdot \rangle_\pi$, the orthonormality of the eigenvector basis under this inner product, and the fact that $T_k(\lambda_i)^2 \in [0, 1]$.

Let $\delta_x$ denote the point mass at $x$. By Cauchy-Schwarz and (2.60),

$$T_k(P)(x,y) = \langle \delta_x, T_k(P) \delta_y \rangle_\pi \leq \frac{\|\delta_x\|_\pi \|\delta_y\|_\pi}{\pi(x)} = \sqrt{\frac{\pi(x)}{\pi(x)}} \sqrt{\frac{\pi(y)}{\pi(y)}},$$

for $k \geq \rho(x,y)$. ■

Combining the two lemmas gives the result. ■

**Remark 2.72.** The local finiteness assumption is made for simplicity only. The result holds for any countable-space, reversible chain. See [LP, Section 13.2].

### Lower bound on mixing

Let $(X_t)$ be an irreducible, aperiodic Markov chain with finite state space $V$ and stationary distribution $\pi$. Recall that, for a fixed $0 < \varepsilon < 1/2$, the mixing time is

$$t_{\text{mix}}(\varepsilon) = \min \{ t : d(t) \leq \varepsilon \},$$

where

$$d(t) = \max_{x \in V} \| P^t(x, \cdot) - \pi \|_{\text{TV}}.$$

It is intuitively clear that $t_{\text{mix}}(\varepsilon)$ is at least of the order of the “diameter” of the transition graph of $P$. For $x, y \in V$, let $\rho(x, y)$ be the graph distance between $x$ and $y$ on the undirected version of the transition graph, i.e., ignoring the orientation of the edges. With this definition, a shortest directed path from $x$ to $y$ contains at least $\rho(x, y)$ edges. Here we define the diameter of the transition graph as $\Delta := \max_{x, y \in V} \rho(x, y)$. Let $x_0, y_0$ be a pair of vertices achieving the diameter. Then we claim that $P^{\lfloor (\Delta-1)/2 \rfloor}(x_0, \cdot)$ and $P^{\lfloor (\Delta-1)/2 \rfloor}(y_0, \cdot)$ are supported on disjoint sets. To see this let

$$A = \{ z \in S : \rho(x_0, z) < \rho(y_0, z) \}.$$

See Figure 2.4.7. By the triangle inequality for $\rho$, any $z$ such that $\rho(x_0, z) \leq \lfloor (\Delta-1)/2 \rfloor$ is in $A$, otherwise we would have $\rho(y_0, z) \leq \rho(x_0, z) \leq \lfloor (\Delta-1)/2 \rfloor$
and hence $\rho(x_0, y_0) \leq \rho(x_0, z) + \rho(y_0, z) \leq 2\lfloor(\Delta - 1)/2\rfloor < \Delta$. Similarly, if $\rho(y_0, z) \leq \lfloor(\Delta - 1)/2\rfloor$, then $z \in A^c$. By the triangle inequality for the total variation distance,

$$
\begin{align*}
\text{d}(\lfloor(\Delta - 1)/2\rfloor) & \geq \frac{1}{2} \left\| P^{\lfloor(\Delta - 1)/2\rfloor}(x_0, \cdot) - P^{\lfloor(\Delta - 1)/2\rfloor}(y_0, \cdot) \right\|_{TV} \\
& \geq \frac{1}{2} \left\{ P^{\lfloor(\Delta - 1)/2\rfloor}(x_0, A) - P^{\lfloor(\Delta - 1)/2\rfloor}(y_0, A) \right\} \\
& = \frac{1}{2} \{1 - 0\} = \frac{1}{2},
\end{align*}
$$

so that:

**Claim 2.73.**

\[ t_{\text{mix}}(\varepsilon) \geq \frac{\Delta}{2}. \]

This bound is often far from the truth. Consider for instance simple random walk on a cycle of size $n$. The diameter is $\Delta = n/2$. But Lemma 2.33 suggests that it takes time of order $\Delta^2$ to even reach the antipode of the starting point, let alone achieve stationarity.

More generally, when $P$ is reversible, we use the Varopoulos-Carne bound to show that the mixing time does indeed scale at least as the square of the diameter.
Assume that $P$ is reversible with respect to $\pi$ and has diameter $\Delta$. Letting $n = |V|$ and $\pi_{\min} = \min_{x \in V} \pi(x)$, we have the following.

**Claim 2.74.**

$$t_{\text{mix}}(\varepsilon) \geq \frac{\Delta^2}{12 \log n + 4|\log \pi_{\min}|}, \quad \text{if } n \geq \frac{16}{(1 - 2\varepsilon)^2}.$$  

**Proof.** The proof is based on the same argument we used to derive our first diameter-based bound, except that the Varopoulos-Carne bound gives a better dependence on the diameter. Namely, let $x_0$, $y_0$, and $A$ be as above. By the Varopoulos-Carne bound,

$$P^t(x_0, A^c) = \sum_{z \in A^c} P^t(x_0, z) \leq \sum_{z \in A^c} 2 \sqrt{\frac{\pi(z)}{\pi(x_0)}} e^{-\frac{\rho^2(x_0, z)}{4t}} \leq 2n\pi_{\min}^{-1/2} e^{-\frac{\Delta^2}{8t}},$$

where we used that $|A^c| \leq n$ and $\rho(x_0, z) \geq \frac{\Delta}{2}$ for $z \in A^c$. For any

$$t < \frac{\Delta^2}{12 \log n + 4|\log \pi_{\min}|},$$

we get that

$$P^t(x_0, A^c) \leq 2n\pi_{\min}^{-1/2} \exp\left(-\frac{3 \log n + |\log \pi_{\min}|}{2}\right) = \frac{2}{\sqrt{n}}.$$

Similarly, $P^t(y_0, A) \leq \frac{2}{\sqrt{n}}$ so that arguing as in (2.61)

$$d((\Delta - 1)/2) \geq \frac{1}{2} \left(1 - \frac{2}{\sqrt{n}} - \frac{2}{\sqrt{n}}\right) = \frac{1}{2} - \frac{2}{\sqrt{n}} \geq \varepsilon,$$

for $n$ as in the statement. 

**Remark 2.75.** The dependence on $\Delta$ and $\pi_{\min}$ in Claim 2.74 cannot be improved. See [LP, Section 13.3].

### 2.4.8 Data science: classification, empirical risk minimization and VC dimension

In binary classification, one is given samples $S_n = \{(X_i, C(X_i))\}_{i=1}^n$ where $X_i \in \mathbb{R}^d$ is a feature vector and $C(X_i) \in \{0, 1\}$ is a label. The feature vectors are assumed to be independent samples from an unknown probability measure $\mu$ and $C : \mathbb{R}^d \to \{0, 1\}$ is a measurable Boolean function. For instance, the feature vector
might be an image (encoded as a vector) and the label might indicate cat (label 0) or dog (label 1). Our goal is “learn” the concept $C$. More precisely, we seek to construct a hypothesis $h : \mathbb{R}^d \to \{0, 1\}$ that is a good approximation to $C$ in the sense that it predicts the label well on a fresh sample.

Put differently, we want $h$ to have small risk, or generalization error,

$$R(h) = \mathbb{P}[h(X) \neq C(X)]$$

where $X \sim \mu$. Because we only have access to the distribution $\mu$ through the samples, it is natural to estimate the risk of the hypothesis $h$ using the samples as

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} 1\{h(X_i) \neq C(X_i)\},$$

which is called the empirical risk. Indeed observe that $\mathbb{E}R_n(h) = R(h)$ and, by the law of large numbers, $R_n(h) \to R(h)$ almost surely as $n \to +\infty$. Ignoring computational considerations, one can then formally define an empirical risk minimizer

$$h^* \in \text{ERM}_H(\mathcal{S}_n) = \{h \in \mathcal{H} : R_n(h') \leq R_n(h'), \forall h' \in \mathcal{H}\}$$

where $\mathcal{H}$, the hypothesis class, is a collection of Boolean functions over $\mathbb{R}^d$. We assume further that $h^*$ can be defined as a measurable function of the samples.

**Overfitting** Why restrict the hypothesis class? The following theorem shows that minimizing the empirical risk over all Boolean functions makes it impossible to achieve an arbitrarily small risk. Intuitively, considering too rich a class of functions, that is, functions that too intricately model the data, leads to overfitting: the learned hypothesis will fit the sampled data, but it may not generalize well to unseen examples. A learner $A$ is a map from samples to measurable Boolean functions over $\mathbb{R}^d$, that is, for any $n$ and any $\mathcal{S}_n = (\mathbb{R}^d \times \{0, 1\})^n$, the learner outputs a function $A(\cdot, \mathcal{S}_n) : \mathbb{R}^d \to \{0, 1\}$.

**Theorem 2.76 (No Free Lunch).** For any learner $A$ and any $\mathcal{X} \subseteq \mathbb{R}^d$ with $|\mathcal{X}| = 2m > 4$, there exist a concept $C : \mathcal{X} \to \{0, 1\}$ and a distribution $\mu$ over $\mathcal{X}$ such that

$$\mathbb{P}[R(A(\cdot, \mathcal{S}_m)) \geq 1/8] \geq 1/8,$$

where $\mathcal{S}_m = \{(X_i, C(X_i))\}_{i=1}^{m}$ with independent $X_i \sim \mu$. 

84
**Proof.** We let \( \mu \) be uniform over \( \mathcal{X} \). To prove the existence of a concept satisfying (2.62), we use the probabilistic method (Section 2.2.1) and pick \( C \) at random. For each \( x \in \mathcal{X} \), we set \( C(x) = Y_x \) where the \( Y_x \)'s are i.i.d. uniform in \{0, 1\}.

We first bound \( \mathbb{E}[R(A(\cdot, S_m))] \), where the expectation runs over both random labels \( \{Y_x\}_{x \in \mathcal{X}} \) and the samples \( S_m = \{(X_i, C(X_i))\}_{i=1}^m \). For an additional independent sample \( X \sim \mu \), we will need the event that the learner, given samples \( S_m \), makes an incorrect prediction on

\[
B = \{A(X, S_m) \neq Y_X\},
\]

and the event that \( X \) is observed in the samples \( S_m \)

\[
O = \{X \in \{X_1, \ldots, X_m\}\}.
\]

By the law of total expectation,

\[
\mathbb{E}[R(A(\cdot, S_m))] = \mathbb{P}[B] = \mathbb{E}[\mathbb{P}[B | S_m]] = \mathbb{E}[\mathbb{P}[B | O, S_m] \mathbb{P}[O | S_m] + \mathbb{P}[B | O^c, S_m] \mathbb{P}[O^c | S_m]] \\
\geq \mathbb{E}[\mathbb{P}[B | O^c, S_m] \mathbb{P}[O^c | S_m]] \geq \frac{1}{2} \times \frac{1}{2},
\]

where we used that:

- \( \mathbb{P}[O^c | S_m] \geq 1/2 \) because \( |\mathcal{X}| = 2m \) and \( \mu \) is uniform, and;

- \( \mathbb{P}[B | O^c, S_m] = 1/2 \) because for any \( x \notin \{X_1, \ldots, X_m\} \) the prediction \( A(x, S_m) \in \{0, 1\} \) is independent of \( Y_x \) and the latter is uniform.

Conditioning over the concept, we have proved that

\[
\mathbb{E}[\mathbb{E}[R(A(\cdot, S_m)) | \{Y_x\}_{x \in \mathcal{X}}]] \geq \frac{1}{4}.
\]

Hence, by the first moment principle (Theorem 2.6),

\[
\mathbb{P}[\mathbb{E}[R(A(\cdot, S_m)) | \{Y_x\}_{x \in \mathcal{X}}] \geq 1/4] > 0,
\]

where the probability is taken over \( \{Y_x\}_{x \in \mathcal{X}} \). That is, there exists a choice \( \{y_x\}_{x \in \mathcal{X}} \in \{0, 1\}^{\mathcal{X}} \) such that

\[
\mathbb{E}[R(A(\cdot, S_m)) | \{Y_x = y_x\}_{x \in \mathcal{X}}] \geq 1/4. \tag{2.63}
\]
Finally, to prove (2.62), we use a variation on Markov’s inequality (Theorem 2.1) for \([0, 1]\)-valued random variables. If \(Z \in [0, 1]\) is a random variable with \(\mathbb{E}[Z] = \mu\) and \(\alpha \in [0, 1]\), then
\[
\mathbb{E}[Z] \leq \alpha \times \mathbb{P}[Z < \alpha] + 1 \times \mathbb{P}[Z \geq \alpha] \leq \mathbb{P}[Z \geq \alpha] + \alpha.
\]
Taking \(\alpha = \mu/2\) gives
\[
\mathbb{P}[Z \geq \mu/2] \geq \mu/2.
\]
Going back to (2.63), we obtain
\[
\mathbb{P} \left[ R(A(\cdot, S_m)) \geq 1 \left\| \{ Y_x = y_x \} \right\|_{X} \right] \geq \frac{1}{8},
\]
establishing the claim.

The gist of the proof is intuitive. In essence, if the target concept is arbitrary and we only get to see half of the possible instances, then we have learned nothing about the other half and cannot expect low generalization error.

So, as we mentioned before, one way out is to limit the complexity of the hypotheses. For instance, we could restrict ourselves to half-spaces
\[
\mathcal{H}_H = \left\{ h(x) = 1 \{ x^T u \geq \alpha \} : u \in \mathbb{R}^d, \alpha \in \mathbb{R} \right\},
\]
or axis-aligned boxes
\[
\mathcal{H}_B = \left\{ h(x) = 1 \left\{ x_i \in [\alpha_i, \beta_i], \forall i \right\} : -\infty \leq \alpha_i \leq \beta_i \leq \infty, \forall i \right\}.
\]
In order for the empirical risk minimizer \(h^*\) to have a generalization error close to the best achievable error, we need the empirical risk of the learned hypothesis \(R_n(h^*)\) to be close to its expectation \(R(h^*)\), which is guaranteed by the law of large numbers for sufficiently large \(n\). But that is not enough, we also need that same property to hold for all hypotheses in \(\mathcal{H}\) simultaneously. Otherwise we could be fooled by a poorly performing hypothesis with unusually good empirical risk on the samples. The hypothesis class is typically infinite and, therefore, controlling empirical risk deviations from their expectations uniformly over \(\mathcal{H}\) is not straightforward.

**Uniform deviations** Our goal in this section is to show how to bound
\[
\mathbb{E} \left[ \sup_{h \in \mathcal{H}} R_n(h) - R(h) \right] = \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h, X) - \mathbb{E}[\ell(h, X)] \right] \tag{2.64}
\]
in terms of a measure of complexity of the class \( \mathcal{H} \), where we defined the loss 
\[ \ell(h, x) = 1\{h(x) \neq C(x)\} \] 
to simplify the notation. We assume that \( \mathcal{H} \) is countable. Observe for instance that, for \( \mathcal{H}_H \) and \( \mathcal{H}_B \), nothing is lost by assuming that the parameters defining the hypotheses are rational-valued.

Controlling deviations uniformly over \( \mathcal{H} \) as in (2.64) allows one to provide guarantees on the empirical risk minimizer. Indeed, for any \( h' \in \mathcal{H} \),

\[
R(h^*) \leq R_n(h^*) + \sup_{h \in \mathcal{H}} (R(h) - R_n(h)) 
\]

\[
\leq R_n(h') + \sup_{h \in \mathcal{H}} (R(h) - R_n(h)) 
\]

\[
\leq R(h') + \sup_{h \in \mathcal{H}} (R_n(h) - R(h)) + \sup_{h \in \mathcal{H}} (R(h) - R_n(h)), 
\]

where, on the second line, we used the definition of the empirical risk minimizer. Taking an infimum over \( h' \) and then taking an expectation gives

\[
\mathbb{E}[R(h^*)] \leq \inf_{h \in \mathcal{H}} R(h) + \mathbb{E} \left[ \sup_{h \in \mathcal{H}} (R_n(h) - R(h)) \right] + \mathbb{E} \left[ \sup_{h \in \mathcal{H}} (R(h) - R_n(h)) \right].
\]

Observe that the supremum in (2.64) is inside the expectation and that the random variables \( R_n(h) - R(h) \) are highly correlated. Indeed, two similar hypotheses will produce similar predictions. While the absence of independence in some sense makes bounding this type expectation harder, the correlation is ultimately what allows us to tackle infinite classes \( \mathcal{H} \).

We will use the methods of Section 2.4.5. As we did in that section, we assume that there is a countable collection \( \mathcal{H}_0 \subseteq \mathcal{H} \) such that all suprema considered here can be taken over \( \mathcal{H}_0 \). As a first step, we use symmetrization, which we introduced in Section 2.4.3 to give a proof of Hoeffding’s lemma (Lemma 2.42). Let \((\varepsilon_i)_{i=1}^n\) be i.i.d. uniform random variables in \([-1, +1]\) and let \((X'_i)_{i=1}^n\) be an independent
copy of \((X_i)_{i=1}^n\). Then

\[
\mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(h, X_i) - \mathbb{E}[\ell(h, X)] \right]
\]

\[
= \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(h, X_i) - \mathbb{E}[\ell(h, X_i) \mid (X_i)_{i=1}^n] \right]
\]

\[
\leq \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n [\ell(h, X_i) - \ell(h, X_i')] \right]
\]

\[
= \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i [\ell(h, X_i) - \ell(h, X_i')] \right]
\]

\[
\leq \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell(h, X_i) + \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n -\varepsilon_i \ell(h, X_i') \right]
\]

\[
\leq 2 \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \ell(h, X_i) \right],
\]

where, on the third line, we used \(\sup_h \mathbb{E} Y_h \leq \mathbb{E}[\sup_h Y_h]\) and the tower property, on the fourth line, we used that \([\ell(h, X_i) - \ell(h, X_i')]\) is symmetric and independent of \(\varepsilon_i\), and on the fifth line, we used that \(\varepsilon_i\) and \(-\varepsilon_i\) are identically distributed.

1. Define

\[
Z_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \ell(h, X_i).
\]

(2.65)

Our task reduces to bounding

\[
\mathbb{E} \left[ \sup_{h \in \mathcal{H}} Z_n(h) \right].
\]

(2.66)

**VC dimension** We make two observations about \(Z_n(h)\):

1. As a weighted sum of independent random variables in \([-1, 1]\), \(Z_n(h)\) is sub-Gaussian with variance factor 1 by the general Hoeffding inequality (Theorem 2.39) and Hoeffding’s lemma (Lemma 2.42).

2. The random variable \(Z_n(h)\) only depends on the values of \(h\) at a finite number of points, \(X_1, \ldots, X_n\). Hence, while the supremum in (2.66) is over a potentially infinite class of functions \(\mathcal{H}\), it is in effect a supremum over at most \(2^n\) functions, that is, all the possible restrictions of the \(h\)’s to \((X_i)_{i=1}^n\).
A naive application of the maximal inequality in Lemma 2.50, together with the two observations above, unfortunately gives the very weak bound

\[ \mathbb{E} \left[ \sup_{h \in \mathcal{H}} Z_n(h) \right] \leq \sqrt{2 \log 2^n} = \sqrt{2n \log 2} \]

To obtain a better bound, we show that in general the number of distinct restrictions of \( \mathcal{H} \) to \( n \) points can grow much slower than \( 2^n \).

**Definition 2.77 (Shattering).** Let \( \Lambda = \{ \ell_1, \ldots, \ell_n \} \subseteq \mathbb{R}^d \) be a finite set and let \( \mathcal{H} \) be a class of Boolean functions on \( \mathbb{R}^d \). The restriction of \( \mathcal{H} \) to \( \Lambda \) is

\[ \mathcal{H}_\Lambda = \{ (h(\ell_1), \ldots, h(\ell_n)) : h \in \mathcal{H} \} \]

We say that \( \Lambda \) is shattered by \( \mathcal{H} \) if \( |\mathcal{H}_\Lambda| = 2^{|\Lambda|} \), that is, if all Boolean functions over \( \Lambda \) can be obtained by restricting a function in \( \mathcal{H} \) to the points in \( \Lambda \).

**Definition 2.78 (VC dimension).** Let \( \mathcal{H} \) be a class of Boolean functions on \( \mathbb{R}^d \). The **VC dimension** of \( \mathcal{H} \), denoted \( \text{vc}(\mathcal{H}) \), is the maximum cardinality of a set shattered by \( \mathcal{H} \).

We prove the following combinatorial lemma at the end of this section.

**Lemma 2.79 (Sauer’s lemma).** Let \( \mathcal{H} \) be a class of Boolean functions on \( \mathbb{R}^d \). For any finite set \( \Lambda = \{ \ell_1, \ldots, \ell_n \} \subseteq \mathbb{R}^d \),

\[ |\mathcal{H}_\Lambda| \leq \left( \frac{en}{\text{vc}(\mathcal{H})} \right)^{\text{vc}(\mathcal{H})} \]

That is, the number of distinct restrictions of \( \mathcal{H} \) to any \( n \) points grows at most as \( \sim n^{\text{vc}(\mathcal{H})} \).

Returning to \( \mathbb{E}[\sup_{h \in \mathcal{H}} Z_n(h)] \), we get the following inequality using Lemma 2.50.

**Lemma 2.80.** There exists a constant \( 0 < C < +\infty \) such that, for any countable class of measurable Boolean functions \( \mathcal{H} \) over \( \mathbb{R}^d \),

\[ \mathbb{E} \left[ \sup_{h \in \mathcal{H}} R_n(h) - R(h) \right] \leq C \sqrt{\frac{\text{vc}(\mathcal{H}) \log n}{n}}. \tag{2.67} \]

**Proof.** Let \( Z_n(h) \) be as in (2.65) and recall that \( Z_n(h) \in \mathcal{G}(1) \). Since the supremum over \( \mathcal{H} \), when seen as restricted to \( \{X_1, \ldots, X_n\} \), is in fact a supremum over at most \( (en)^{\text{vc}(\mathcal{H})} \) functions by Sauer’s lemma (Lemma 2.79), we have by Lemma 2.50

\[ \mathbb{E} \left[ \sup_{h \in \mathcal{H}} Z_n(h) \right] \leq \sqrt{2 \log (en)^{\text{vc}(\mathcal{H})}}. \]

Symmetrization then proves the claim. \( \blacksquare \)
We give some examples.

**Example 2.81** (VC dimension of half-spaces). Consider the class of half-spaces.

**Claim 2.82.**

\[ \text{vc}(\mathcal{H}_H) = d + 1. \]

We only prove the case \( d = 1 \), where \( \mathcal{H}_H \) reduces to half-lines \((-\infty, \gamma] \) or \([\gamma, +\infty)\). Clearly any set \( \Lambda = \{\ell_1, \ell_2\} \subseteq \mathbb{R} \) with elements is shattered by \( \mathcal{H}_H \). On the other hand, for any \( \Lambda = \{\ell_1, \ell_2, \ell_3\} \) with \( \ell_1 < \ell_2 < \ell_3 \), any half-line containing \( \ell_1 \) and \( \ell_3 \) necessarily includes \( \ell_2 \) as well. Hence no set of size 3 is shattered by \( \mathcal{H}_H \).

For the proof in general dimension \( d \), see e.g. [SSBD14, Section 9.1.3].

**Example 2.83** (VC dimension of boxes). Consider the class of axis-aligned boxes.

**Claim 2.84.**

\[ \text{vc}(\mathcal{H}_B) = 2d. \]

We only prove the case \( d = 2 \), where \( \mathcal{H}_B \) reduces to rectangles. The set \( \Lambda = \{(-1,0), (1,0), (0,-1), (0,1)\} \) is shattered by \( \mathcal{H}_B \). Indeed, the rectangle \([-1,1] \times [-1,1]\) contains \( \Lambda \), with each side of the rectangle containing one the points. Moving any side inward by \( \varepsilon \) removes the corresponding point from the rectangle. Hence, any subset of \( \Lambda \) can be obtained by this procedure.

On the other hand, let \( \Lambda = \{\ell_1, \ldots, \ell_5\} \subseteq \mathbb{R}^2 \) be any set with five distinct points. Consider the rectangle with smallest area containing \( \Lambda \). Then there must be four distinct points in \( \Lambda \), each touching a different side of the rectangle (including degenerate sides). Now observe that any rectangle containing those four points must also contain the fifth one. That proves the claim.

The proof in general dimension \( d \) follows along the same lines.

These two examples also provide insight into Sauer’s lemma. Consider the case of rectangles for instance. Over a collection of \( n \) sample points, a rectangle defines the same \( \{0, 1\} \)-labeling as the minimal-area rectangle containing the same points. Because each side of a minimal-area rectangle must touch at least one point in the sample, there are at most \( n^4 \) such rectangles, and hence there are at most \( n^4 \ll 2^n \) restrictions of \( \mathcal{H}_B \) to these sample points.

**Chaining** It turns out that the \( \sqrt{\log n} \) factor in (2.67) is not optimal. We use chaining (Section 2.4.5) to improve the bound. We return to symmetrization, that is, we consider again the process \( Z_n(h) \) defined in (2.65).
We claim that \( (Z_n(h))_{h \in \mathcal{H}} \) has sub-Gaussian increments under an appropriately defined pseudo-metric. Indeed, conditioning on \( (X_i)_{i=1}^n \), by the general Hoeffding inequality (Theorem 2.39) and Hoeffding’s lemma (Lemma 2.42), we have

\[
Z_n(g) - Z_n(h) = \sum_{i=1}^n \varepsilon_i \frac{\ell(g, X_i) - \ell(h, X_i)}{\sqrt{n}},
\]

is sub-Gaussian increments with variance factor

\[
\sum_{i=1}^n \left( \frac{\ell(g, X_i) - \ell(h, X_i)}{\sqrt{n}} \right)^2 \times 1 = \frac{1}{n} \sum_{i=1}^n [\ell(g, X_i) - \ell(h, X_i)]^2.
\]

Define the pseudometric

\[
\rho_n(g, h) = \left[ \frac{1}{n} \sum_{i=1}^n [\ell(g, X_i) - \ell(h, X_i)]^2 \right]^{1/2} = \left[ \frac{1}{n} \sum_{i=1}^n [g(X_i) - h(X_i)]^2 \right]^{1/2},
\]

which satisfies the triangle inequality since it can be expressed as a Euclidean norm. In fact, it will be useful to recast it in a more general setting. For a probability measure \( \eta \) over \( \mathbb{R}^d \), define

\[
\|g - h\|_{L^2(\eta)} = \int_{\mathbb{R}^d} (f(x) - g(x))^2 \, d\eta(x).
\]

Let \( \mu_n \) be the empirical measure

\[
\mu_n = \mu\{X_i\}_{i=1}^n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i},
\]

(2.68)

where \( \delta_x \) is the probability measure that puts mass 1 on \( x \). Then, we can re-write

\[
\rho_n(g, h) = \|g - h\|_{L^2(\mu_n)}.
\]

Hence we have shown that \( (Z_n(h))_{h \in \mathcal{H}} \) has sub-Gaussian increments with respect to \( \| \cdot \|_{L^2(\mu_n)} \). Note that the pseudometric here is random as it depends on the samples. However, by the law of large numbers, \( \|g - h\|_{L^2(\mu_n)} \) approaches its expectation, \( \|g - h\|_{L^2(\mu)} \), as \( n \to +\infty \).

Applying the discrete Dudley’s inequality (Theorem 2.59), we obtain the following bound.

**Lemma 2.85.** There exists a constant \( 0 < C < +\infty \) such that, for any countable class of measurable Boolean functions \( \mathcal{H} \) over \( \mathbb{R}^d \),

\[
\mathbb{E} \left[ \sup_{h \in \mathcal{H}} R_n(h) - R(h) \right] \leq \frac{C}{\sqrt{n}} \mathbb{E} \left[ \sum_{k=0}^{+\infty} 2^{-k} \sqrt{\log \mathcal{N}(\mathcal{H}, \| \cdot \|_{L^2(\mu_n)}, 2^{-k})} \right],
\]

where \( \mu_n \) is the empirical measure over the samples \( (X_i)_{i=1}^n \).
Proof. Because $\mathcal{H}$ comprises only Boolean functions, it follows that under the pseudometric $\| \cdot \|_{L^2(\mu_n)}$, the diameter is bounded by 1. We apply the discrete Dudley’s inequality (Theorem 2.59) conditioned on $(X_i)_{i=1}^n$. Then we take an expectation over the samples which, together with symmetrization, gives the claim.

Our use of symmetrization above is more intuitive than it may appear at first. The central limit theorem indicates that the fluctuations of centered averages such as

$$(R_n(g) - R(g)) - (R_n(h) - R(h))$$

tend cancel out and that, in the limit, the variance alone characterizes the overall behavior. The $\varepsilon_i$’s in some sense explicitly capture the canceling part of this phenomenon while $\rho_n$ captures the scale of the resulting global fluctuations in the increments.

Bounding the covering numbers using the VC dimension

Our final task is to bound the covering numbers $\mathcal{N}(\mathcal{H}, \| \cdot \|_{L^2(\mu_n)}, 2^{-k})$.

**Theorem 2.86 (Covering numbers via VC dimension).** There exists a constant $0 < C < +\infty$ such that, for any class of measurable Boolean functions $\mathcal{H}$ over $\mathbb{R}^d$, any probability measure $\eta$ over $\mathbb{R}^d$ and any $\varepsilon \in (0, 1)$,

$$\mathcal{N}(\mathcal{H}, \| \cdot \|_{L^2(\eta)}, \varepsilon) \leq \left( \frac{2}{\varepsilon} \right)^{C \text{vc}(\mathcal{H})}.$$

Before proving Theorem 2.86, we derive its implications on uniform deviations. Compare the following bound to Lemma 2.80.

**Lemma 2.87.** There exists a constant $0 < C < +\infty$ such that, for any countable class of measurable Boolean functions $\mathcal{H}$ over $\mathbb{R}^d$,

$$\mathbb{E} \left[ \sup_{h \in \mathcal{H}} R_n(h) - R(h) \right] \leq C \sqrt{\frac{\text{vc}(\mathcal{H})}{n}}.$$
Proof. By Lemma 2.85,

\[ \mathbb{E} \left[ \sup_{h \in \mathcal{H}} R_n(h) - R(h) \right] \leq \frac{C}{\sqrt{n}} \mathbb{E} \left[ \sum_{k=0}^{+\infty} 2^{-k} \sqrt{\log N(\mathcal{H}, \| \cdot \|_{L^2(\mu_n)}, 2^{-k})} \right] \leq C \mathbb{E} \left[ \sum_{k=0}^{+\infty} 2^{-k} \sqrt{\log \left( \frac{2}{2^{-k}} \right)} C' \text{vc}(\mathcal{H}) \right] \leq C \sqrt{\text{vc}(\mathcal{H})} \mathbb{E} \left[ \sum_{k=0}^{+\infty} 2^{-k} \sqrt{k + 1} \sqrt{\log 2} \right] \leq C'' \sqrt{\text{vc}(\mathcal{H})}, \]

for some \( 0 < C'' < +\infty \).

It remains to prove Theorem 2.86.

Proof of Theorem 2.86. Let \( \mathcal{G} = \{ g_1, \ldots, g_N \} \subseteq \mathcal{H} \) be a maximal \( \varepsilon \)-packing of \( \mathcal{H} \) with \( N \geq \mathcal{N}(\mathcal{H}, \| \cdot \|_{L^2(\eta)}, \varepsilon) \), which exists by Lemma 2.53. We use the probabilistic method (Section 2.2) and Hoeffding’s inequality (Theorem 2.40) to show that there exists a small number of points \( \{ x_1, \ldots, x_m \} \) such that \( \mathcal{G} \) is still a good packing when \( \mathcal{H} \) is restricted to the \( x_i \)'s. Then we use Sauer’s lemma (Lemma 2.79) to conclude.

1. Restriction. By construction, the collection \( \mathcal{G} \) satisfies

\[ \| g_i - g_j \|_{L^2(\eta)} > \varepsilon, \quad \forall i \neq j. \]

For an integer \( m \) that we will choose as small as possible below, let \( X = \{ X_1, \ldots, X_m \} \) be i.i.d. samples from \( \eta \) and let \( \eta_X \) be the corresponding empirical measure (as defined in (2.68)). Observe that, for any \( i \neq j \),

\[ \mathbb{E} \left[ \| g_i - g_j \|_{L^2(\eta_X)} \right] = \mathbb{E} \left[ \frac{1}{m} \sum_{k=1}^{m} (g_i(X_k) - g_j(X_k))^2 \right] = \| g_i - g_j \|_{L^2(\eta)}. \]

Moreover \( [g_i(X_k) - g_j(X_k)]^2 \in [0, 1] \). Hence, by Hoeffding’s inequality (Theorem 2.40), there exists a constant \( 0 < C < +\infty \) and an \( m \leq C \varepsilon^{-4} \log N \) such that

\[ \mathbb{P} \left[ \| g_i - g_j \|_{L^2(\eta)} - \| g_i - g_j \|_{L^2(\eta_X)} \geq \frac{3\varepsilon^2}{4} \right] \leq \exp \left( -\frac{2(m \cdot 3\varepsilon^2/4)^2}{m} \right) = \exp \left( -\frac{9}{8} \varepsilon^4 m \right) < \frac{1}{N^2}. \]
That implies that, for this choice of \( m \),

\[
P \left[ \| g_i - g_j \|_{L^2(\mu)}^2 < \frac{\varepsilon}{2}, \forall i \neq j \right] > 0,
\]

where the probability is over the samples. Therefore, there must be a set \( \mathcal{X} = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^d \) such that

\[
\| g_i - g_j \|_{L^2(\mu_X)}^2 > \frac{\varepsilon}{2}, \forall i \neq j.
\] (2.69)

2. **VC bound.** In particular, by (2.69), the functions in \( \mathcal{G} \) restricted to \( \mathcal{X} \) are distinct. By Sauer’s lemma (Lemma 2.79),

\[
N = |\mathcal{G}_X| \leq |\mathcal{H}_X| \leq \left( \frac{em}{\text{vc}(\mathcal{H})} \right)^{\text{vc}(\mathcal{H})} \leq \left( \frac{eC\varepsilon^{-4} \log N}{\text{vc}(\mathcal{H})} \right)^{\text{vc}(\mathcal{H})}. \] (2.70)

Using that \( \frac{1}{2D} \log N = \log N^{1/2D} \leq N^{1/2D} \) where \( D = \text{vc}(\mathcal{H}) \), we get

\[
\left( \frac{eC\varepsilon^{-4} \log N}{\text{vc}(\mathcal{H})} \right)^{\text{vc}(\mathcal{H})} \leq (C' \varepsilon^{-4})^{\text{vc}(\mathcal{H})} N^{1/2}, \] (2.71)

where \( C' = 2eC \). Plugging (2.71) back into (2.70) and rearranging gives

\[
N \leq \left( C' \varepsilon^{-4} \right)^{2\text{vc}(\mathcal{H})}. \]

That concludes the proof.  

**Proof of Sauer’s lemma**  To be written. See [Ver18, Section 8.3.3].

**Exercises**

**Exercise 2.1** (Bonferroni inequalities). Let \( A_1, \ldots, A_n \) be events and \( B_n := \cup_i A_i \). Define

\[
S(r) := \sum_{1 \leq i_1 < \cdots < i_r \leq n} \mathbb{P}[A_{i_1} \cap \cdots \cap A_{i_r}],
\]

and

\[
X_n := \sum_{i=1}^{n} 1_{A_i}.
\]

a) Let \( x_0 \leq x_1 \leq \cdots \leq x_s \geq x_{s+1} \geq \cdots \geq x_m \) be a unimodal sequence of non-negative reals such that \( \sum_{j=0}^{m} (-1)^j x_j = 0 \). Show that \( \sum_{j=0}^{\ell} (-1)^j x_j \) is \( \geq 0 \) for even \( \ell \) and \( \leq 0 \) for odd \( \ell \).
b) Show that, for all \( r \),
\[
\sum_{1 \leq i_1 < \cdots < i_r \leq n} 1_{A_{i_1}} 1_{A_{i_2}} \cdots 1_{A_{i_r}} = \binom{X_n}{r}.
\]

c) Use a) and b) to show that when \( \ell \in [n] \) is odd
\[
\mathbb{P}[B_n] \leq \sum_{r=1}^{\ell} (-1)^{r-1} S^{(r)},
\]
and when \( \ell \in [n] \) is even
\[
\mathbb{P}[B_n] \geq \sum_{r=1}^{\ell} (-1)^{r-1} S^{(r)}.
\]

These inequalities are called Bonferroni inequalities. The case \( \ell = 1 \) is Boole’s inequality.

Exercise 2.2 (Percolation on \( \mathbb{Z}^2 \): better bound [Ste]). Let \( E_1 \) be the event that all edges are open in \([-N, N] \times [-N, N]\) and \( E_2 \) be the event that there is no closed self-avoiding dual cycle surrounding \([-N, N]^2\). By looking at \( E_1 \cap E_2 \), show that \( \theta(p) > 0 \) for \( p > 2/3 \).

Exercise 2.3 (Percolation on \( \mathbb{Z}^d \): existence of critical threshold). Consider bond percolation on \( \mathbb{L}^d \).

a) Show that \( p_c(\mathbb{L}^d) > 0 \). [Hint: Count self-avoiding paths.]
b) Show that \( p_c(\mathbb{L}^d) < 1 \). [Hint: Use the result for \( \mathbb{L}^2 \).]

Exercise 2.4 (Sums of uncorrelated variables). Centered random variables \( X_1, \ldots, X_n \) are pairwise uncorrelated if
\[
\mathbb{E}[X_r X_s] = 0, \quad \forall r \neq s.
\]
Assume further that \( \text{Var}[X_r] \leq C < +\infty \) for all \( r \). Show that
\[
\mathbb{P}\left[ \frac{1}{n} \sum_{r \leq n} X_r \geq \beta \right] \leq \frac{C^2}{\beta^2 n}.
\]

Exercise 2.5 (Pairwise independence: lack of concentration [LW06]). Let \( U = (U_1, \ldots, U_\ell) \) be uniformly distributed over \( \{0, 1\}^\ell \). Let \( n = 2^\ell - 1 \). For all \( v \in \{0, 1\}^\ell \setminus \{0\} \), define
\[
X_v = (U \cdot v) \mod 2.
\]

95
a) Show that the random variables $X_v, v \in \{0, 1\}^\ell \setminus \{0\}$, are uniformly distributed in $\{0, 1\}$ and pairwise independent.

b) Show that for any event $A$ measurable with respect to $\sigma(X_v, v \in \{0, 1\}^\ell \setminus \{0\})$, $P[A]$ is either 0 or $\geq 1/(n+1)$.

Exercise 2.4 shows that pairwise independence implies “polynomial concentration” of the average of square-integrable $X_v$s. On the other hand, the current exercise suggests that in general pairwise independence cannot imply “exponential concentration.”

**Exercise 2.6** (Chernoff bound for Poisson trials). Using the Chernoff-Cramér method, prove part (a) of Theorem 2.37. Show that part (b) follows from part (a).

**Exercise 2.7** (A proof of Pólya’s theorem). Let $(X_t)$ be simple random walk on $\mathbb{L}^d$ started at the origin 0.

a) For $d = 1$, use Stirling’s formula to show that $P_0[X_{2n} = 0] = \Theta(n^{-1/2})$.

b) For $j = 1, \ldots, d$, let $N^{(j)}_t$ be the number of steps in the $j$-th coordinate by time $t$. Show that

$$P \left[ N^{(j)}_n \in \left[ \frac{n}{2d}, \frac{3n}{2d} \right], \forall j \right] \geq 1 - \exp(-\kappa_d n),$$

for some constant $\kappa_d > 0$.

c) Use a) and b) to show that, for any $d \geq 3$, $P_0[X_{2n} = 0] = O(n^{-d/2})$.

**Exercise 2.8** (Maximum degree). Let $G_n = (V_n, E_n) \sim G_{n,p_n}$ be an Erdős-Rényi graph with $n$ vertices and density $p_n$. Suppose $np_n = C \log n$ for some $C > 0$. Let $D_n$ be the maximum degree of $G_n$. Use Bernstein’s inequality to show that for any $\varepsilon > 0$

$$P[D_n \geq (n-1)p_n + \max\{C, 4(1+\varepsilon)\} \log n] \to 0,$$

as $n \to +\infty$.

**Exercise 2.9** (RIP v. orthogonality). Show that a $(k, 0)$-RIP matrix with $k \geq 2$ is orthogonal, i.e., its columns are orthonormal.

**Exercise 2.10** (Compressed sensing: linear programming formulation). Formulate (2.52) as a linear program, i.e., the minimization of a linear objective subject to linear inequalities.
Exercise 2.11 (Compressed sensing: almost sparse case). Prove the almost sparse case of Claim 2.66 by adapting the proof of the sparse case.

Exercise 2.12 (Chaining tail inequality). Prove Theorem 2.60.

Exercise 2.13 (Poisson convergence: method of moments). Let $A_1, \ldots, A_n$ be events and $A := \cup_i A_i$. Define

$$S^{(r)} := \sum_{1 \leq i_1 < \cdots < i_r \leq n} \mathbb{P}[A_{i_1} \cap \cdots \cap A_{i_r}],$$

and

$$X_n := \sum_{i=1}^n A_i.$$ 

Assume that there is $\mu > 0$ such that, for all $r$,

$$S^{(r)} \to \frac{\mu^r}{r!}.$$ 

Use Exercise 2.1 and a Taylor expansion of $e^{-\mu}$ to show that

$$\mathbb{P}[X_n = 0] \to e^{-\mu}.$$ 

In fact, $X_n \xrightarrow{d} \text{Poi}(\mu)$ (no need to prove this). This is a special case of the method of moments. See e.g. [Dur10, Section 3.3.5] and [JLR11, Section 6.1].

Exercise 2.14 (Connectivity: critical window). Using Exercise 2.13 show that, when $p_n = \frac{\log n + s}{n}$, the probability that an Erdős-Rényi graph $G_n \sim G_{n,p_n}$ contains no isolated vertex converges to $e^{-e^{-s}}$.

Bibliographic remarks

Section 2.2 The examples in Section 2.2.1 are taken from [AS11, Sections 2.4, 3.2]. A fascinating account of the longest increasing subsequence problem is given in [Rom14], from which the material in Section 2.2.3 is taken. The contour lemma, Lemma 2.17, is attributed to Whitney [Whi32] and is usually proved “by picture” [Gri10a, Figure 3.1]. A formal proof of the lemma can be found in [Kes82, Appendix A]. For much more on percolation, see [Gri10b]. A gentler introduction is provided in [Ste].
Section 2.3 The presentation in Section 2.3.2 follows [AS11, Section 4.4] and [JLR11, Section 3.1]. The result for general subgraphs is due to Bollobás [Bol81]. A special case (including cliques) was proved by Erdős and Rényi [ER60]. For variants of the small subgraph containment problem involving copies that are induced, disjoint, isolated etc., see e.g. [JLR11, Chapter 3]. For corresponding results for larger graphs, such as cycles or matchings, see e.g. [Bol01]. The result in that section is due to Erdős and Rényi [ER60]. The connectivity threshold in Section 2.3.2 is also due to the same authors [ER59]. The presentation here follows [vdH14, Section 5.2]. Theorem 2.30 is due to R. Lyons [Lyo90].

Section 2.4 The use of the moment-generating function to derive tail bounds for sums of independent random variables was pioneered by Cramér [Cra38], Bernstein [Ber46], and Chernoff [Che52]. For much more on concentration inequalities, see e.g. [BLM13]. The basics of large deviation theory are covered in [Dur10, Section 2.6]. See also [RAS] and [DZ10]. The presentation in Section 2.4.6 is based on [Har, Lectures 6 and 8] and [Tao]. Section 2.4.4 is based partly on [Ver18] and [Lug, Section 3.2]. Very insightful, and much deeper, treatment of the material in Section 2.4.5 can be found in [Ver18, vH16]. The Johnson-Lindenstrauss lemma was first proved by Johnson and Lindenstrauss using non-probabilistic arguments [JL84]. The idea of using random projections to simplify the proof was introduced by Frankl and Maehara [FM88] and the proof presented here based on Gaussian projections is due to Indyk and Motwani [IM98]. See [Ach03] for an overview of the various proofs known. For more on the random projection method, see [Vem04]. For algorithmic applications of the Johnson-Lindenstrauss lemma, see e.g. [Har, Lecture 7]. Compressed sensing emerged in the works of Donoho [Don06] and Candès, Romberg and Tao [CRT06a, CRT06b]. The restricted isometry property was introduced by Candès and Tao [CT05]. Claim 2.66 is due to Candés, Romberg and Tao [CRT06b]. The proof of Claim 2.67 presented here is due to Baraniuk et al. [BDDW08]. A survey of compressed sensing can be found in [CW08]. A thorough mathematical introduction to compressed sensing can be found in [FR13]. The presentation in Section 2.4.7 follows [KP, Section 3] and [LP, Section 13.3]. The Varopoulos-Carne bound is due to Carne [Car85] and Varopoulos [Var85]. For a probabilistic approach to the Varopoulos-Carne bound see Peyre’s proof [Pey08]. The application to mixing times is from [LP]. The material in Section 2.4.3 can be found in [BLM13, Chapter 2]. Hoeffding’s lemma and inequality are due to Hoeffding [Hoe63]. Bennett’s inequality is due to Bennett [Ben62]. Section 2.4.8 borrows from [Ver18, vH16, SSBD14, Haz16].
Chapter 3

Martingales and potentials

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

3.1 Background

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

**Definition 3.1** (Filtered space). A filtered space is a tuple \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathbb{Z}^+_0}, \mathbb{P})\) where:

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

  where each \(\mathcal{F}_t\) is a \(\sigma\)-field.

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

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

Fix \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathbb{Z}^+_0}, \mathbb{P})\). A process \((W_t)_t\) is adapted if \(W_t \in \mathcal{F}_t\) for all \(t\). A process \(\{C_t\}_{t \geq 1}\) is predictable if \(C_t \in \mathcal{F}_{t-1}\) for all \(t\).
Example 3.3 (i.i.d. random variables (continued)). The process \((S_t)_t\), where \(S_t = \sum_{i \leq t} X_i\), is adapted. The process \((C_t)_t\), where \(C_t = 1\{S_{t-1} \leq k\}\), is predictable.

3.1.1 Stopping times

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

**Definition 3.4 (Stopping time).** A random variable \(\tau : \Omega \to \mathbb{Z}_+ := \{0, 1, \ldots, +\infty\}\) is called a stopping time if

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

or, equivalently,

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

(To see the equivalence, note that \(\{\tau = t\} = \{\tau \leq t\} \setminus \{\tau \leq t - 1\}\), and \(\{\tau \leq t\} = \bigcup_{i \leq t} \{\tau = i\}\).)

**Example 3.5 (Hitting time).** Let \((A_t)_{t \in \mathbb{Z}_+}\), with values in \((E, \mathcal{E})\), be adapted and \(B \in \mathcal{E}\). Then

\[
\tau = \inf\{t \geq 0 : A_t \in B\},
\]

is a stopping time known as a hitting time. In contrast, the last visit to a set is typically not a stopping time.

Let \(\tau\) be a stopping time. Denote by \(\mathcal{F}_\tau\) the set of all events \(F\) such that \(\forall t \in \mathbb{Z}_+\)

\[
F \cap \{\tau = t\} \in \mathcal{F}_t.
\]

Roughly speaking, the \(\sigma\)-field \(\mathcal{F}_\tau\) captures the information up to time \(\tau\). The following lemmas help clarify the definition of \(\mathcal{F}_\tau\).

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

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

For the third claim note that

\[
F = \bigcup_{s \in \mathbb{Z}_+} F \cap \{\tau = s\} \in \mathcal{F}_\infty.
\]

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

\[\blacksquare\]

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

Proof. Let $F \in \mathcal{F}_{\sigma \land \tau}$. Note that
\[
F \cap \{\tau = t\} = \bigcup_{s \leq t} \left( (F \cap \{\sigma \land \tau = s\}) \cap \{\tau = t\} \right) \in \mathcal{F}_t.
\]
Indeed, the expression in parenthesis is in $\mathcal{F}_s \subseteq \mathcal{F}_t$ and $\{\tau = t\} \in \mathcal{F}_t$.  

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

Definition 3.9 (First visit and return). The first visit time and first return time to first visit time, first return time
\[
\tau_x := \inf\{t \geq 0 : X_t = x\} \quad \text{and} \quad \tau_x^+ := \inf\{t \geq 1 : X_t = x\}.
\]
Similarly, $\tau_B$ and $\tau_B^+$ are the first visit and first return to $B \subseteq V$.

Definition 3.10 (Cover time). Assume $V$ is finite. The cover time of $(X_t)$ is the first time that all states have been visited, i.e.,
\[
\tau_{\text{cov}} := \inf\{t \geq 0 : \{X_0, \ldots, X_t\} = V\}.
\]

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

Theorem 3.11 (Strong Markov property). Let $f_t : V^\infty \to \mathbb{R}$ be a sequence of measurable functions, uniformly bounded in $t$ and let $F_t(x) := \mathbb{E}_x[f_t((X_t)_{t \geq 0})]$, then
\[
\mathbb{E}[f_\tau((X_{\tau+t})_{t \geq 0}) \mid \mathcal{F}_\tau] = F_\tau(X_\tau) \quad \text{on} \{\tau < +\infty\}.
\]
Proof. Let $A \in \mathcal{F}_\tau$. Summing over the possible values of $\tau$ and using the Markov property

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

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

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

That concludes the proof. ■

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

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

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

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

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

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

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

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

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

Summing over $a > 0$ and rearranging gives

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

**Proof.** Reflect the sub-path after the first visit to $b$ across the line $y = b$. ■
We record another related result that will be useful later (see e.g. [Dur10, Theorem 4.3.2]).

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

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

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

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

\[
\mathbb{P}_x[\tau^+_y < +\infty] = \rho_{xy}\rho_{yy}^{k-1}.
\]

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

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

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

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

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

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

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

\[
\mathbb{P}_0[S_{2t} = 0] = \binom{2t}{t} 2^{-2t} \sim 2^{-2t} \frac{(2t)^{2t}}{(t!)^2} \frac{\sqrt{2t}}{\sqrt{2\pi t}} \sim \frac{1}{\sqrt{\pi t}}.
\]
Thus
\[ E_0[N_0] = \sum_{t > 0} P_0[S_t = 0] = +\infty, \]
and the chain is recurrent.

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

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

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

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

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

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

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

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

$$G_\sigma(x, z) = \mathbb{E}_x \left[ \sum_{0 \leq t < \sigma} 1_{X_t = z} \right] = \mathbb{E}_x \left[ \sum_{0 \leq t < \sigma} 1_{X_{t+1} = z} \right] = \sum_{t \geq 0} \mathbb{P}_x[X_{t+1} = z, \sigma > t].$$

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

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

That proves the claim.

Here is a typical application of this lemma.

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

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

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

Many more results of this form can be derived from the occupation measure identity. See e.g. [AF, Chapter 2].
3.1.2  

Markov chains: exponential tail of hitting times and some cover time bounds

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

Lemma 3.25. Let \((X_t)\) be a finite, irreducible Markov chain with state space \(V\) and initial distribution \(\mu\). For \(A \subseteq V\), there is \(\beta_1 > 0\) and \(0 < \beta_2 < 1\) depending on \(A\) such that

\[
P_{\mu}[\tau_A > t] \leq \beta_1 \beta_2^t,
\]

In particular, \(E_{\mu}[\tau_A] < +\infty\) for any \(\mu, A\).

Proof. For any integer \(m\), for some distribution \(\theta\),

\[
P_{\mu}[\tau_A > ms | \tau_A > (m - 1)s] = P_{\theta}[\tau_A > s] \leq \max_x P_x[\tau_A > s] =: 1 - \alpha_s.
\]

Choose \(s\) large enough that, from any \(x\), there is a path to \(A\) of length at most \(s\) of positive probability. Such an \(s\) exists by irreducibility. In particular \(\alpha_s > 0\). By induction,

\[
P_{\mu}[\tau_A > ms] \leq (1 - \alpha_s)^m \text{ or } P_{\mu}[\tau_A > t] \leq (1 - \alpha_s)^{|t|} \leq \beta_1 \beta_2^t \text{ for } \beta_1 > 0 \text{ and } 0 < \beta_2 < 1 \text{ depending on } \alpha_s.
\]

The result for the expectation follows from

\[
E_{\mu}[\tau_A] = \sum_{k \geq 0} P_{\mu}[\tau_A > k] \leq \sum_t \beta_1 \beta_2^t < +\infty.
\]

That concludes the proof.  

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

Lemma 3.26. Let \((X_t)\) be a finite, irreducible Markov chain with state space \(V\) and initial distribution \(\mu\). For \(A \subseteq V\), let \(\bar{t}_A := \max_x E_x[\tau_A]\). Then

\[
P_{\mu}[\tau_A > t] \leq \exp \left( - \left| \frac{t}{\bar{t}_A} \right| \right).
\]

Proof. For any integer \(m\), for some distribution \(\theta\),

\[
P_{\mu}[\tau_A > ms | \tau_A > (m - 1)s] = P_{\theta}[\tau_A > s] \leq \max_x P_x[\tau_A > s] \leq \frac{\bar{t}_A}{s},
\]

by the Markov property and Markov’s inequality (Theorem 2.1). By induction,

\[
P_{\mu}[\tau_A > ms] \leq \left( \frac{\bar{t}_A}{s} \right)^m \text{ or } P_{\mu}[\tau_A > t] \leq \left( \frac{\bar{t}_A}{t} \right)^{|t|} \text{. By differentiating w.r.t. } s, \text{ it can be checked that a good choice is } s = \left[ e \bar{t}_A \right].
\]
Application to cover times We give an application of the previous bound to cover times. Let \((X_t)\) be a finite, irreducible Markov chain on \(V\) with \(n := |V| > 1\). Recall that the cover time is \(\tau_{\text{cov}} := \max_y \tau_y\). We bound the mean cover time in terms of \(\bar{t}_{\text{hit}} := \max_{x \neq y} E_x \tau_y\).

Claim 3.27. 
\[
\max_x E_x [\tau_{\text{cov}}] \leq (3 + \ln n) \lceil e \bar{t}_{\text{hit}} \rceil.
\]

Proof. By a union bound (Corollary 2.11) over all states to be visited and Lemma 3.26,
\[
\max_x P_x [\tau_{\text{cov}} > t] \leq \min \left\{ 1, n \cdot \exp \left( - \left\lfloor \frac{t}{e \bar{t}_{\text{hit}}} \right\rfloor \right) \right\}.
\]
Summing over \(t\) and appealing to the sum of a geometric series,
\[
\max_x E_x [\tau_{\text{cov}}] \leq (\ln(n) + 1) \lceil e \bar{t}_{\text{hit}} \rceil + \frac{1}{1 - e^{-1}} \lceil e \bar{t}_{\text{hit}} \rceil.
\]

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

Theorem 3.28 (Matthews’ cover time bounds). Let \(\bar{t}_{\text{hit}}^A := \min_{x,y \in A, x \neq y} E_x \tau_y\) and \(h_n := \sum_{m=1}^n \frac{1}{m}\). Then
\[
\max_x E_x [\tau_{\text{cov}}] \leq h_n \bar{t}_{\text{hit}},
\]
and
\[
\min_x E_x [\tau_{\text{cov}}] \geq \max_{A \subseteq V} h_{|A|-1} \bar{t}_{\text{hit}}^A.
\]

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

Proof. We prove the lower bound for \(A = V\). The other cases are similar. Let \((J_1, \ldots, J_n)\) be a uniform random ordering of \(V\), let \(C_m := \max_{i \leq m} \tau_i\), and let \(L_m\) be the last state visited among \(J_1, \ldots, J_m\). Then
\[
E_x [C_m - C_{m-1} \mid J_1, \ldots, J_m, \{X_t, t \leq C_{m-1}\}] \geq \bar{t}_{\text{hit}}^V 1_{\{L_m = J_m\}}.
\]
By symmetry, \(P[L_m = J_m] = \frac{1}{m}\). To see this, first pick the set of vertices corresponding to \(\{J_1, \ldots, J_m\}\), wait for all of these vertices to be visited, then pick the ordering. Moreover observe that \(E_x C_1 \geq (1 - \frac{1}{n}) \bar{t}_{\text{hit}}^V\) accounts for the probability that \(J_1 \neq x\). Taking expectations above and summing over \(m\) gives the result.
Remark 3.29. The bounds (3.1) and (3.2) are tight up to smaller order for the coupon collector problem, which can be stated in terms of the cover time of a lazy random walk on the complete graph.

3.1.3 Martingales

Definition We first recall the definition of a martingale.

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

$$\mathbb{E}[M_{t+1} | F_t] = M_t, \quad \forall t \geq 0$$

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

Note that for a martingale, by the tower property (Lemma A.5), we have

$$\mathbb{E}[M_n | F_{m}] = M_m$$

for all $n > m$, and similarly for supermartingales and submartingales.

Jensen’s inequality immediately implies:

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

We start with a straightforward example.

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

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

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

Martingales can also be a little more hidden.

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

$$\mathbb{E}[M_t | F_{t-1}] = \mathbb{E}[(X_t + S_{t-1})^2 - t\sigma^2 | F_{t-1}]$$

$$= \mathbb{E}[X_t^2 + 2X_tS_{t-1} + S_{t-1}^2 - t\sigma^2 | F_{t-1}]$$

$$= \sigma^2 + 0 + S_{t-1}^2 - t\sigma^2 = M_{t-1},$$

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

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

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

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

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

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

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

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

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

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

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

**Pólya’s urn** Define $M_t$ to be the fraction of green balls after the $t$th draw. Then

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

Some useful properties We collect here a few more properties of martingales.

**Theorem 3.38 (Doob’s submartingale inequality).** Let $(M_t)$ be a nonnegative submartingale. Then for $b > 0$
\[
P\left[\sup_{1 \leq s \leq t} M_s \geq b\right] \leq \frac{\mathbb{E}[M_t]}{b}.
\]
Observe that Markov’s inequality (Theorem 2.1) implies only that
\[
\sup_{1 \leq s \leq t} \mathbb{P}[M_s \geq b] \leq \frac{\mathbb{E}[M_t]}{b},
\]
where we used that $\mathbb{E}[M_t] \geq \mathbb{E}[M_s]$ for all $1 \leq s \leq t$.

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

A useful consequence of the previous inequality:

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

We will also need the following orthogonality property.

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

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

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

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

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

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

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

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

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

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

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

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

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

if one of the following holds:

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

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

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

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

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

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

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

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

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

We illustrate Wald’s identities on an important example.

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

**Claim 3.47.** We have:

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

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

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

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

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

1) We argue that in fact $\mathbb{E}[\tau] < \infty$. That follows immediately from the exponential tail of hitting times in Lemma 3.25 for the chain $(S_{\tau \wedge t})$ whose state space, $\{a, a+1, \ldots, b\}$, is finite.
2) By Wald’s first identity, $E[S_\tau] = 0$ or
\[ a \mathbb{P}[S_\tau = a] + b \mathbb{P}[S_\tau = b] = 0, \]
that is,
\[ \mathbb{P}[\tau_\alpha < \tau_b] = \frac{b}{b-a} \quad \text{and} \quad \mathbb{P}[\tau_\alpha < +\infty] \geq \mathbb{P}[\tau_\alpha < \tau_b] \rightarrow 1, \]
where we took $b \rightarrow \infty$ in the first expression to obtain the second one.

3) Because $\sigma^2 = 1$, Wald’s second identity says that $E[S^2_\tau] = E[\tau]$. Furthermore, we have by 2)
\[ E[S^2_\tau] = \frac{b}{b-a}a^2 + \frac{-a}{b-a}b^2 = -ab. \]
Thus $E[\tau] = -ab$.

4) The first claim was proved in 2). When $b \rightarrow +\infty$, $\tau_\alpha \wedge \tau_\beta \uparrow \tau_\alpha$ and monotone convergence applied to 3) gives that $E[\tau_\alpha] = +\infty$.

That concludes the proof.

Note that 4) shows that the $L^1$ condition on the stopping time in Wald’s second identity is necessary. Indeed, we have shown $a^2 = E[S^2_\tau_\alpha] \neq a^2E[\tau_\alpha] = +\infty$.

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

Claim 3.49. We have:

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

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

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

4) $\tau_\alpha = +\infty$ with positive probability.

Proof. Let $\psi_t(x) := x - (p-q)t$. We use two martingales: $(\phi(S_t))$ and $(\psi_t(S_t))$. Observe that indeed
\[ E[\phi(S_t) | F_{t-1}] = p(q/p)^{S_{t-1}+1} + q(q/p)^{S_{t-1}-1} = \phi(S_{t-1}), \]
and
\[ E[\psi_t(S_t) | F_{t-1}] = p[S_{t-1} + 1 - (p-q)t] + q[S_{t-1} - 1 - (p-q)t] = \psi_{t-1}(S_{t-1}). \]
1) This claim follows by the same argument as in the unbiased case.

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

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

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

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

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

3) By Theorem 3.42 again,

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

By monotone convergence, \(\mathbb{E}[\tau_b \wedge t] \uparrow \mathbb{E}[\tau_b]\). Furthermore, observe that \(-\inf_t S_t \geq 0\) a.s. since \(S_0 = 0\) and, for \(x \geq 0\), by (3.3)

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

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

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

4) That claim was proved in 2).

That concludes the proof.

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

\[\nabla\]

3.1.4 Percolation on trees: critical regime

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

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

Theorem 3.50. \(|C_0| < +\infty\) a.s.
Proof. Let $b := d - 1$ be the branching ratio. Because the root has a different branching ratio, we consider the descendants of its children. Let $Z_n$ be the number of vertices in the open cluster of the first child of the root $n$ levels below it and let $\mathcal{F}_n = \sigma(Z_0, \ldots, Z_n)$. Then $Z_0 = 1$ and

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

So $(Z_n)$ is a nonnegative, integer-valued martingale and it converges to an a.s. finite limit. But, clearly, for any integer $k > 0$ and $N \geq 0$

$$
P[Z_n = k, \forall n \geq N] = 0,
$$
so $Z_\infty \equiv 0$.

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

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

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

- $A_t$: active vertices
- $E_t$: explored vertices
- $N_t$: neutral vertices

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

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

Let $M_t := |A_t|$. Revealing the edges as they are explored and letting $(\mathcal{F}_t)$ be the corresponding filtration, we have $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1} + bp - 1 = M_{t-1}$ on
\{M_{t-1} > 0\} so \(M_t\) is a nonnegative martingale. Let \(\sigma^2 := bp(1 - p) \geq \frac{1}{3}\), \(\tau := \inf\{t \geq 0 : M_t = 0\}\), and \(Y_t := M_{t \wedge \tau} - \sigma^2(t \wedge \tau)\). Then, on \(\{M_{t-1} > 0\}\),

\[
E[Y_t \mid F_{t-1}] = E[(M_{t-1} + (M_t - M_{t-1}))^2 - \sigma^2 t \mid F_{t-1}]
\]

\[
= E[M_{t-1}^2 + 2M_{t-1}(M_t - M_{t-1}) + (M_t - M_{t-1})^2 - \sigma^2 t \mid F_{t-1}]
\]

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

so \((Y_t)\) is also a martingale.

For \(h > 0\), let

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

Note that \(\tau'_h \leq \tau = |\tilde{C}_1| < +\infty\) a.s. We use

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

By Markov’s inequality (Theorem 2.1),

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

and

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

To compute \(E M_{\tau'_h}\), we use Theorem 3.42 to obtain

\[
1 = E[M_{\tau'_h \wedge s}] \to E[M_{\tau'_h}],
\]

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

To compute \(E \tau'_h\), we use Theorem 3.42 again

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

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

\[
E[M_{\tau'_h}^2 \mid M_{\tau'_h} \geq h] \leq (h + b)^2,
\]

we have

\[
E\tau'_h \leq \frac{1}{\sigma^2} \left\{ \frac{1}{h} E[M_{\tau'_h}^2 \mid M_{\tau'_h} \geq h] \right\} \leq \frac{(h + b)^2}{\sigma^2 h} \leq \frac{2(h + b)^2}{h}.
\]

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

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

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

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

3.2.1 Azuma-Hoeffding inequality

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

**Theorem 3.52** (Maximal Azuma-Hoeffding inequality). Let $(Z_t)_{t \in \mathbb{Z}^+}$ be a martingale with respect to the filtration $(\mathcal{F}_t)_{t \in \mathbb{Z}^+}$. 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,

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

Then for all $\beta > 0$

$$\mathbb{P}\left[ \sup_{0 \leq i \leq t} (Z_i - Z_0) \geq \beta \right] \leq \exp\left( -\frac{2\beta^2}{\sum_{i=1}^{t} c_i^2} \right).$$

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

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

\textsuperscript{*}Requires: Section 2.4.3.

\textsuperscript{†}Quoting [vH].
Hence, for $s > 0$, by Theorem 3.38

$$
P \left[ \sup_{0 \leq t \leq T} (Z_t - Z_0) \geq \beta \right] = P \left[ \sup_{0 \leq i \leq t} e^{s(Z_i - Z_0)} \geq e^{s\beta} \right] \leq \frac{E \left[ e^{s(Z_t - Z_0)} \right]}{e^{s\beta}} = \frac{E \left[ e^{s \sum_{r=1}^{t} (Z_r - Z_{r-1})} \right]}{e^{s\beta}}. \quad (3.4)
$$

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

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

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

$$
E \left[ e^{s(Z_t - Z_{t-1}) \big| F_{t-1}} \right] \leq \exp \left( \frac{s^2 c_t^2}{2} \right) = \exp \left( \frac{s^2 c_t^2}{8} \right). \quad (3.5)
$$

Arguing by induction, we obtain

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

so that we can write

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

120
so that
\[ |\mathbb{E}[f(X_1, \ldots, X_n) \mid X_1] - \mathbb{E}[f(X_1, \ldots, X_n)]| = n > 2. \]

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

For a less extreme example which has support over all of \((-1, 1)^n\), let
\[ U_i = \begin{cases} 1, & \text{w.p. } 1 - \varepsilon \\ -1, & \text{w.p. } \varepsilon, \end{cases} \]
for some \( \varepsilon > 0 \) independently for all \( i = 1, \ldots, n - 1 \). Let again \( X_1 \) be a uniform random variable on \((-1, 1)\) and, for \( i = 2, \ldots, n \), define the random variable \( X_i = U_{i-1}X_{i-1} \), that is, \( X_i \) is the same as \( X_{i-1} \) with probability \( \varepsilon \) and otherwise is flipped. Then,
\[ \mathbb{E}[f(X_1, \ldots, X_n)] = \mathbb{E}[X_1 + \cdots + X_n] \]
\[ = \mathbb{E}
\left[
X_1
\left(1 + \sum_{i = 1}^{n-1} \prod_{j \leq i} U_j\right)\right] \]
\[ = \mathbb{E}[X_1]\mathbb{E}
\left[1 + \sum_{i = 1}^{n-1} \prod_{j \leq i} U_j\right] \]
\[ = 0, \]
by the independence of \( X_1 \) and the \( U_i \)'s. Similarly
\[ \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1] = X_1 \mathbb{E}
\left[1 + \sum_{i = 1}^{n-1} \prod_{j \leq i} U_j\right] = X_1 \left(\sum_{i = 1}^{n} (1 - 2\varepsilon)^i\right), \]
so that
\[ |\mathbb{E}[f(X_1, \ldots, X_n) \mid X_1] - \mathbb{E}[f(X_1, \ldots, X_n)]| = \left(\sum_{i = 1}^{n} (1 - 2\varepsilon)^i\right) > 2, \]
for \( \varepsilon \) small enough and \( n \geq 3 \). In particular, the corresponding Doob martingale does not have increments bounded by 2.

Bounds on \( \|D_if\|_\infty \) are often expressed in terms of a Lipschitz condition under an appropriate metric. The Hamming distance is defined as
\[ \rho(x, x') := \sum_{i = 1}^{n} 1_{\{x_i \neq x'_i\}}, \]

121
for $x, x' \in X_1 \times \cdots \times X_n$. Let $0 < c < +\infty$. A function $f : X_1 \times \cdots \times X_n \to \mathbb{R}$ is $c$-Lipschitz (with respect to the Hamming distance) if for all $i = 1, \ldots, n$, all $(x_1, \ldots, x_n) \in X_1 \times \cdots \times X_n$ and all $y, y' \in X_i$

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

Lemma 3.54. If $f$ is $c$-Lipschitz, then

$$\|D_if\|_\infty \leq c, \quad \forall i.$$ 

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

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

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

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

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

Lemma 3.56.

$$\mathbb{E}[\mathbb{E}[f(X) | G_i] | F_i] = \mathbb{E}[f(X) | F_i].$$

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

$$\mathbb{E}[f(X) | F_i] = \mathbb{E}[\mathbb{E}[f(X) | G_i] | F_i].$$

Moreover, $\sigma(X_i)$ is independent of $\sigma(G_i, F_{i-1})$ so by the role of independence (Lemma A.4), we have

$$\mathbb{E}[\mathbb{E}[f(X) | G_i] | F_{i-1}, X_i] = \mathbb{E}[\mathbb{E}[f(X) | G_i] | F_i].$$

Combining the last two displays gives the result. 

122
Again, we take advantage of the orthogonality of increments (Lemma 3.40) to write
\[
\text{Var}[f(X)] = \sum_{i=1}^{n} \mathbb{E} \left[ (Z_i - Z_{i-1})^2 \right].
\]

By the lemma above,
\[
(Z_i - Z_{i-1})^2 = (\mathbb{E} [f(X) | F_i] - \mathbb{E} [f(X) | F_{i-1}])^2
= (\mathbb{E} [f(X) | F_i] - \mathbb{E} [\mathbb{E} [f(X) | G_i] | F_i])^2
= (\mathbb{E} [f(X) - \mathbb{E} [f(X) | G_i] | F_i])^2
\leq \mathbb{E} \left[ (f(X) - \mathbb{E} [f(X) | G_i])^2 \right] F_i,
\]
where we used Jensen’s inequality on the last line. Taking expectations
\[
\text{Var}[f(X)] = \sum_{i=1}^{n} \mathbb{E} \left[ (Z_i - Z_{i-1})^2 \right]
\leq \sum_{i=1}^{n} \mathbb{E} \left[ (f(X) - \mathbb{E} [f(X) | G_i])^2 \right] F_i
= \sum_{i=1}^{n} \mathbb{E} \left[ (f(X) - \mathbb{E} [f(X) | G_i])^2 \right] G_i
= \sum_{i=1}^{n} \mathbb{E} \left[ \text{Var} [f(X) | G_i] \right].
\]

That concludes the proof.

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

**Theorem 3.57 (Efron-Stein inequality).** Let $X_1, \ldots, X_n$ be independent random variables where $X_i$ is $X_i$-valued for all $i$ and let $X = (X_1, \ldots, X_n)$. Assume that $f : X_1 \times \cdots \times X_n \to \mathbb{R}$ is a measurable function with $\mathbb{E}[f(X)^2] < +\infty$. Let $X' = (X'_1, \ldots, X'_n)$ be an independent copy of $X$ and
\[
X^{(i)} = (X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n).
\]

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

123
Proof. Observe that if $Y'$ is an independent copy of $Y \in L^2$, then $\text{Var}[Y] = \frac{1}{2} \mathbb{E}[(Y - Y')^2]$, which can be seen by adding and subtracting the mean, expanding and using independence. Hence,

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

where we used the independence of the $X_i$s and $X'_i$s.

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

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

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

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

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

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

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

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

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

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

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

\[\Box\]

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

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

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

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

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

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

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

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

and

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

125
Indeed, since $\sigma(X_i)$ is independent of $\mathcal{F}_{i-1}$ and $\mathcal{G}_i$, by the role of independence (Lemma A.4)

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

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

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

**Examples**

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

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

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

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

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

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

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

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

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

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 application in the next section is more striking.

One more example:

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

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

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

**Claim 3.65.**

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

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

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

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

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

or, rearranging and using our assumption on \( A \),

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

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 3.65 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 \( A \). The smallest \( r \) such that \( |A_r| \geq (1 - \varepsilon)2^n \) in general depends on \( A \). Here are two extremes.

For \( \gamma > 0 \), let

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

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

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

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

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

where \( Z \sim N(0, 1) \). Let \( \varepsilon < \varepsilon' < 1/2 \) and let \( \gamma_{\varepsilon'} \) be such that \( \mathbb{P}[Z \leq -\gamma_{\varepsilon'}] = \varepsilon' \).

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

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

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

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

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

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

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

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

is known as a vertex exposure martingale. An alternative way to define the filtration is to consider instead the random variables \( X_i = (1_{\{i,j\} \in E} : 1 \leq j \leq i) \) for \( i = 2, \ldots, n \). In words, \( X_i \) is a vector whose entries indicate the status (present or absent) of all potential edges incident to \( i \) and a vertex preceding it. Hence, \( \mathcal{H}_i = \sigma(X_2, \ldots, X_i) \) for \( i = 2, \ldots, n \) (and note that \( \mathcal{H}_1 \) is trivial as it corresponds to a graph with a single vertex and no edge). This representation has an important property: the \( X_i \)'s are independent as they pertain to disjoint subsets of edges. We are then in the setting of the method of bounded differences. Re-writing \( F(G) = \)
\( f(X_1 \ldots X_n) \), the vertex exposure martingale coincides with the martingale (3.6) used in that context.

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

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

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

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

**Claim 3.68.**

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

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

**Remark 3.69.** Note that Claim 3.68 tells us nothing about the expectation of \( \chi(G) \). It turns out that, up to logarithmic factors, \( \mathbb{E}_{n,p_n}[\chi(G)] \) 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 \( \alpha > \frac{1}{2} \), \( \chi(G) \) is in fact concentrated on two values! We give a partial result along those lines which illustrates a less straightforward choice of martingale in the Azuma-Hoeffding inequality (Theorem 3.52).
Figure 3.1: All but $O(\sqrt{n})$ vertices are colored using $\varphi_n$ colors. The remaining vertices are colored using 3 additional colors.

**Claim 3.70.** Let $p_n = n^{-\alpha}$ with $\alpha > \frac{5}{6}$ and let $G_n \sim G_{n,p_n}$. Then for any $\varepsilon > 0$ there is $\varphi_n := \varphi_n(\alpha,\varepsilon)$ such that

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

for all $n$ large enough.

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

$$
P_{n,p_n}[\chi(G_n) \leq \varphi_n] > \frac{\varepsilon}{3}. \tag{3.9}$$

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

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

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

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

\[
P_{n,p} \left[ F_n(G_n) - E_{n,p}[F_n(G_n)] \right] \leq -b \varepsilon \sqrt{n - 1}
\]

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

\[
E_{n,p}[F_n(G_n)] \leq b \varepsilon \sqrt{n - 1}.
\]

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

\[
P_{n,p} \left[ F_n(G_n) \geq 2b \varepsilon \sqrt{n - 1} \right]
\]

\[
\leq P_{n,p} \left[ F_n(G_n) - E_{n,p}[F_n(G_n)] \geq b \varepsilon \sqrt{n - 1} \right]
\]

\[
\leq \frac{\varepsilon}{3}.
\]

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

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

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

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

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

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

\[
\mathbb{P}_{n,p}[Y_n > 0] \leq E_{n,p}Y_n
\]

\[
\leq \sum_{\ell=4}^{\frac{c\sqrt{n}}{2}} \frac{n}{\ell} \left( \frac{\ell}{\frac{\ell}{2}} \right)^{\frac{3\ell}{2}} p_n
\]

\[
\leq \sum_{\ell=4}^{\frac{c\sqrt{n}}{2}} \frac{en}{\ell} \left( \frac{e\ell}{3} \right)^{\frac{3\ell}{2}} n^{-\frac{3\ell}{2}}
\]

\[
\leq \sum_{\ell=4}^{\frac{c\sqrt{n}}{2}} \left( \frac{e^{\frac{3}{2}} n^{1-\frac{3\alpha}{2}} \ell^{\frac{1}{2}}}{3^{\frac{3}{2}}} \right)^{\ell}
\]

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

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

\[
\rightarrow 0,
\]

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

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

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

By (3.11) and (3.12),

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

So, overall,

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

### 3.2.4 Preferential attachment: degree sequence

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

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

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

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

Claim 3.73.

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

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

Lemma 3.74 (Convergence of the mean).

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

Lemma 3.75 (Concentration around the mean).

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

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

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

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

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

To form $G_s$, attach a new edge $e_s$ to the vertex of $G_{s-1}$ corresponding to $X_s$. A vertex of degree $d'$ in $G_{s-1}$ is selected with probability $\frac{d'}{2(s-1)}$, as it should. Note that $X_s$ can be picked in advance independently of the sequence $(G_{s'})_{s' < s}$. 

134
Figure 3.2: Graph obtained when $x_2 = (1, \text{head})$, $x_3 = (2, \text{tail})$ and $x_4 = (3, \text{head})$.

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

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

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

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

$$
\mathbb{P}[|N_d(t) - \mathbb{E}N_d(t)| \geq \beta] \leq 2 \exp \left(-\frac{2\beta^2}{(2)^2(t-1)}\right),
$$
which, choosing $\beta = \sqrt{2t \log \delta^{-1}}$, we can re-write as

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

\[ \blacksquare \]

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

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

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

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

We re-write (3.14) as

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

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

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

(3.15)

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

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

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

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

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

as $t \to +\infty$.

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

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

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

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

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

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

where we used (3.13). That concludes the proof of Lemma 3.74. 

\[\square\]
To prove Claim 3.73, we combine Lemmas 3.74 and 3.75. Fix any \( \delta, \varepsilon > 0 \). Choose \( t' \) large enough that for all \( t \geq t' \)

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

Then

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

for all \( t \geq t' \). That proves convergence in probability.

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

**Proof of Lemma 3.76.** By induction on \( t \), we have

\[
f(t + 1) = \left(1 - \frac{\alpha}{t}\right) f(t) + g(t)
\]

\[
= \left(1 - \frac{\alpha}{t}\right) \left[ \left(1 - \frac{\alpha}{t-1}\right) f(t-1) + g(t-1) \right] + g(t)
\]

\[
= \left(1 - \frac{\alpha}{t}\right) g(t-1) + g(t) + \left(1 - \frac{\alpha}{t}\right) \left(1 - \frac{\alpha}{t-1}\right) f(t-1)
\]

\[
= \ldots
\]

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

or

\[
f(t + 1) = \sum_{s=t_0}^{t} g(s) \prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) + f(t_0) \prod_{r=t_0}^{t} \left(1 - \frac{\alpha}{r}\right). \tag{3.16}
\]

To guess the answer note that, for large \( s \), \( g(s) \) is roughly constant and that the product in the first term behaves like

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

So approximating the sum by an integral we get that \( f(t + 1) \approx \frac{gt}{\alpha+1} \).

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

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

138
and that by a Taylor expansion, for \( |z| \leq 1/2 \),
\[
\log (1 - z) = -z + \Theta(z^2).
\]

Fix \( \eta > 0 \) small and take \( t \) large enough that \( \eta t > 2\alpha \) and \( |g(s) - g| < \eta \) for all \( s \geq \eta t \). Then, for \( s + 1 \geq t_0 \),
\[
\sum_{r=s+1}^{t} \log \left(1 - \frac{\alpha}{r}\right) = - \sum_{r=s+1}^{t} \left\{ \frac{\alpha}{r} + \Theta(r^{-2}) \right\}
= -\alpha (\log t - \log s) + \Theta(s^{-1}),
\]
so, taking exponentials,
\[
\prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) = \frac{s^{\alpha}}{t^{\alpha}} (1 + \Theta(s^{-1})).
\]

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

Moreover
\[
\frac{1}{t} \sum_{s=\eta t}^{t} g(s) \prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) \leq \frac{1}{t} \sum_{s=\eta t}^{t} (g + \eta) \frac{s^{\alpha}}{t^{\alpha}} (1 + \Theta(s^{-1}))
\leq O(\eta) + (1 + \Theta(t^{-1})) \frac{g}{t^{\alpha+1}} \sum_{s=\eta t}^{t} s^{\alpha}
\leq O(\eta) + (1 + \Theta(t^{-1})) \frac{g}{t^{\alpha+1}} \frac{(t + 1)^{\alpha+1}}{\alpha + 1}
\to O(\eta) + \frac{g}{\alpha + 1},
\]
and, similarly,
\[
\frac{1}{t} \sum_{s=t_0}^{\eta t - 1} g(s) \prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) \leq \frac{1}{t} \sum_{s=t_0}^{\eta t - 1} (g + \eta) \frac{s^{\alpha}}{t^{\alpha}} (1 + \Theta(s^{-1}))
\leq \frac{\eta t}{t} (g + \delta) \frac{(\eta t)^{\alpha}}{t^{\alpha}} (1 + \Theta(t_0^{-1}))
\to O(\eta^{\alpha+1}).
\]
Plugging these inequalities back into (3.16), we get
\[
\limsup_t \frac{1}{t} f(t + 1) \leq \frac{g}{1 + \alpha} + O(\eta).
\]
A similar inequality holds in the other direction. Taking \(\eta \to 0\) concludes the proof. \(\blacksquare\)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\overline{R}_n \leq \frac{2\alpha^2}{\Delta_s} \log n + \Delta_s C_{\alpha},$$

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

This bound should not come entirely as a surprise. Indeed a simple, alternative approach to UCB is to (1) first pull each arm $m_n = o(n)$ times and then (2) use the arm with largest estimated mean for the remainder. Assuming there is a known lower bound on $\Delta_s$, then Hoeffding’s inequality (Theorem 2.40) guarantees that $m_n$ can be chosen of the order of $\frac{1}{\Delta_s^2} \log n$ 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 $\Delta_s \frac{1}{\Delta_s^2} \log n + \Delta_s \approx \frac{1}{\Delta_s^2} \log n$. The UCB strategy, on the other hand, elegantly adapts to the gap $\Delta_s$ and the horizon $n$.

We break down the proof into a sequence of lemmas. We first rewrite the
pseudo-regret as
\[ R_n = n\mu^* - E \left[ \sum_{t=1}^{n} \mu_{I_t} \right] = E \left[ \sum_{t=1}^{n} (\mu^* - \mu_{I_t}) \right] = E \left[ \sum_{t=1}^{n} \sum_{i=1,2} \mathbb{1}_{\{I_t = i\}} \Delta_i \right] = \sum_{i=1,2} \Delta_i E[T_i(n)]. \tag{3.18} \]

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

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

\[ E_{t,1} = \{ \hat{\mu}_{i^*,T_{i^*}(t-1)} + \alpha H(T_{i^*}(t-1), 1/t) \leq \mu^* \}, \tag{3.19} \]

\[ E_{t,2} = \{ \hat{\mu}_{i,T_i(t-1)} - \alpha H(T_i(t-1), 1/t) > \mu_i \}, \tag{3.20} \]

\[ E_{t,3} = \{ \alpha H(T_i(t-1), 1/t) > \frac{\Delta_i}{2} \}. \tag{3.21} \]

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

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

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

That implies that arm \( i \) would not be chosen. \[ \square \]

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

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

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

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

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

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

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

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

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

which proves the claim. \hfill \blacksquare

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

$$\mathbb{P}[\hat{\mu}_{i,T_i(t-1)} - \alpha \mathcal{H}(T_i(t-1), 1/t) > \mu_i] \leq \mathbb{P} \left[ \bigcup_{s \leq t-1} \{\hat{\mu}_{i,s} - \alpha \mathcal{H}(s, 1/t) > \mu_i\} \right]. \quad (3.22)$$

144
The maximal Azuma-Hoeffding inequality (Theorem 3.52) allows us to bound such a probability without a union bound. We re-write

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

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

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

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

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

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

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

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

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

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

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

\[
= P\left[ \sup_{\gamma^{k-1} \leq s < \gamma^k} M_s \geq \beta \omega(\gamma^{k-1}) \right].
\]
Now we use the maximal Azuma-Hoeffding inequality (Theorem 3.52) to obtain
\[
P\left[ \sup_{\gamma^{k-1} \leq s < \gamma^k} M_s \geq \beta \omega(\gamma^{k-1}) \right] \leq \exp\left( -\frac{2(\beta \omega(\gamma^{k-1}))^2}{\gamma^k} \right)
= \exp\left( -\frac{2\beta^2}{\gamma} \right).
\]

Plugging this back above we get
\[
P\left[ \sup_{1 \leq s < t} \frac{M_s}{\omega(s)} \geq \beta \right] \leq \frac{\log t}{\log \gamma} \exp\left( -\frac{2\beta^2}{\gamma} \right).
\]

Combining (3.22), (3.23), and (3.24) with \( \beta = \alpha \sqrt{\frac{\log t}{2}} \), we obtain
\[
P[\hat{\mu}_{T_i(t-1)} + \alpha H(T_i(t-1), 1/t) > \mu_i] \leq \frac{\log t}{\log \gamma} \exp\left( -\frac{\alpha^2 \log t}{\gamma} \right)
= \frac{\log t}{\log \gamma} \frac{1}{t^{\alpha^2/\gamma}}.
\]

**Lemma 3.81.** For any \( \gamma > 1 \), it holds that
\[P[\mathcal{E}_{t,1}] \leq \frac{1}{\log \gamma} t^{-\alpha^2/\gamma} \log t,\]
and similarly for \( P[\mathcal{E}_{t,2}] \).

We are ready to prove the main result.

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

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

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

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

\[
\mathbb{P}_x[\tau_A < \tau_Z],
\]

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

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

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

\[
\mathbb{P}_x[\tau_A < \tau_Z] = \sum_{y \not\in A \cup Z} P(x, y) \mathbb{P}_y[\tau_A < \tau_Z] + \sum_{y \in A} P(x, y) \cdot 1 + \sum_{y \in Z} P(x, y) \cdot 0
\]

\[
= \sum_y P(x, y) \mathbb{P}_y[\tau_A < \tau_Z].
\]

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

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

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

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

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

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

### 3.3.1 Martingales and the Dirichlet problem

Although the rest of Section 3.3 is concerned with reversible Markov chains, the current subsection applies to the non-reversible case as well. The following definition will be useful below. Let $\sigma$ be a stopping time for a Markov chain $(X_t)$. The Green function of the chain stopped at $\sigma$ is given by

$$G_{\sigma}(x, y) = \mathbb{E}_x \left[ \sum_{0 \leq t < \sigma} 1_{\{X_t = y\}} \right], \quad x, y \in V$$

i.e., it is the expected number of visits to $y$ before $\sigma$ when started at $x$. We will use the notation $h|_Z$ for the function $h$ restricted to the subset $Z$.

**Existence and uniqueness of a harmonic extension** We begin with a general problem.

**Theorem 3.83** (Existence and uniqueness). Let $P$ be an irreducible transition matrix on a finite or countable state space $V$. Let $W$ be a finite, proper subset of $V$ and let $h : W^c \to \mathbb{R}$ be a bounded function on $W^c = V \setminus W$. Then there exists a unique extension of $h$ to $W$ that is harmonic on $W$, i.e., it satisfies

$$h(x) = \sum_y P(x, y)h(y), \quad \forall x \in W. \quad (3.29)$$

The solution is given by $h(x) = \mathbb{E}_x[h(X_{\tau_{W^c}})]$.

**Proof.** We first argue about uniqueness. Suppose $h$ is defined over all of $V$ and satisfies (3.29). Let $\tau^* := \tau_{W^c}$. Then the process $(h(X_{t \wedge \tau^*}))_t$ is a martingale: on \( \{\tau^* \leq t\} \),

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

and on \( \{\tau^* > t\} \)

$$\mathbb{E}[h(X_{(t+1) \wedge \tau^*}) | \mathcal{F}_t] = \sum_y P(X_t, y)h(y) = h(X_t) = h(X_{t \wedge \tau^*}).$$
Because \( W \) is finite and the chain is irreducible, we have \( \tau^* < +\infty \) a.s. See Lemma 3.25. Moreover the process is bounded because \( h \) is bounded on \( W^c \) and \( W \) is finite. Hence by the bounded convergence theorem (or the optional stopping theorem)

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

which implies that \( h \) is unique.

For the existence, simply define

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

and use the Markov property as in (3.27).

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

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

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

Introduce the Laplacian operator on \( N \)

\[
\Delta_N f(x) = \left[ \sum_y P(x, y)f(y) \right] - f(x) = \sum_y P(x, y)[f(y) - f(x)].
\]

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

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

and that solution is given by \( f(x) = \mathbb{E}_x[h(X_{\tau^*_{W^c}})], \) for \( x \in W \cup \partial_N W \). The system (3.30) is called a Dirichlet problem. The Laplacian above can be interpreted as a discretized version of the standard Laplacian. For instance, for simple random walk on \( \mathbb{Z} \) (with \( \pi \equiv 1 \)), \( \Delta_N f(x) = \frac{1}{2}\{[f(x+1) - f(x)] - [f(x) - f(x-1)]\} \)

which is a discretized second derivative.
Applications Before developing the electrical network theory, we point out that Theorem 3.83 has many more applications. One of its consequences is that harmonic functions on finite, connected networks are constant.

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

Proof. Fix the value of $h$ at an arbitrary vertex $z$ and set $W = V \setminus \{z\}$. Applying Theorem 3.83, for all $x \in W$, $h(x) = \mathbb{E}_x[h(X_{\tau_W})] = h(z)$.

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

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

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

does not in fact depend on $x$.

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

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

Substituting into $h(x)$ gives

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

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

Rearranging, we get

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

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

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

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

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

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

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

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

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

(3.31)

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

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

(3.32)

Moreover

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

(3.33)

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

**Proof.** Set $h(x) = v(x)$ on $A \cup Z$. Theorem 3.83 gives the result. ■

151
Let \( v \) be a voltage function on \( \mathcal{N} \) with source \( a \) and sink \( Z \). The Laplacian-based formulation of harmonicity, (3.30), can be interpreted in flow terms as follows. We define the current function \( i(x, y) := c(x, y)[v(x) - v(y)] \) or, equivalently, \( v(x) - v(y) = r(x, y)i(x, y) \). The latter definition is usually referred to as Ohm’s “law.” Notice that the current function is defined on ordered pairs of vertices and is anti-symmetric, i.e., \( i(x, y) = -i(y, x) \). In terms of the current function, the harmonicity of \( v \) is then expressed as

\[
\sum_{y \sim x} i(x, y) = 0, \quad \forall x \in W,
\]
i.e., \( i \) is a flow on \( W \). This set of equations is known as Kirchhoff’s node law. We also refer to these constraints as flow-conservation constraints. To be clear, the current function is not just any flow. It is a flow that can be written as a potential difference according to Ohm’s law. Such a current also satisfies Kirchhoff’s cycle law: if \( x_1 \sim x_2 \sim \cdots \sim x_k \sim x_{k+1} = x_1 \) is a cycle, then

\[
\sum_{j=1}^{k} i(x_j, x_{j+1}) r(x_j, x_{j+1}) = 0,
\]
as can be seen by substituting Ohm’s law. The strength of the current is defined as

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

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

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

**Example 3.88** (Network reduction: birth-and-death chain). Let \( \mathcal{N} \) be the line on \( \{0, 1, \ldots, n\} \) with \( j \sim k \iff |j - k| = 1 \) and arbitrary (positive) conductances
on the edges. Let \((X_t)\) be the corresponding walk. We use the principle above to compute \(\mathbb{P}_x[\tau_0 < \tau_n]\) for \(1 \leq x \leq n - 1\). Consider the voltage function \(v\) when \(v(0) = 1\) and \(v(n) = 0\) with current \(i\). The desired quantity is \(v(x)\) by Corollary 3.86. Note that because \(i\) is a flow on \(\mathcal{N}\), 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 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 \(\mathcal{N}\)

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

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

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

Some special cases:

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

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

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

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

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

**Example 3.89** (Network reduction: binary tree). Let $\mathcal{N}$ be the rooted binary tree with $n$ levels $\mathbb{T}_2^n$ and equal conductances on all edges. Let 0 be the root. Pick an arbitrary leaf and denote it by $n$. The remaining vertices on the path between 0
and \(n\), which we refer to as the main path, will be denoted by \(1, \ldots, n-1\) moving away from the root. We claim that, for all \(0 < x < n\), it holds that
\[
P_x[\tau_0 < \tau_n] = \frac{n-x}{n}.
\]
Indeed let \(v\) be the voltage with values 1 and 0 at \(a=0\) and \(Z = \{z\}\) with \(z = n\) respectively. Let \(i\) be the corresponding current. Notice that, for each \(0 \leq y < n\), the current—as a flow—has nowhere to go on the subtree \(T_y\) hanging from \(y\) away from the main path. The leaves of the subtree are dead ends. Hence the current must be 0 on \(T_y\) and by Ohm’s law the voltage must be constant on it, i.e., every vertex in \(T_y\) has voltage \(v(y)\). Imagine collapsing all vertices in \(T_y\), including \(y\), into a single vertex (and removing the self-loops so created). Doing this for every vertex on the main path results in a new reduced network which is formed of a single path as in Example 3.88. Note that the voltage and the current can be taken to be the same as they were previously on the main path. Indeed, with this choice, Ohm’s law is automatically satisfied. Moreover, because there is no current on the hanging subtrees in the original network, Kirchhoff’s node law is also satisfied on the reduced network, as no current is lost. Hence the answer can be obtained from Example 3.88. That proves the claim. (You should convince yourself that this result is obvious from a probabilistic point of view.)

We gave a probabilistic interpretation of the voltage. What about the current? The following result says that, roughly speaking, \(i(x, y)\) is the net traffic on the edge \(\{x, y\}\) from \(x\) to \(y\). We start with an important formula for the voltage at \(a\). For the walk started at \(a\), we use the shorthand
\[
P[a \to Z] := P_{\tau_Z < \tau_a^+},
\]
for the escape probability.

**Lemma 3.90.** Let \(v\) be a voltage on \(\mathcal{N}\) with source \(a\) and sink \(Z\). Let \(i\) be the associated current. Then
\[
\frac{v(a)}{||i||} = \frac{1}{c(a) P[a \to Z]}. \tag{3.34}
\]
Proof. Using the usual one-step trick,

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

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

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

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

and

\[
i(x, y) = \mathbb{E}_a[N_{x \rightarrow y}^Z - N_{y \rightarrow x}^Z], \quad \forall x \sim y.
\]

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

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

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

\[
\frac{1}{c(x)} \sum_{y \sim x} \mathbb{E}_a[N_{y \rightarrow x}^Z] = \frac{\mathcal{G}_{\tau_Z}(a, x)}{c(x)}. \tag{3.36}
\]
On the other hand,

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

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

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

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

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

(3.37)

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

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

(3.38)

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

Finally, by (3.37),

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

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

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

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

That concludes the proof.

\[ \blacksquare \]

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

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

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

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

**Effective resistance**  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 $P_a[\tau_a^+ < +\infty] < 1$.

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, by Lemma 3.25, $P_a[\tau_{Z_n} \wedge \tau_a^+ = +\infty] = 0$ for all $n$ by our assumptions on $(G_n)$. Hence, the remaining possibilities are

\[
1 = P_a[\exists n, \tau_a^+ < \tau_{Z_n}] + P_a[\forall n, \tau_{Z_n} < \tau_a^+] = P_a[\tau_a^+ < +\infty] + \lim_n P[a \to Z_n].
\]

Therefore $a$ is transient if and only if $\lim_n P[a \to Z_n] > 0$. 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. (Exercise.) Hence we define

\[
P[a \to \infty] := \lim_n P[a \to Z_n] > 0.
\]

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

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

**Definition 3.94 (Effective resistance and conductance).** Let $\mathcal{N} = (G, c)$ be a finite or countable, locally finite, connected network. Let $A = \{a\}$ and $Z$ be disjoint

158
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

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

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

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

$$\mathcal{R}(a \leftrightarrow \infty) := \lim_{n} \mathcal{R}(a \leftrightarrow Z_n),$$

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

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

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

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

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

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

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

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

So $0$ is transient if and only if $p > 1/2$. \hfill \triangle
**Example 3.97** (Biased walk on the $b$-ary tree). Fix $\lambda \in (0, +\infty)$. Consider the rooted, infinite $b$-ary tree with conductance $\lambda^j$ on all edges between level $j - 1$ and $j$, $j \geq 1$. We determine when this chain is transient. Because it is irreducible, all states have the same type and it suffices to look at the root. Denote the root by 0. For an exhaustive sequence, let $G_n$ be the root together with the first $n - 1$ levels. Let $Z_n$ be as before. To compute $R(0 \leftrightarrow Z_n)$:

1) glue together all vertices of $Z_n$;
2) glue together all vertices on the same level of $G_n$;
3) replace parallel edges with a single edge whose conductance is the sum of the conductances; 4) let the current on this edge be the sum of the currents; and 5) leave the voltages unchanged. Note that Ohm’s law and Kirchhoff’s node law are still satisfied. Hence we have not changed the effective resistance. (This is an application of the parallel law.) The reduced network is now a line. Denote the new vertices $0, 1, \ldots, n$. The conductance on the edge between $j$ and $j + 1$ is $b^{j+1}\lambda^j = b(b\lambda)^j$. So this is the chain from the previous example with $(p/q) = b\lambda$ where all conductances are scaled by a factor of $b$. Hence

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

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

### 3.3.3 Bounding the effective resistance

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

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

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

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

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

The unique minimizer is the unit current.

**Proof.** It will be convenient to work in matrix form. Choose an arbitrary orientation of \( \mathcal{N} \), i.e., replace each edge \( \{x, y\} \) with either \( \langle x, y \rangle \) or \( \langle y, x \rangle \). Let \( \overrightarrow{G} \) be the corresponding directed graph. Think of the flow \( \theta \) as a vector with one component for each oriented edge. Then the flow constraint can be written as a linear system \( A\theta = b \). Here the matrix \( A \) has a column for each edge and a row for each vertex except those in \( Z \). The entries of \( A \) are of the form \( A_{x,\langle x, y \rangle} = 1, A_{y,\langle x, y \rangle} = -1 \), and 0 otherwise. The vector \( b \) has 0s everywhere except for \( b_a = 1 \). Let \( r \) be the vector of resistances and let \( R \) be the diagonal matrix with diagonal \( r \). In matrix form the optimization problem (3.39) reads

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

where \( \theta' \) denotes the transpose.

Introduce the Lagrangian

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

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

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

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

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

Substituting into the Lagrangian and simplifying, we have proved that

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

161
This inequality is a statement of weak duality.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

\[
\sum_{y \in V_{\Pi} : y \sim x} \theta(x, y) = \sum_{y \in V : y \sim a} \theta(x, y) = 0, \tag{3.43}
\]

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

\[
\sum_{y \in V_{\Pi} : y \sim a} \theta(a, y) = \sum_{y \in V : y \sim a} \theta(a, y) = \|\theta\|. \tag{3.44}
\]

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

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

as \( x \in Z_{\Pi}, y \in W_{\Pi}, y \sim x \) implies \( \{x, y\} \in \Pi \). That concludes the proof. \( \blacksquare \)

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

\[
\sum_{e \in \Pi_j} c(e) \sum_{e \in \Pi_j} r(e) |\theta(e)|^2 \geq \left( \sum_{e \in \Pi_j} \sqrt{c(e)r(e)} |\theta(e)| \right)^2 \\
= \left( \sum_{e \in \Pi_j} |\theta(e)| \right)^2 \\
\geq 1.
\]

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

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

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

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

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

More generally:

**Corollary 3.105 (Rayleigh’s principle).** Let \( \mathcal{N} \) 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 \( Z \),

\[
\mathcal{R}_{\mathcal{N}}(a \leftrightarrow Z) \leq \mathcal{R}_{\mathcal{N}'}(a \leftrightarrow Z).
\]

**Proof.** Compare the energies of an arbitrary flow on \( \mathcal{N} \) and \( \mathcal{N}' \), and apply Thomson’s principle.

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

Proof. One direction is immediate. Suppose such a flow exists and has energy bounded by $B < +\infty$. Let $(G_n)$ be an exhaustive sequence with associated sinks $(Z_n)$. A unit flow from $a$ to $\infty$ 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 $R(a \leftrightarrow Z_n) \leq B$ for all $n$ and transience follows from Theorem 3.95.

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

$$v_\infty(x) := \lim_n v_n(x),$$

and then

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

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

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

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

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

‡Requires: Section 2.3.3.

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

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

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

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

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

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

$$\mu(e) := P[\langle x, y \rangle \in \Gamma \lor \langle y, x \rangle \in \Gamma] = P[\langle x, y \rangle \in \Gamma] + P[\langle y, x \rangle \in \Gamma] \geq \theta(x, y),$$

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

Claim 3.109 (Method of random paths).

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

(3.45)

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

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

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

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

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

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

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

Plugging these bounds into (3.45), we get

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

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

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

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

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

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

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

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

Example 3.112 (Independent coordinates walk). Let \( \mathcal{N} = \mathbb{L}^d \) with unit resistances and let \( \mathcal{N}' \) be the network on the subset of \( \mathbb{Z}^d \) corresponding to \( (Y_t^{(1)}, \ldots, Y_t^{(d)}) \), where the \( (Y_t^{(i)}) \)'s are independent simple random walks on \( \mathbb{Z} \) started at 0. Note that \( \mathcal{N}' \) contains only those points of \( \mathbb{Z}^d \) with coordinates of identical parities. We claim that the networks \( \mathcal{N} \) and \( \mathcal{N}' \) are roughly equivalent.

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

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

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

Theorem 3.113 (Recurrence and rough equivalence). Let \( \mathcal{N} \) and \( \mathcal{N}' \) be roughly equivalent, locally finite, connected networks. Then \( \mathcal{N} \) is transient if and only if \( \mathcal{N}' \) is transient.
Figure 3.5: The flow on \( \langle x', y' \rangle \) is the sum of the flows on \( \langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \) and \( \langle x_3, y_3 \rangle. \)

**Proof.** Assume \( \mathcal{N} \) is transient and let \( \theta \) be a unit flow from some \( a \) to \( \infty \) of finite energy. The existence of this flow is guaranteed by Theorem 3.106. Let \( \phi, \Phi \) be a rough embedding with parameters \( \alpha \) and \( \beta. \)

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

\[
\theta'(x', y') := \sum_{(x, y):(x', y') \in \overrightarrow{\Phi(x, y)}} \theta(x, y). \tag{3.46}
\]

See Figure 3.5.

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

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

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

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

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

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

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

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

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

Example 3.114 (Independent coordinates walk (continued)). Consider again the networks \( \mathcal{N} \) and \( \mathcal{N}' \) in Example 3.112. Because they are roughly equivalent, they have the same type. This leads to yet another proof of Pólya’s theorem. Recall that, because the number of returns to 0 is geometric with success probability equal to the escape probability, random walk on \( \mathcal{N}' \) is recurrent if and only if the expected number of visits to 0 is finite. By independence of the coordinates, this expectation can be written as
\[
\sum_{t \geq 0} \left( \Pr_0 \left[ Y_{2t}^{(1)} = 0 \right] \right)^d = \sum_{t \geq 0} \left( \frac{2t}{t} \right)^d = \sum_{t \geq 0} \Theta(t^{-d/2}),
\]
where we used Stirling’s formula. The r.h.s. is finite if and only if \( d \geq 3 \). That implies random walk on \( \mathcal{N}' \) is transient under the same condition. By rough equivalence, the same is true of \( \mathcal{N} \).

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

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

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

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

Example 3.116 (Random walk on the torus). Consider random walk on the \(d\)-dimensional torus \(\mathbb{L}_n^d\) 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 Theorem 3.115, note that by symmetry \(\mathbb{E}_x[\tau_y] = \mathbb{E}_y[\tau_x]\) so that

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

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

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

From (3.47), we get:

Claim 3.117.

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

Remark 3.118. The bounds in the previous example are tight up to constants. See [LPW06, Proposition 10.13].
3.3.4 Random walk on supercritical percolation clusters

In this section, we apply the random paths approach to random walk on percolation clusters. To be written. See [LP, Section 5.5].

3.3.5 Uniform spanning trees: Wilson’s method

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

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

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

Claim 3.119.

\[
P[e \in T \mid e' \in T] \leq P[e \in T], \quad \forall e \neq e' \in G.
\]

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

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

\[
P[e \in T] = R(x \leftrightarrow y).
\]

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

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

\[
P[e \in T \mid e' \notin T] \geq P[e \in T],
\]

(3.48)
unless \( \mathbb{P}[e' \in T] \in \{0, 1\} \) or \( \mathbb{P}[e \in T] \in \{0, 1\} \) in which case the claim is vacuous. (In fact these probabilities cannot be 0. Why? Can they be equal to 1?) Picking a uniform spanning tree on \( \mathcal{N} \) conditioned on \( \{e' \notin T\} \) is the same as picking a uniform spanning tree on the modified network \( \mathcal{N}' \) where \( e' \) is removed. By Rayleigh’s principle,

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

and Kirchhoff’s resistance formula gives (3.48).

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

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

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

(3.49)

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

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

(3.50)

Indeed, started at \( x \), \( N^Z_{y \rightarrow x} = 0 \) and \( N^Z_{x \rightarrow y} \in \{0, 1\} \). 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 Exercise 3.11.

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

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

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

To state Wilson’s method, we need the notion of loop erasure. Let $P = x_0 \sim \ldots \sim x_k$ be a path in $\mathcal{N}$. The loop erasure of $P$ is obtained by removing cycles in the order they appear. That is, let $j^*$ be the smallest $j$ such that $x_j = x_\ell$ for some $\ell < j$. Remove the subpath $x_{\ell+1} \sim \ldots \sim x_j$ from $P$, and repeat. The resulting self-avoiding path is denoted by $\text{LE}(P)$.

Let $\rho$ be an arbitrary vertex of $G$, which refer to as the root, and let $T_0$ be the subtree made up of $\rho$ alone. Order arbitrarily the vertices $v_0, \ldots, v_{n-1}$ of $G$, starting with the root $v_0 := \rho$. Wilson’s method constructs an increasing sequence of subtrees as follows. See Figure 3.6. Let $T := T_0$.

1. Let $v$ be the vertex of $G$ not in $T$ with lowest index. Perform random walk on $\mathcal{N}$ started at $v$ until the first visit to a vertex of $T$. Let $P$ be the resulting path.
2. Add the loop erasure \( LE(P) \) to \( T \).

3. Repeat until all vertices of \( G \) are in \( T \).

Let \( T_0, \ldots, T_m \) be the sequence of subtrees produced by Wilson’s method.

**Claim 3.122.** Forgetting the root, \( T_m \) is a uniform spanning tree on \( \mathcal{N} \).

This claim is far from obvious. Before proving it, we finish the proof of Kirchhoff’s resistance formula.

**Proof of Theorem 3.120.** From (3.49) and (3.50), it suffices to prove that, for \( e = \{x, y\} \),

\[
\mathbb{P}_x [\langle x, y \rangle \text{ is traversed before } \tau_y] = \mathbb{P}[e \in T],
\]

where the probability on the l.h.s. refers to random walk on \( \mathcal{N} \) started at \( x \) and the probability on the r.h.s. refers to a uniform spanning tree \( T \) on \( \mathcal{N} \). Generate \( T \) using Wilson’s method started at root \( \rho = y \) with the choice \( v_1 = x \). If the sample path from \( x \) to \( y \) during the first iteration of Wilson’s method includes \( \langle x, y \rangle \), then the loop erasure is simply \( x \sim y \) and \( e \) is in \( T \). On the other hand, if the sample path from \( x \) to \( y \) does not include \( \langle x, y \rangle \), then \( e \) cannot be used at a later stage because it would create a cycle. That immediately proves the claim.

It remains to prove the claim.

**Proof of Claim 3.122.** The idea of the proof is to cast Wilson’s method in the more general framework of cycle popping algorithms. We begin by explaining how such algorithms work.

Let \( P \) be the transition matrix corresponding to random walk on \( \mathcal{N} = (G, c) \) with \( G = (V, E) \). To each vertex \( x \neq \rho \) in \( V \), we assign an independent stack of directed edges

\[
S^x_0 := (\langle x, Y_1^x \rangle, \langle x, Y_2^x \rangle, \ldots)
\]

where each \( Y_j^x \) is chosen independently at random from the distribution \( P(x, \cdot) \). In particular all \( Y_j^x \)s are neighbors of \( x \) in \( \mathcal{N} \). The index \( j \) in \( \langle x, Y_j^x \rangle \) is usually referred to as the color of the edge. It keeps track of the position of the edge in the original stack. (Picture \( S^x \) as a spring-loaded plate dispenser located on vertex \( x \).)

We consider a process which involves popping edges off the stacks. We use the notation \( S^x \) to denote the current stack at \( x \). The initial assignment of the stack is \( S^x := S^x_0 \) as above. Given the current stacks \( (S^x) \), we call visible graph the directed graph over \( V \) with edges \( \text{Top}(S^x) \) for all \( x \neq \rho \), where \( \text{Top}(S^x) \) is the first edge in the current stack \( S^x \). The latter are referred to as visible edges.

We denote the current visible graph by \( \overline{G} \). Note that \( \overline{G} \) has out-degree 1 for
all \( x \neq \rho \) and the root has out-degree 0. In particular all (undirected) cycles in \( \overrightarrow{G} \) are in fact directed cycles. Indeed, a set of edges forming a cycle that is not directed must have a vertex of out-degree 2. Recall the following characterization (see Lemma A.8): a cycle-free undirected graph with \( n \) vertices and \( n - 1 \) edges is a spanning tree. Hence, if there is no cycle in \( \overrightarrow{G} \) then it must be a spanning tree where, furthermore, all edges point towards the root. Such a tree is also known as a spanning arborescence.

As the name suggests, a cycle popping algorithm proceeds by popping cycles in \( \overrightarrow{G} \) off the tops of the stacks until a spanning arborescence is produced. That is, at every iteration, if \( \overrightarrow{G} \) 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 \( \overrightarrow{G} \) is revealed. See Figure 3.7 for an illustration.

With these definitions in place, the proof of the claim involves the following steps.

1. **Wilson’s method is a cycle popping algorithm.** We can think of the initial stacks \( (S^G)_0 \) as corresponding to picking—ahead of time—all potential transitions in the random walks used by Wilson’s method. With this representation, Wilson’s method boils down to a recipe for choosing which cycle to pop next. Indeed, at each iteration, we start from a vertex \( v \) not in the current tree \( T \). Following the visible edges from \( v \) traces a path whose distribution is that of random walk on \( N \). Loop erasure then corresponds to popping cycles. We pop only those visible edges on the removed cycles as they originate from vertices that will be visited again by the algorithm and for which a new transition will then be needed. Those visible edges in the remaining loop-erased path are not popped—they are part of the final arborescence.

2. **The popping order does not matter.** We just argued that Wilson’s method is a cycle popping algorithm. In fact we claim that any cycle popping algorithm, i.e., no matter what popping choices are made along the way, produces the same final arborescence. To make this precise, we identify the popped cycles uniquely. This is where the colors come in. A colored cycle is a directed cycle over \( \overrightarrow{V} \) made of colored edges from the stacks (not necessarily of the same color and not necessarily in the current visible graph). We say that a colored cycle \( \overrightarrow{C} \) is poppable for a visible graph \( \overrightarrow{G} \) if there exists a sequence of colored cycles \( \overrightarrow{C}_1, \ldots, \overrightarrow{C}_r = \overrightarrow{C} \) that can be popped in that order starting from \( \overrightarrow{G} \). Note that, by this definition, \( \overrightarrow{C}_1 \) is a directed cycle in \( \overrightarrow{G} \). Now we claim that if \( \overrightarrow{C}_1' \) were popped first instead of \( \overrightarrow{C}_1 \), producing the new visible graph \( \overrightarrow{G}' \), then \( \overrightarrow{C} \) would still be poppable for \( \overrightarrow{G}' \). This claim implies
Figure 3.7: A realization of a cycle popping algorithm (from top to bottom). In all three figures, the underlying graph is $G$ while the arrows depict the visible edges.
that, in any cycle popping algorithm, either an infinite number of cycles are popped or eventually all poppable cycles are popped—independently of the order—producing the same outcome. To prove the claim, note first that if \( \vec{C}_1' = \vec{C} \) or if \( \vec{C}_1' \) does not share a vertex with any of \( \vec{C}_1, \ldots, \vec{C}_r \), there is nothing to prove. So let \( \vec{C}_j \) be the first cycle in the sequence sharing a vertex with \( \vec{C}_1' \), say \( x \). Let \( \langle x, y \rangle_c \) and \( \langle x, y' \rangle_c' \) be the colored edges emanating from \( x \) in \( \vec{C}_j \) and \( \vec{C}_1' \) respectively. By definition, \( x \) is not on any of \( \vec{C}_1, \ldots, \vec{C}_{j-1} \) 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 \( \vec{C}_j \) and \( \vec{C}_1' \), and the same argument applies to it. Proceeding by induction leads to the conclusion that \( \vec{C}_1' = \vec{C}_j \) as colored cycles. But then \( \vec{C} \) is clearly poppable for the visible graph resulting from popping \( \vec{C}_1' \) first, because it can be popped with the rearranged sequence \( \vec{C}_1' = \vec{C}_j, \vec{C}_1, \ldots, \vec{C}_{j-1}, \vec{C}_{j+1}, \ldots, \vec{C}_r = \vec{C} \), where we used the fact that \( \vec{C}_1' \) does not share a vertex with \( \vec{C}_1, \ldots, \vec{C}_{j-1} \).

3. **Termination occurs in finite time almost surely.** We have shown so far that, in any cycle popping algorithm, either an infinite number of cycles are popped or eventually all poppable cycles are popped. But Wilson’s method—a cycle popping algorithm as we have shown—stops after a finite amount of time with probability 1. Indeed, because the network is finite and connected, the random walk started at each iteration hits the current \( T \) in finite time almost surely (by Lemma 3.25). To sum up, all cycle popping algorithms terminate and produce the same spanning arborescence. It remains to compute the distribution of the outcome.

4. **The arborescence has the desired distribution.** Let \( \mathcal{A} \) be the spanning arborescence produced by any cycle popping algorithm on the stacks \( (S_0^x) \). To compute the distribution of \( \mathcal{A} \), we first compute the distribution of a particular cycle popping realization leading to \( \mathcal{A} \). Because the popping order does not matter, by “realization” we mean a collection \( \mathcal{C} \) of colored cycles together with a final spanning arborescence \( \mathcal{A} \). Notice that what lies in the stacks under \( \mathcal{A} \) is not relevant to the realization, i.e., the same outcome is produced no matter what is under \( \mathcal{A} \). So, from the distribution of the stacks, the probability of observing \( (\mathcal{C}, \mathcal{A}) \) is simply the product of the transitions corresponding to the directed edges in \( \mathcal{C} \) and \( \mathcal{A} \), i.e.,

\[
\prod_{\vec{e} \in \mathcal{C} \cup \mathcal{A}} P(\vec{e}) = \Psi(\mathcal{A}) \prod_{\vec{c} \in \mathcal{C}} \Psi(\vec{c}),
\]

181
where the function $\Psi$ returns the product of the transition probabilities of a set of directed edges. Thanks to the product form on the r.h.s., summing over all possible $C$s gives that the probability of producing $A$ is proportional to $\Psi(A)$. For this argument to work though, there are two small details to take care of. First, note that we want the probability of the uncolored arborescence. But observe that, in fact, there is no need to keep track of the colors on the edges of $A$ because these are determined by $C$. Secondly, we need for the collection of possible $C$s not to vary with $A$. But it is clear that any arborescence could lie under any $C$.

To see that we are done, let $T$ be the undirected spanning tree corresponding to the outcome, $A$, of Wilson’s method. Then, because $P(x, y) = \frac{c(x, y)}{c(x)}$, we get

$$\Psi(A) = \frac{W(T)}{\prod_{x \neq \rho} c(x)},$$

where note that the denominator does not depend on $T$. So if we forget the orientation of $A$, which is determined by the root, we get a spanning tree whose distribution is proportional to $W(T)$, as required.

### 3.3.6 Ising model on trees: the reconstruction problem

To be written. See [Per99, Section 16].

**Exercises**

**Exercise 3.1** (Azuma-Hoeffding: a second proof). This exercise leads the reader through an alternative proof of the Azuma-Hoeffding inequality.

1. Show that for all $x \in [-1, 1]$ and $a > 0$

   $$e^{ax} \leq \cosh a + x \sinh a.$$

2. Use a Taylor expansion to show that for all $x$

   $$\cosh x \leq e^{x^2/2}.$$

3. 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

   $$E[X_{i_1} \cdots X_{i_k}] = 0, \quad \forall 1 \leq k \leq n, \forall 1 \leq i_1 < \cdots < i_k \leq n. \quad (3.51)$$
Show, using a) and b), that for all $b > 0$

$$
\Pr \left[ \sum_{i=1}^{n} X_i \geq b \right] \leq \exp \left( -\frac{b^2}{2\sum_{i=1}^{n} c_i^2} \right).
$$

d) Prove that c) implies the Azuma-Hoeffding inequality as stated in Theorem 3.52.

e) Show that the random variables in Exercise 2.5 do not satisfy (3.51) (without using the claim in part b) of that exercise).

Exercise 3.2 (Kirchhoff’s laws). Consider a finite, connected network with a source and a sink. Show that an anti-symmetric function on the edges satisfying Kirchhoff’s two laws is a current function (i.e., it corresponds to a voltage function through Ohm’s law).

Exercise 3.3 (Dirichlet problem: non-uniqueness). Let $(X_t)$ be the birth-and-death chain on $\mathbb{Z}_+$ with $P(x, x+1) = p$ and $P(x, x-1) = 1-p$ for all $x \geq 1$, and $P(0,1) = 1$, for some $0 < p < 1$. Fix $h(0) = 1$.

a) When $p > 1/2$, show that there is more than one bounded extension of $h$ to $\mathbb{Z}_+ \setminus \{0\}$ that is harmonic on $\mathbb{Z}_+ \setminus \{0\}$. [Hint: Consider $\mathbb{P}_x[\tau_0 = +\infty]$.]

b) When $p \leq 1/2$, show that there exists a unique bounded extension of $h$ to $\mathbb{Z}_+ \setminus \{0\}$ that is harmonic on $\mathbb{Z}_+ \setminus \{0\}$.

Exercise 3.4 (Maximum principle). Let $\mathcal{N} = (G, c)$ be a finite or countable, connected network with $G = (V,E)$. Let $W$ be a finite, connected, proper subset of $V$.

a) Let $h : V \rightarrow \mathbb{R}$ be a function on $V$. Prove the maximum principle: if $h$ is harmonic on $W$, i.e., it satisfies

$$
h(x) = \frac{1}{c(x)} \sum_{y \sim x} c(x,y)h(y), \quad \forall x \in W,
$$

and if $h$ achieves its supremum on $W$, then $h$ is constant on $W \cup \partial_V W$, where

$$
\partial_V W = \{ z \in V \setminus W : \exists y \in W, y \sim z \}.
$$

b) Let $h : W^c \rightarrow \mathbb{R}$ be a bounded function on $W^c := V \setminus W$. Let $h_1$ and $h_2$ be extensions of $h$ to $W$ that are harmonic on $W$. Use part a) to prove that $h_1 \equiv h_2$. 

183
Exercise 3.5 (Effective resistance: metric). Show that effective resistances between pairs of vertices form a metric.

Exercise 3.6 (Dirichlet principle: proof). Prove Theorem 3.100.

Exercise 3.7 (Random walk on $L^2$: effective resistance). Consider random walk on $L^2$, 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 $R(0 \leftrightarrow Z_n) = \Theta(\log n)$. [Hint: Use the Nash-Williams inequality and the method of random paths.]

Exercise 3.8 (Random walk on regular graphs: effective resistance). Let $G$ be a $d$-regular graph with $n$ vertices and $d > n/2$. Let $\mathcal{N}$ be the network $(G, c)$ with unit conductances. Let $a$ and $z$ be arbitrary distinct vertices.

(a) Show that there are at least $2d - n$ vertices $x \neq a, z$ such that $a \sim x \sim z$ is a path.

(b) Prove that

$$R(a \leftrightarrow z) \leq \frac{2rn}{2r - n}.$$ 

Exercise 3.9 (Rough isometries). Graphs $G = (V, E)$ and $G' = (V', E')$ are roughly isometric (or quasi-isometric) if there is a map $\phi : V \to V'$ and constants $0 < \alpha, \beta < +\infty$ such that for all $x, y \in V$

$$\alpha^{-1}d(x, y) - \beta \leq d'(\phi(x), \phi(y)) \leq \alpha d(x, y) + \beta,$$

where $d$ and $d'$ are the graph distances on $G$ and $G'$ respectively, and furthermore all vertices in $G'$ are within distance $\beta$ of the image of $V$. Let $\mathcal{N} = (G, c)$ and $\mathcal{N'} = (G', c')$ be countable, connected networks with uniformly bounded conductances, resistances and degrees. Prove that if $G$ and $G'$ are roughly isometric then $\mathcal{N}$ and $\mathcal{N'}$ are roughly equivalent. [Hint: Start by proving that being roughly isometric is an equivalence relation.]

Exercise 3.10 (Random walk on the cycle: hitting time). Use the commute time identity to compute $E_x[\tau_y]$ in Example 3.116 in the case $d = 1$. Give a second proof using a direct martingale argument.

Exercise 3.11 (Markov chain tree theorem). Let $P$ be the transition matrix of a finite, irreducible Markov chain with stationary distribution $\pi$. Let $G$ be the directed graph corresponding to the positive transitions of $P$. For an arborescence $A$ of $G$, define its weight as

$$\Psi(A) = \prod_{\vec{e} \in A} P(\vec{e}).$$
Consider the following process on spanning arborescences over $G$. Let $\rho$ be the root of the current spanning arborescence $A$. Pick an outgoing edge $\vec{e} = \langle \rho, x \rangle$ of $\rho$ according to $P(\rho, \cdot)$. Add $\vec{e}$ to $A$. This creates a cycle. Remove the edge of this cycle originating from $x$, producing a new arborescence $A'$ with root $x$. Repeat the process.

a) Show that this chain is irreducible.

b) Show that $\Psi$ is a stationary measure for this chain.

c) Prove the Markov chain tree theorem: The stationary distribution $\pi$ of $P$ is proportional to

$$
\pi_x = \sum_{A : \text{root}(A) = x} \Psi(A).
$$

**Bibliographic remarks**

**Section 3.1** The textbooks [Dur10] and [Wil91] contain excellent introductions to martingales. The upper bound in Theorem 3.28 was first proved by Matthews [Mat88]. Section 3.1.4 is based on [Per09, Sections 2 and 3].

**Section 3.2** The Azuma-Hoeffding inequality is due to Hoeffding [Hoe63] and Azuma [Azu67]. The version of the inequality in Exercise 3.1 is from [Ste97]. The method of bounded differences has its origins in the works of Yurinskii [Yur76], Maurey [Mau79], Milman and Schechtman [MS86], Rhee and Talagrand [RT87], and Shamir and Spencer [SS87]. It was popularized by McDiarmid [McD89]. Example 3.63 is taken from [MU05, Section 12.5]. Claim 3.68 is due to Shamir and Spencer [SS87]. The 2-point concentration result alluded to in Section 3.2.3 is due to Alon and Krivelevich [AK97]. For the full story on the chromatic number of Erdős-Rényi graphs, see [JLR11, Chapter 7]. Claim 3.73 is due to Bollobás, Riordan, Spencer, and Tusnády [BRST01]. It confirmed simulations of Barabási and Albert [BA99]. The expectation was analyzed by Dorogovtsev, Mendes, and Samukhin [DMS00]. For much more on preferential attachment models see [Dur06] and [vdH14]. General references on the concentration of measure phenomenon and concentration inequalities are [Led01] and [BLM13]. See [BCB12] for an introduction to bandit problems. The slicing argument in Section 3.2.5 is based on [Bub10]. A more general discussion of the slicing method, whose best known application is the proof of the law of the iterated logarithm (e.g. [Wil91, Section 14.7]), can be found in [vH].
Section 3.3  The material in Sections 3.3.1-3.3.5 borrows heavily from [LPW06, Chapters 9, 10], [AF, Chapters 2, 3] and, especially, [LP, Sections 2.1-2.6, 4.1-4.2, 5,5]. The classical reference on potential theory and its probabilistic counterpart is [Doo01]. For the discrete case and the electrical network point of view, the book of Doyle and Snell is excellent [DS84]. In particular the series and parallel laws are defined and illustrated. See also [KSK76]. For an introduction to convex optimization and duality, see e.g. [BV04]. The Nash-Williams inequality is due to Nash-Williams [NW59]. The result in Example 3.107 is due to R. Lyons [Lyo90]. An elementary proof of Pólya’s theorem can be found in [Dur10, Section 4.2]. The flow we used in the proof of Pólya’s theorem is essentially due to T. Lyons [Lyo83]. Theorem 3.113 is due to Kanai [Kan86]. The commute time identity was proved by Chandra, Raghavan, Ruzzo, Smolensky and Tiwari [CRR+89]. Wilson’s method is due to Wilson [Wil96]. A related method for generating uniform spanning trees was introduced by Aldous [Ald90] and Broder [Bro89]. A connection between loop-erased random walks and uniform spanning trees had previously been established by Pemantle [Pem91] using the Aldous-Broder method. For more on negative correlation in uniform spanning trees, see e.g. [LP, Section 4.2]. For a proof of the matrix tree theorem using Wilson’s method, see [KRS]. For a discussion of the running time of Wilson’s method and other spanning tree generation approaches, see [Wil96].
Chapter 4

Coupling

Coupling is a quintessential probabilistic technique with a wide range of applications. The idea behind the coupling method is that, to compare two probability measures $\mu$ and $\nu$, it is sometimes useful to construct a joint probability space with marginals $\mu$ and $\nu$. For instance, in the classical application of coupling to the convergence of Markov chains (Theorem 1.22), one simultaneously constructs two copies of a Markov chain—one of which is at stationarity—and shows that they can be made to coincide after a random amount of time called the coupling time.

In this chapter, we discuss a number of applications of the coupling method in discrete probability, including applications of the important concept of stochastic domination and its connection to correlation inequalities.

4.1 Background

Throughout this section, we will denote by $\mu_Z$ the law of random variable $Z$.

4.1.1 Basic definitions

A formal definition of coupling follows. Recall that for measurable spaces $(S_1, \mathcal{S}_1)$ $(S_2, \mathcal{S}_2)$, we can consider the product space $(S_1 \times S_2, \mathcal{S}_1 \times \mathcal{S}_2)$ where

$$S_1 \times S_2 := \{(s_1, s_2) : s_1 \in S_1, s_2 \in S_2\}$$

is the Cartesian product of $S_1$ and $S_2$, and $\mathcal{S}_1 \times \mathcal{S}_2$ is the smallest $\sigma$-field $S_1 \times S_2$ containing the rectangles $A_1 \times A_2$ for all $A_1 \in \mathcal{S}_1$ and $A_2 \in \mathcal{S}_2$. 
Definition 4.1 (Coupling). Let $\mu$ and $\nu$ be probability measures on the same measurable space $(S, \mathcal{S})$. A coupling of $\mu$ and $\nu$ is a probability measure $\gamma$ on the product space $(S \times S, \mathcal{S} \times \mathcal{S})$ such that the marginals of $\gamma$ coincide with $\mu$ and $\nu$, i.e.,

$$
\gamma(A \times S) = \mu(A) \quad \text{and} \quad \gamma(S \times A) = \nu(A), \quad \forall A \in \mathcal{S}.
$$

Similarly, for two random variables $X$ and $Y$ taking values in $(S, \mathcal{S})$, a coupling of $X$ and $Y$ is a joint variable $(X', Y')$ taking values in $(S \times S, \mathcal{S} \times \mathcal{S})$ whose law is a coupling of the laws of $X$ and $Y$. Note that, under this definition, $X$ and $Y$ need not be defined on the same probability space—but $X'$ and $Y'$ do need to. We also say that $(X', Y')$ is a coupling of $\mu$ and $\nu$ if the law of $(X', Y')$ is a coupling of $\mu$ and $\nu$.

We give a few examples.

Example 4.2 (Coupling of Bernoulli variables). Let $X$ and $Y$ be Bernoulli random variables with parameters $0 \leq q < r \leq 1$ respectively. That is, $P[X = 0] = 1 - q$ and $P[X = 1] = q$, and similarly for $Y$. Here $S = \{0, 1\}$ and $\mathcal{S} = 2^S$.

- **(Independent coupling)** One coupling of $X$ and $Y$ is $(X', Y')$ where $X' \overset{d}{=} X$ and $Y' \overset{d}{=} Y$ are independent. Its law is

$$
\left( P[(X', Y') = (i, j)] \right)_{i,j \in \{0, 1\}} = \begin{pmatrix}
(1 - q)(1 - r) & (1 - q)r \\
q(1 - r) & qr
\end{pmatrix}.
$$

- **(Monotone coupling)** Another possibility is to pick $U$ uniformly at random in $[0, 1]$, and set $X'' = 1_{\{U \leq q\}}$ and $Y'' = 1_{\{U \leq r\}}$. The law of coupling $(X'', Y'')$ is Then $(X'', Y'')$ is a coupling of $X$ and $Y$ with law

$$
\left( P[(X'', Y'') = (i, j)] \right)_{i,j \in \{0, 1\}} = \begin{pmatrix}
1 - r & r - q \\
0 & q
\end{pmatrix}.
$$

Example 4.3 (Bond percolation: monotonicity). Let $G = (V, E)$ be a countable graph. Denote by $\mathbb{P}_p$ the law of bond percolation on $G$ with density $p$. Let $x \in V$ and assume $0 \leq q < r \leq 1$. Using the coupling in the previous example on each edge independently produces a coupling of $\mathbb{P}_q$ and $\mathbb{P}_r$. More precisely:

- Let $\{U_e\}_{e \in E}$ be independent uniforms on $[0, 1]$.

- For $p \in [0, 1]$, let $W_p$ be the set of edges $e$ such that $U_e \leq p$. 

188
Thinking of $W_p$ as specifying the open edges in the percolation process on $G$ under $\mathbb{P}_p$, we see that $(W_q, W_r)$ is a coupling of $\mathbb{P}_q$ and $\mathbb{P}_r$ with the property that $\mathbb{P}[W_q \subseteq W_r] = 1$. Let $C_x^{(q)}$ and $C_x^{(r)}$ be the open clusters of $x$ under $W_q$ and $W_r$ respectively. Because $C_x^{(q)} \subseteq C_x^{(r)}$,

$$\theta(q) := \mathbb{P}_q[|C_x| = +\infty] = \mathbb{P}_r[|C_x| = +\infty] = \theta(r),$$

as claimed in Section 2.2.4.

**Example 4.4** (Biased random walk on $\mathbb{Z}$). For $p \in [0, 1]$, let $(S_t^{(p)})$ be nearest-neighbor random walk on $\mathbb{Z}$ started at 0 with probability $p$ of jumping to the right and probability $1 - p$ of jumping to the left. Assume $0 \leq q < r \leq 1$. Using again the coupling of Bernoulli variables above we produce a coupling of $S_t^{(q)}$ and $S_t^{(r)}$.

- Let $(X''_i', Y''_i')$ be an infinite sequence of i.i.d. monotone Bernoulli couplings with parameters $q$ and $r$ respectively.

- Define $(Z_i^{(q)}, Z_i^{(r)}) := (2X''_i' - 1, 2Y''_i' - 1)$. Note that $\mathbb{P}[2X''_i' - 1 = 1] = \mathbb{P}[X''_i' = 1] = q$ and $\mathbb{P}[2X''_i' - 1 = -1] = \mathbb{P}[X''_i' = 0] = 1 - q$.

- Let $\hat{S}_t^{(q)} = \sum_{i \leq t} Z_i^{(q)}$ and $\hat{S}_t^{(r)} = \sum_{i \leq t} Z_i^{(r)}$.

Then $(\hat{S}_t^{(q)}, \hat{S}_t^{(r)})$ is a coupling of $(S_t^{(q)}, S_t^{(r)})$ such that $\hat{S}_t^{(q)} \leq \hat{S}_t^{(r)}$ for all $n$ almost surely. So for all $y$ and all $t$

$$\mathbb{P}[S_t^{(q)} \leq y] = \mathbb{P}[\hat{S}_t^{(q)} \leq y] \geq \mathbb{P}[\hat{S}_t^{(r)} \leq y] = \mathbb{P}[S_t^{(r)} \leq y].$$

**4.1.2 Harmonic functions on lattices and trees**

Let $(X_t)$ be a Markov chain on a (finite or) countable state space $V$ with transition matrix $P$ and let $\mathbb{P}_x$ be the law of $(X_t)$ started at $x$.* Recall that a function $h : V \to \mathbb{R}$ is $P$-harmonic on $V$ (or harmonic for short) if

$$h(x) = \sum_{y \in V} P(x, y)h(y), \quad \forall x \in V.$$

We first give a coupling-based criterion for harmonic functions to be constant.

---

*Requires: Section 3.3.1.
Lemma 4.5 (Coupling and bounded harmonic functions). If, for all \( y, z \in V \), there is a coupling \(((Y_t), (Z_t))\) of \( \mathbb{P}_y \) and \( \mathbb{P}_z \) such that

\[
\lim_{t \to \infty} \mathbb{P}[Y_t \neq Z_t] = 0,
\]

then all bounded harmonic functions on \( V \) are constant.

Proof. Let \( h \) be bounded and harmonic on \( V \) with \( \sup_x |h(x)| = M < +\infty \). Let \( y, z \) be any points in \( V \). By harmonicity, \( (h(Y_t)) \) and \( (h(Z_t)) \) are martingales and, in particular,

\[
\mathbb{E}[h(Y_t)] = \mathbb{E}[h(Y_0)] = h(y) \quad \text{and} \quad \mathbb{E}[h(Z_t)] = \mathbb{E}[h(Z_0)] = h(z).
\]

So by Jensen’s inequality and the boundedness assumption

\[
|h(y) - h(z)| = |\mathbb{E}[h(Y_t)] - \mathbb{E}[h(Z_t)]| \leq \mathbb{E}|h(Y_t) - h(Z_t)| \leq 2M \mathbb{P}[Y_t \neq Z_t] \to 0.
\]

So \( h(y) = h(z) \).

Harmonic functions on \( \mathbb{Z}^d \) Consider simple random walk on \( \mathbb{Z}^d \), \( d \geq 1 \). In that case, we show that all bounded harmonic functions are constant.

Theorem 4.6 (Bounded harmonic functions on \( \mathbb{Z}^d \)). All bounded harmonic functions on \( \mathbb{Z}^d \) are constant.

Proof. Clearly, \( h \) is harmonic with respect to simple random walk if and only if it is harmonic with respect to lazy simple random walk. Let \( \mathbb{P}_y \) and \( \mathbb{P}_z \) be the laws of lazy simple random walk on \( \mathbb{Z}^d \) started at \( y \) and \( z \). We construct a coupling \(((Y_t), (Z_t)) \) of \( \mathbb{P}_y \) and \( \mathbb{P}_z \) as follows: at time \( t \), pick a coordinate \( I \in [d] \) uniformly at random, then

- if \( Y_t^{(I)} = Z_t^{(I)} \) then do nothing with probability \( 1/2 \) and otherwise pick \( W \in \{-1, +1\} \) uniformly at random, set \( Y_{t+1}^{(I)} = Z_{t+1}^{(I)} := Z_t^{(I)} + W \) and leave the other coordinates unchanged;

- if instead \( Y_t^{(I)} \neq Z_t^{(I)} \), pick \( W \in \{-1, +1\} \) uniformly at random, and with probability \( 1/2 \) set \( Y_{t+1}^{(I)} := Y_t^{(I)} + W \) and leave \( Z_t \) and the other coordinates of \( Y_t \) unchanged, or otherwise set \( Z_{t+1}^{(I)} := Z_t^{(I)} + W \) and leave \( Y_t \) and the other coordinates of \( Z_t \) unchanged.
It is straightforward to check that \((Y_t, (Z_t))\) is indeed a coupling of \(P_y\) and \(P_z\). To apply the previous lemma, it remains to bound \(P[Y_t \neq Z_t]\).

The key is to note that, for each coordinate \(i\), the difference \((Y_t^{(i)} - Z_t^{(i)})\) is itself a random walk on \(\mathbb{Z}\) started at \(y^{(i)} - z^{(i)}\) with holding probability \(1 - \frac{1}{d}\) until it hits 0. Simple random walk on \(\mathbb{Z}\) is irreducible and recurrent. The holding probability does not affect the type of the walk, as can be seen for instance from the characterization in terms of effective resistance. So \((Y_t^{(i)} - Z_t^{(i)})\) hits 0 in finite time with probability 1. Hence, letting \(\tau^{(i)}\) be the first time \(Y_t^{(i)} - Z_t^{(i)} = 0\), we have \(P[Y_t^{(i)} \neq Z_t^{(i)}] \leq P[\tau^{(i)} > t] \rightarrow P[\tau^{(i)} = +\infty] = 0\).

By a union bound,

\[ P[Y_t \neq Z_t] \leq \sum_{i \in [d]} P[Y_t^{(i)} \neq Z_t^{(i)}] \rightarrow 0, \]

as desired. 

\[ \blacksquare \]

**Harmonic functions on \(T_d\)** On trees, the situation is different. Let \(T_d\) be the infinite \(d\)-regular tree with root \(\rho\). For \(x \in T_d\), we let \(T_x\) be the subtree, rooted at \(x\), of descendants of \(x\).

**Theorem 4.7** (Bounded harmonic functions on \(T_d\)). For \(d \geq 3\), let \((X_t)\) be simple random walk on \(T_d\) and let \(P\) be the corresponding transition matrix. Let \(a\) be a neighbor of the root and consider the function

\[ h(x) = P_x[X_t \in T_a \text{ for all but finitely many } t]. \]

Then \(h\) is a non-constant, bounded \(P\)-harmonic function on \(T_d\).

**Proof.** The function \(h\) is clearly bounded and by the usual one-step trick

\[ h(x) = \sum_{y \sim x} \frac{1}{d} P_y[X_t \in T_0 \text{ for all but finitely many } t] = \sum_y P(x, y)h(y), \]

so \(h\) is \(P\)-harmonic.

Let \(b \neq a\) be a neighbor of the root. The key of the proof is the following lemma.

**Lemma 4.8.**

\[ q := P_a[\tau_\rho = +\infty] = P_b[\tau_\rho = +\infty] > 0. \]
Proof: The second equality follows by symmetry. To see that \( q > 0 \), let \((Z_t)\) be simple random walk on \(\mathbb{T}_d\) started at \(a\) until the walk hits \(0\) and let \(L_t\) be the graph distance between \(Z_t\) and the root. Then \((L_t)\) is a biased random walk on \(\mathbb{Z}\) started at \(1\) jumping to the right with probability \(1 - \frac{1}{d}\) and jumping to the left with probability \(\frac{1}{d}\). The probability that \((L_t)\) hits \(0\) in finite time is \(< 1\) because \(1 - \frac{1}{d} > 2\) when \(d \geq 3\).

Note that

\[
h(\rho) \leq \left(1 - \frac{1}{d}\right) (1 - q) < 1.
\]

Indeed if on the first step the random walk started at \(\rho\) moves away from \(a\), an event of probability \(1 - \frac{1}{d}\), then it must come back to \(\rho\) in finite time to reach \(T_a\). Similarly, by the strong Markov property,

\[
h(a) = q + (1 - q) h(\rho).
\]

Since \(h(\rho) \neq 1\) and \(q > 0\), this shows that \(h(a) > h(\rho)\).

4.2 Coupling inequality

In the examples above, we used coupling to prove monotonicity statements. Coupling is also useful to bound the distance between probability measures.

4.2.1 Bounding the total variation distance via coupling

Let \(\mu\) and \(\nu\) be probability measures on \((S, S)\). Recall the definition of the total variation distance

\[
\|\mu - \nu\|_{TV} := \sup_{A \in S} |\mu(A) - \nu(A)|.
\]

Lemma 4.9 (Coupling inequality). Let \(\mu\) and \(\nu\) be probability measures on \((S, S)\). For any coupling \((X, Y)\) of \(\mu\) and \(\nu\),

\[
\|\mu - \nu\|_{TV} \leq \mathbb{P}[X \neq Y].
\]

Proof. For any \(A \in \mathcal{S}\),

\[
\begin{align*}
\mu(A) - \nu(A) &= \mathbb{P}[X \in A] - \mathbb{P}[Y \in A] \\
&= \mathbb{P}[X \in A, X = Y] + \mathbb{P}[X \in A, X \neq Y] \\
&\quad - \mathbb{P}[Y \in A, X = Y] - \mathbb{P}[Y \in A, X \neq Y] \\
&= \mathbb{P}[X \in A, X \neq Y] - \mathbb{P}[Y \in A, X \neq Y] \\
&\leq \mathbb{P}[X \neq Y],
\end{align*}
\]

192
and, similarly, \( \nu(A) - \mu(A) \leq \mathbb{P}[X \neq Y] \). Hence

\[
|\mu(A) - \nu(A)| \leq \mathbb{P}[X \neq Y].
\]

Here is a simple example.

**Example 4.10** (A coupling of Poisson random variables). Let \( X \sim \text{Poi}(\lambda) \) and \( Y \sim \text{Poi}(\nu) \) with \( \lambda > \nu \). Recall that a sum of independent Poissons is Poisson. This fact leads to a natural coupling: let \( \hat{Y} \sim \text{Poi}(\nu) \), \( \hat{Z} \sim \text{Poi}(\lambda - \nu) \) independently of \( Y \), and \( \hat{X} = \hat{Y} + \hat{Z} \). Then \((\hat{X}, \hat{Y})\) is a coupling and

\[
\|\mu_X - \mu_Y\|_{\text{TV}} \leq \mathbb{P}[\hat{X} \neq \hat{Y}] = \mathbb{P}[\hat{Z} > 0] = 1 - e^{-(\lambda - \nu)} \leq \lambda - \nu.
\]

\[\Box\]

In fact, the inequality in Lemma 4.9 is tight. For simplicity, we prove this in the finite case only.

**Lemma 4.11** (Maximal coupling). Assume \( S \) is finite and let \( S = 2^S \). Let \( \mu \) and \( \nu \) be probability measures on \((S, S)\). Then,

\[
\|\mu - \nu\|_{\text{TV}} = \inf \{\mathbb{P}[X \neq Y] : \text{coupling } (X, Y) \text{ of } \mu \text{ and } \nu\}.
\]

**Proof.** We construct a coupling which achieves equality in the coupling inequality. Such a coupling is called a **maximal coupling**.

Let \( A = \{x \in S : \mu(x) > \nu(x)\} \), \( B = \{x \in S : \mu(x) \leq \nu(x)\} \) and

\[
p := \sum_{x \in S} \mu(x) \wedge \nu(x), \quad \alpha := \sum_{x \in A} [\mu(x) - \nu(x)], \quad \beta := \sum_{x \in B} [\nu(x) - \mu(x)].
\]

First, two lemmas. See Figure 4.1 for a proof by picture.

**Lemma 4.12.**

\[
\sum_{x \in S} \mu(x) \wedge \nu(x) = 1 - \|\mu - \nu\|_{\text{TV}}.
\]
Figure 4.1: Proof by picture that: \( 1 - p = \alpha = \beta = \|\mu - \nu\|_{TV} \).

Proof. We have

\[
2\|\mu - \nu\|_{TV} = \sum_{x \in S} |\mu(x) - \nu(x)|
\]

\[
= \sum_{x \in A} [\mu(x) - \nu(x)] + \sum_{x \in B} [\nu(x) - \mu(x)]
\]

\[
= \sum_{x \in A} \mu(x) + \sum_{x \in B} \nu(x) - \sum_{x \in S} \mu(x) \land \nu(x)
\]

\[
= 2 - \sum_{x \in B} \mu(x) - \sum_{x \in A} \nu(x) - \sum_{x \in S} \mu(x) \land \nu(x)
\]

\[
= 2 - 2 \sum_{x \in S} \mu(x) \land \nu(x).
\]

Lemma 4.13.

\[
\sum_{x \in A} [\mu(x) - \nu(x)] = \sum_{x \in B} [\nu(x) - \mu(x)] = \|\mu - \nu\|_{TV} = 1 - p.
\]

Proof. The first equality is immediate. The second equality follows from the second line in the proof of the previous lemma.

\[
\]

194
The maximal coupling is defined as follows:

- With probability $p$, pick $X = Y$ from $\gamma_{\min}$ where
  \[ \gamma_{\min}(x) := \frac{1}{p} \mu(x) \land \nu(x), \quad x \in S. \]

- Otherwise, pick $X$ from $\gamma_A$ where
  \[ \gamma_A(x) := \frac{\mu(x) - \nu(x)}{1 - p}, \quad x \in A, \]
  and, independently, pick $Y$ from
  \[ \gamma_B(x) := \frac{\nu(x) - \mu(x)}{1 - p}, \quad x \in B. \]

Note that $X \neq Y$ in that case because $A$ and $B$ are disjoint.

The marginal law of $X$ at $x \in S$ is
\[ p \gamma_{\min}(x) + (1 - p) \gamma_A(x) = \mu(x), \]
and similarly for $Y$. Finally $\mathbb{P}[X \neq Y] = 1 - p = \|\mu - \nu\|_{TV}$.

\[ \text{Remark 4.14. A proof of this result for general Polish spaces can be found in [dH, Section 2.5].} \]

We return to our coupling of Bernoulli variables.

**Example 4.15** (Coupling of Bernoulli variables (continued)). Let $X$ and $Y$ be Bernoulli random variables with parameters $0 \leq q < r \leq 1$ respectively. That is, $\mathbb{P}[X = 0] = 1 - q$ and $\mathbb{P}[X = 1] = q$, and similarly for $Y$. Here $S = \{0, 1\}$ and $S = 2^S$. Let $\mu$ and $\nu$ be the laws of $X$ and $Y$ respectively. To construct the maximal coupling as above, we note that
\[ p := \sum_x \mu(x) \land \nu(x) = (1 - r) + q, \quad 1 - p = \alpha = \beta := r - q, \]
\[ A := \{0\}, \quad B := \{1\}, \]
\[ (\gamma_{\min}(x))_{x=0,1} = \left( \frac{1 - r}{1 - r + q}, \frac{q}{1 - r + q} \right), \quad \gamma_A(0) := 1, \quad \gamma_B(1) := 1. \]

The law of the maximal coupling $(X^m, Y^m)$ is given by
\[ \left( \mathbb{P}[(X^m, Y^m) = (i, j)] \right)_{i,j \in \{0,1\}} = \begin{pmatrix} p \gamma_{\min}(0) & (1 - p) \gamma_A(0) \gamma_B(1) \\ 0 & p \gamma_{\min}(1) \end{pmatrix} = \begin{pmatrix} 1 - r & r - q \\ 0 & q \end{pmatrix}, \]
which coincides with the monotone coupling. ▷
Poisson approximation  Here is a classical application of coupling. Let \( X_1, \ldots, X_n \) be independent Bernoulli random variables with parameters \( p_1, \ldots, p_n \) respectively. We are interested in the case where the \( p_i \)s are “small.” Let \( S_n := \sum_{i \leq n} X_i \). We approximate \( S_n \) with a Poisson random variable \( Z_n \) as follows: let \( W_1, \ldots, W_n \) be independent Poisson random variables with means \( \lambda_1, \ldots, \lambda_n \) respectively and define \( Z_n := \sum_{i \leq n} W_i \). We choose \( \lambda_i = -\log(1 - p_i) \) so as to ensure
\[
(1 - p_i) = \mathbb{P}[X_i = 0] = \mathbb{P}[W_i = 0] = e^{-\lambda_i}.
\]
Note that \( Z_n \sim \text{Poi}(\lambda) \) where \( \lambda = \sum_{i \leq n} \lambda_i \).

**Theorem 4.16** (Poisson approximation).
\[
\|\mu_n - \text{Poi}(\lambda)\|_{TV} \leq \frac{1}{2} \sum_{i \leq n} \lambda_i^2.
\]

**Proof.** We couple the pairs \((X_i, W_i)\) independently for \( i \leq n \). Let
\[
W'_i \sim \text{Poi}(\lambda_i) \quad \text{and} \quad X'_i = W'_i \wedge 1.
\]
Because \( \lambda_i = -\log(1 - p_i) \), \((X'_i, W'_i)\) is a coupling of \((X_i, W_i)\). Let \( S'_n := \sum_{i \leq n} X'_i \) and \( Z'_n := \sum_{i \leq n} W'_i \). Then \((S'_n, Z'_n)\) is a coupling of \((S_n, Z_n)\). By the coupling inequality
\[
\|\mu_n - \mu_{Z_n}\|_{TV} \leq \mathbb{P}[S'_n \neq Z'_n] \leq \sum_{i \leq n} \mathbb{P}[X'_i \neq W'_i] = \sum_{i \leq n} \lambda_i \mathbb{P}[W'_i \geq 2]
\]
\[
= \sum_{i \leq n} \sum_{j \geq 2} e^{-\lambda_i} \frac{\lambda_i^j}{j!} \leq \sum_{i \leq n} \frac{\lambda_i^2}{2} \sum_{\ell \geq 0} e^{-\lambda_i} \frac{\lambda_i^{\ell}}{\ell!} = \sum_{i \leq n} \frac{\lambda_i^2}{2}.
\]

Maps reduce total variation distance  The following lemma will be useful.

**Lemma 4.17** (Mappings). Let \( X \) and \( Y \) be random variables taking values in \((S, S)\), let \( h \) be a measurable map from \((S, S)\) to \((S', S')\), and let \( X' := h(X) \) and \( Y' := h(Y) \). It holds that
\[
\|\mu_{X'} - \mu_{Y'}\|_{TV} \leq \|\mu_X - \mu_Y\|_{TV}.
\]

**Proof.** It holds that
\[
\sup_{A' \in S'} |\mathbb{P}[X' \in A'] - \mathbb{P}[Y' \in A']| = \sup_{A' \in S'} |\mathbb{P}[h(X) \in A'] - \mathbb{P}[h(Y) \in A']|
\]
\[
= \sup_{A' \in S'} |\mathbb{P}[X \in h^{-1}(A')] - \mathbb{P}[Y \in h^{-1}(A')]|
\]
\[
= \sup_{A \in S} |\mathbb{P}[X \in A] - \mathbb{P}[Y \in A]|.
\]
4.2.2 Erdős-Rényi graphs: degree sequence

Let \( G_n \sim \mathbb{G}_{n,p} \) be an Erdős-Rényi graph with \( p_n := \frac{\lambda n}{n} \) and \( \lambda > 0 \). For \( i \in [n] \), let \( D_i(n) \) be the degree of vertex \( i \) and define

\[
N_d(n) := \sum_{i=1}^{n} 1\{D_i(n) = d\}.
\]

**Theorem 4.18** (Erdős-Rényi graphs: degree sequence).

\[
\frac{1}{n} N_d(n) \rightarrow_p f_d := e^{-\lambda} \frac{\lambda^d}{d!}, \quad \forall d \geq 1.
\]

**Proof.** We proceed in two steps:

1. we use the coupling inequality (Lemma 4.9) to show that the expectation of \( \frac{1}{n} N_d(n) \) is close to \( f_d \);

2. we use Chebyshev’s inequality (Theorem 2.2) to show that \( \frac{1}{n} N_d(n) \) is close to its expectation.

We prove each step as a lemma below.

**Lemma 4.19** (Convergence of the mean).

\[
\frac{1}{n} \mathbb{E}_{n,p} \left[ N_d(n) \right] \rightarrow f_d, \quad \forall d \geq 1.
\]

**Proof.** Note that the \( D_1(n) \)s are identically distributed (but not independent) so

\[
\frac{1}{n} \mathbb{E}_{n,p} \left[ N_d(n) \right] = \mathbb{P}_{n,p} \left[ D_1(n) = d \right].
\]

Moreover \( D_1(n) \sim \text{Bin}(n - 1, p_n) \). Let \( S_n \sim \text{Bin}(n, p_n) \) and \( Z_n \sim \text{Poi}(\lambda) \). By the Poisson approximation

\[
\| \mu_{S_n} - \mu_{Z_n} \|_{TV} \leq \frac{1}{2} \sum_{i \leq n} \left( - \log(1 - p_n) \right)^2 = \frac{1}{2} \sum_{i \leq n} \left( \frac{\lambda}{n} + O(n^{-2}) \right)^2 = \frac{\lambda^2}{2n} + O(n^{-2}).
\]

We can couple \( D_1(n) \) and \( S_n \) as \( \sum_{i \leq n-1} X_i, \sum_{i \leq n} X_i \) where the \( X_i \)s are i.i.d. Bernoulli with parameter \( \frac{\lambda}{n} \). By the coupling inequality

\[
\| \mu_{D_1(n)} - \mu_{S_n} \|_{TV} \leq \mathbb{P} \left[ \sum_{i \leq n-1} X_i \neq \sum_{i \leq n} X_i \right] = \mathbb{P}[X_n = 1] = \frac{\lambda}{n}.
\]

By the triangle inequality for total variation distance,

\[
\frac{1}{2} \sum_{d \geq 0} | \mathbb{P}_{n,p_n} [ D_1(n) = d ] - f_d | \leq \frac{\lambda + \lambda^2/2}{n} + O(n^{-2}).
\]
Therefore,
\[
\left| \frac{1}{n} E_{n,p_n} [N_d(n)] - f_d \right| \leq \frac{2\lambda + \lambda^2}{n} + O(n^{-2}) \to 0.
\]

Lemma 4.20 (Concentration around the mean).
\[
P_{n,p_n} \left[ \left| \frac{1}{n} N_d(n) - \frac{1}{n} E_{n,p_n} [N_d(n)] \right| \geq \varepsilon \right] \leq \frac{2\lambda + 1}{\varepsilon^2 n}, \quad \forall d \geq 1, \forall n.
\]

Proof. By Chebyshev’s inequality, for all \(\varepsilon > 0\)
\[
P_{n,p_n} \left[ \left| \frac{1}{n} N_d(n) - \frac{1}{n} E_{n,p_n} [N_d(n)] \right| \geq \varepsilon \right] \leq \frac{\text{Var}_{n,p_n} \left[ \frac{1}{n} N_d(n) \right]}{\varepsilon^2}.
\]

Note that
\[
\text{Var}_{n,p_n} \left[ \frac{1}{n} N_d(n) \right]
\]
\[
= \frac{1}{n^2} \left\{ E_{n,p_n} \left[ \left( \sum_{i \leq n} 1 \{ D_i(n) = d \} \right)^2 \right] - n \, E_{n,p_n} [D_1(n) = d]^2 \right\}
\]
\[
= \frac{1}{n^2} \left\{ n(n-1) \, \text{P}_{n,p_n} [D_1(n) = d, D_2(n) = d]
\]
\[
+ n \, \text{P}_{n,p_n} [D_1(n) = d] - n^2 \, \text{P}_{n,p_n} [D_1(n) = d]^2 \right\}
\]
\[
\leq \frac{1}{n} + \left\{ \text{P}_{n,p_n} [D_1(n) = d, D_2(n) = d] - \text{P}_{n,p_n} [D_1(n) = d]^2 \right\}.
\]

We bound the second term using a neat coupling argument. Let \(Y_1\) and \(Y_2\) be independent \(\text{Bin}(n-2, p_n)\) and let \(X_1\) and \(X_2\) be independent \(\text{Ber}(p_n)\). Then the term in curly bracket above is equal to
\[
P[(X_1 + Y_1, X_1 + Y_2) = (d, d)] - P[(X_1 + Y_1, X_2 + Y_2) = (d, d)]
\]
\[
\leq P[(X_1 + Y_1, X_1 + Y_2) = (d, d), (X_1 + Y_1, X_2 + Y_2) \neq (d, d)]
\]
\[
= P[(X_1 + Y_1, X_1 + Y_2) = (d, d), X_2 + Y_2 \neq d]
\]
\[
= \text{P}[X_1 = 0, Y_1 = Y_2 = d, X_2 = 1] + \text{P}[X_1 = 1, Y_1 = Y_2 = d - 1, X_2 = 0]
\]
\[
\leq \frac{2\lambda}{n}.
\]

So \(\text{Var}_{n,p_n} \left[ \frac{1}{n} N_d(n) \right] \leq \frac{2\lambda + 1}{n}.
\]

Combining the lemmas concludes the proof of Theorem 4.18.

198
4.3 Stochastic domination

In comparing two probability measures, a natural relationship to consider is the notion of “domination.” To see why this might be useful, let \((X_i)_{i=1}^n\) be independent \(\mathbb{Z}_+\)-valued random variables with
\[
\mathbb{P}[X_i \geq 1] \geq p,
\]
and let \(S = \sum_{i=1}^n X_i\) be their sum. Consider also a random variable
\[
S_* \sim \text{Bin}(n, p).
\]
Then it is intuitively clear that one should be able to obtain some bounds on \(S\) by studying \(S_*\) instead—which may be easier. Indeed, in some sense, \(S\) “dominates” \(S_*\), that is, \(S\) should have a tendency to be bigger than \(S_*\). One expects more specifically that
\[
\mathbb{P}[S > x] \geq \mathbb{P}[S_* > x].
\]
Coupling provides a handy characterization of this notion, as we detail in this section.

In particular we study an important special case known as positive association. In that case a measure “dominates itself” in the following sense: conditioning on certain events makes other events more likely. This concept, which is formalized in Section 4.3.4, has numerous applications in discrete probability.

4.3.1 Definitions

We start with the case of real random variables.

**Ordering of real random variables** For real random variables, stochastic domination is defined as follows. See Figure 4.2 for an illustration.

**Definition 4.21** (Stochastic domination). Let \(\mu\) and \(\nu\) be probability measures on \(\mathbb{R}\). The measure \(\mu\) is said to stochastically dominate \(\nu\), denoted \(\mu \succeq \nu\), if for all \(x \in \mathbb{R}\)
\[
\mu[(x, +\infty)] \geq \nu[(x, +\infty)].
\]
A real random variable \(X\) stochastically dominates \(Y\), denoted by \(X \succeq Y\), if the law of \(X\) dominates the law of \(Y\).

**Example 4.22** (Bernoulli vs. Poisson). Let \(X \sim \text{Poi}(\lambda)\) be Poisson with mean \(\lambda > 0\) and let \(Y\) be a Bernoulli trial with success probability \(p \in (0, 1)\), i.e.,
Figure 4.2: The law of $X$, represented here by its cumulative distribution function $F_X$ in red, stochastically dominates the law of $Y$, in orange. The construction of a monotone coupling, $(\hat{X}, \hat{Y}) := (F_X^{-1}(U), F_Y^{-1}(U))$ where $U$ is uniform in $[0, 1]$, is also depicted.
\( \mathbb{P}[Y = 1] = 1 - \mathbb{P}[Y = 0] = p \). In order for \( X \) to stochastically dominate \( Y \), it suffices to have
\[
\mathbb{P}[X > \ell] \geq \mathbb{P}[Y > \ell], \quad \forall \ell \geq 0.
\]
This is always true for \( \ell \geq 1 \) since \( \mathbb{P}[X > \ell] > 0 \) but \( \mathbb{P}[Y > \ell] = 0 \). So it remains to consider the case \( \ell = 0 \). We have
\[
1 = e^{-\lambda} = \mathbb{P}[X > 0] \geq \mathbb{P}[Y > 0] = p,
\]
if and only if
\[
\lambda \geq -\log(1 - p).
\]
Note that stochastic domination does not require \( X \) and \( Y \) to be defined on the same probability space. The connection to coupling arises from the following characterization.

**Theorem 4.23** (Coupling and stochastic domination). The real random variable \( X \) stochastically dominates \( Y \) if and only if there is a coupling \((\hat{X}, \hat{Y})\) of \( X \) and \( Y \) such that
\[
\mathbb{P}[\hat{X} \geq \hat{Y}] = 1. \tag{4.1}
\]
We refer to \((\hat{X}, \hat{Y})\) as a monotone coupling of \( X \) and \( Y \).

**Proof.** One direction is clear. Suppose there is such a coupling. Then for all \( x \in \mathbb{R} \)
\[
\mathbb{P}[Y > x] = \mathbb{P}[\hat{Y} > x] = \mathbb{P}[\hat{X} \geq \hat{Y} > x] \leq \mathbb{P}[\hat{X} > x] = \mathbb{P}[X > x].
\]
For the other direction, define the cumulative distribution functions \( F_X(x) = \mathbb{P}[X \leq x] \) and \( F_Y(x) = \mathbb{P}[Y \leq x] \). Assume \( X \succeq X' \). The idea of the proof is to use the following standard way of generating a real random variable. Recall (e.g. [Dur10, Section 1.2]) that
\[
X \overset{d}{=} F_X^{-1}(U), \tag{4.2}
\]
where \( U \) is a \([0, 1]\)-valued uniform random variable and
\[
F_X^{-1}(u) := \inf\{x \in \mathbb{R} : F_X(x) \geq u\},
\]
is a generalized inverse. Indeed \( F_X^{-1}(u) > x \) precisely when \( u > F_X(x) \). It is natural to construct a coupling of \( X \) and \( Y \) by simply using the same uniform random variable \( U \) in this representation, i.e., we define \( \hat{X} = F_X^{-1}(U) \) and \( \hat{Y} = F_Y^{-1}(U) \). See Figure 4.2. By (4.2), this is a coupling of \( X \) and \( Y \). It remains
to check (4.1). Because \( F_X(x) \leq F_Y(x) \) for all \( x \) by definition of stochastic domination, by the definition of the generalized inverse,

\[
P[\hat{X} \geq \hat{Y}] = P[F_X^{-1}(U) \geq F_Y^{-1}(U)] = 1,
\]
as required.

**Example 4.24.** Returning to the example in the first paragraph of Section 4.3, let \((X_i)_{i=1}^n\) be independent \( \mathbb{Z}_+ \)-valued random variables with \( P[X_i \geq 1] \geq p \) and consider their sum \( S := \sum_{i=1}^n X_i \). Further let \( S_\ast \sim \text{Bin}(n, p) \). Write \( S_\ast \) as the sum \( \sum_{i=1}^n Y_i \) where \((Y_i)\) are independent \( \{0,1\}\)-variables with \( P[Y_i = 1] = p \). To couple \( S \) and \( S_\ast \), first set \( \hat{Y}_i := (Y_i) \) and \( \hat{S}_\ast := \sum_{i=1}^n \hat{Y}_i \). Let \( \hat{X}_i = 0 \) whenever \( \hat{Y}_i = 0 \). Otherwise, i.e. if \( \hat{Y}_i = 1 \), generate \( \hat{X}_i \) according to the distribution of \( X_i \) conditioned on \( \{X_i \geq 1\} \), independently of everything else. By construction \( \hat{X}_i \geq \hat{Y}_i \) a.s. for all \( i \) and as a result \( \sum_{i=1}^n \hat{X}_i =: \hat{S} \geq \hat{S}_\ast \) a.s. or \( S \geq S_\ast \) by the previous theorem. That implies for instance that \( P[S > x] \geq P[S_\ast > x] \) as we claimed earlier. A special case of this argument gives the following useful fact about binomials

\[
n \geq m, q \geq p \implies \text{Bin}(n, q) \succeq \text{Bin}(m, p).
\]

**Example 4.25 (Poisson distribution).** Let \( X \sim \text{Poi}(\mu) \) and \( Y \sim \text{Poi}(\nu) \) with \( \mu > \nu \). Recall that a sum of independent Poissons is Poisson (use moment-generating functions or see e.g. [Dur10, Exercise 2.1.14]). This fact leads to a natural coupling: let \( \hat{Y} \sim \text{Poi}(\nu) \), \( \hat{Z} \sim \text{Poi}(\mu - \nu) \) independently of \( Y \), and \( \hat{X} = \hat{Y} + \hat{Z} \). Then \( (\hat{X}, \hat{Y}) \) is a coupling and \( \hat{X} \geq \hat{Y} \) a.s. because \( \hat{Z} \geq 0 \). Hence \( X \succeq Y \).

We record two useful consequences of Theorem 4.23.

**Corollary 4.26.** Let \( X \) and \( Y \) be real random variables with \( X \succeq Y \) and let \( f : \mathbb{R} \to \mathbb{R} \) be a non-decreasing function. Then \( f(X) \succeq f(Y) \) and furthermore, provided \( \mathbb{E}|f(X)|, \mathbb{E}|f(Y)| < +\infty \), we have that

\[
\mathbb{E}[f(X)] \geq \mathbb{E}[f(Y)].
\]

**Proof.** Let \((\hat{X}, \hat{Y})\) be the monotone coupling of \( X \) and \( Y \) whose existence is guaranteed by Theorem 4.23. Then \( f(\hat{X}) \succeq f(\hat{Y}) \) a.s. so that, provided the expectations exist,

\[
\mathbb{E}[f(X)] = \mathbb{E}[f(\hat{X})] \geq \mathbb{E}[f(\hat{Y})] = \mathbb{E}[f(Y)],
\]

and furthermore \((f(\hat{X}), f(\hat{Y}))\) is a monotone coupling of \( f(X) \) and \( f(Y) \). Hence \( f(X) \succeq f(Y) \).
Corollary 4.27. Let $X_1$, $X_2$ be independent random variables. Let $Y_1$, $Y_2$ be independent random variables such that $X_i \succeq Y_i$, $i = 1, 2$. Then

$$X_1 + X_2 \succeq Y_1 + Y_2.$$ 

Proof. Let $(\hat{X}_1, \hat{Y}_1)$ and $(\hat{X}_2, \hat{Y}_2)$ be independent, monotone couplings of $(X_1, Y_1)$ and $(X_2, Y_2)$ (on the same probability space). Then

$$X_1 + X_2 \sim \hat{X}_1 + \hat{X}_2 \geq \hat{Y}_1 + \hat{Y}_2 \sim Y_1 + Y_2.$$ 

Example 4.28 (Binomial vs. Poisson). A sum of $n$ Poisson variables with mean $\lambda$ is Poi($n\lambda$). A sum of $n$ Bernoulli trials with success probability $p$ is Bin($n, p$). Using Example 4.22 and Corollary 4.27, we get

$$\lambda \geq -\log(1 - p) \implies \text{Poi}(n\lambda) \succeq \text{Bin}(n, p).$$  

The following special case will be useful later. Let $0 < \Lambda < 1$ and let $m$ be an integer. Then

$$\frac{\Lambda}{m-1} \geq \frac{\Lambda}{m-\Lambda} = \frac{m}{m-\Lambda} - 1 \geq \log \left( \frac{m}{m-\Lambda} \right) = -\log \left( 1 - \frac{\Lambda}{m} \right),$$ 

where we used that $\log x \leq x - 1$ for all $x \in \mathbb{R}$. So, setting $\lambda := \frac{\Lambda}{m-1}$, $p := \frac{\Lambda}{m}$ and $n := m - 1$ in (4.3), we get

$$\Lambda \in (0, 1) \implies \text{Poi}(\Lambda) \succeq \text{Bin} \left( m - 1, \frac{\Lambda}{m} \right).$$  

Ordering on partially ordered sets The definition of stochastic domination hinges on the totally ordered nature of $\mathbb{R}$. It also extends naturally to posets. Let $(\mathcal{X}, \leq)$ be a poset, i.e., for all $x, y, z \in \mathcal{X}$:

- [Reflexivity] $x \leq x$,
- [Antisymmetry] if $x \leq y$ and $y \leq x$ then $x = y$,
- [Transitivity] if $x \leq y$ and $y \leq z$ then $x \leq z$. 

poset

203
For instance the set \( \{0, 1\}^F \) is a poset when equipped with the relation \( x \leq y \) if and only if \( x_i \leq y_i \) for all \( i \in F \). Equivalently the subsets of \( F \), denoted by \( 2^F \), form a poset with the inclusion relation. (A totally ordered set satisfies in addition that, for any \( x, y \), we have either \( x \leq y \) or \( y \leq x \).)

Let \( \mathcal{F} \) be a \( \sigma \)-field over the poset \( \mathcal{X} \). An event \( A \in \mathcal{F} \) is increasing if \( x \in A \) implies that any \( y \geq x \) is also in \( A \). A function \( f : \mathcal{X} \to \mathbb{R} \) is increasing if \( x \leq y \) implies \( f(x) \leq f(y) \).

**Definition 4.29** (Stochastic domination for posets). Let \( (\mathcal{X}, \leq) \) be a poset and let \( \mathcal{F} \) be a \( \sigma \)-field on \( \mathcal{X} \). Let \( \mu \) and \( \nu \) be probability measures on \( (\mathcal{X}, \mathcal{F}) \). The measure \( \mu \) is said to stochastically dominate \( \nu \), denoted by \( \mu \geq \nu \), if for all increasing \( A \in \mathcal{F} \):

\[
\mu(A) \geq \nu(A).
\]

An \( \mathcal{X} \)-valued random variable \( X \) stochastically dominates \( Y \), denoted by \( X \succeq Y \), if the law of \( X \) dominates the law of \( Y \).

As before, a monotone coupling \( (\hat{X}, \hat{Y}) \) of \( X \) and \( Y \) is one which satisfies \( \hat{X} \geq \hat{Y} \) a.s.

**Example 4.30** (Monotonicity of the percolation function). We have already seen an example of stochastic domination in Section 2.2.4. Consider bond percolation on the \( d \)-dimensional lattice \( \mathbb{L}^d \). Here the poset is the collection of all subsets of edges, specifying the open edges, with the inclusion relation. Recall that the percolation function is given by

\[
\theta(p) := \mathbb{P}_p[|C_0| = +\infty],
\]

where \( C_0 \) is the open cluster of the origin. We argued in Section 2.2.4 that \( \theta(p) \) is non-decreasing by considering the following alternative representation of the percolation process under \( \mathbb{P}_p \): to each edge \( e \), assign a uniform \( [0, 1] \)-valued random variable \( U_e \) and declare the edge open if \( U_e \leq p \). Using the same \( U_e \)'s for two different \( p \)-values, \( p_1 < p_2 \), gives a monotone coupling of the processes for \( p_1 \) and \( p_2 \). It follows immediately that \( \theta(p_1) \leq \theta(p_2) \), where we used that the event \( \{|C_0| = +\infty\} \) is increasing.

The existence of a monotone coupling is perhaps more surprising for posets. We prove the result in the finite case only, which will be enough for our purposes.

**Theorem 4.31** (Strassen’s theorem). Let \( X \) and \( Y \) be random variables taking values in a finite poset \( (\mathcal{X}, \leq) \) with the \( \sigma \)-field \( \mathcal{F} = 2^\mathcal{X} \). Then \( X \succeq Y \) if and only if there exists a monotone coupling \( (\hat{X}, \hat{Y}) \) of \( X \) and \( Y \).
Proof. One direction is clear. Suppose there is such a coupling. Then for all increasing $A$

$$P[Y \in A] = P[\hat{Y} \in A] = P[\hat{X} \geq \hat{Y} \in A] \leq P[\hat{X} \in A] = P[X \in A].$$

The proof in the other direction relies on the max-flow min-cut theorem (Theorem 1.9). To see the connection with flows, let $\mu_X$ and $\mu_Y$ be the laws of $X$ and $Y$ respectively, and denote by $\nu$ their joint distribution under the desired coupling.

Noting that we want $\nu(x, y) = 0$ if $x \leq y$, the marginal conditions on the coupling read

$$\sum_{y \leq x} \nu(x, y) = \mu_X(x), \quad \forall x \in \mathcal{X},$$

and

$$\sum_{x \geq y} \nu(x, y) = \mu_Y(y), \quad \forall y \in \mathcal{X}.$$

These equations can be interpreted as flow-conservation constraints. Consider the following directed graph. There are two vertices, $(w, 1)$ and $(w, 2)$, for each element $w$ in $\mathcal{X}$ with edges connecting each $(x, 1)$ to those $(y, 2)$s with $x \geq y$. These edges have capacity $+\infty$. In addition there is a source $a$ and a sink $z$. The source has a directed edge of capacity $\mu_X(x)$ to $(x, 1)$ for each $x \in \mathcal{X}$ and, similarly, each $(y, 2)$ has a directed edge of capacity $\mu_Y(y)$ to the sink. The existence of a monotone coupling will follow once we show that there is a flow of strength 1 between $a$ and $z$. Indeed, in that case, all edges from the source and all edges to the sink are at capacity. If we let $\nu(x, y)$ be the flow on edge $((x, 1), (y, 2))$, the constraints above then impose the conservation of the flow on the vertices $(\mathcal{X} \times \{1\}) \cup (\mathcal{X} \times \{2\})$. Hence the flow between $\mathcal{X} \times \{1\}$ and $\mathcal{X} \times \{2\}$ yields the desired coupling. See Figure 4.3.

By the max-flow min-cut theorem (Theorem 1.9), it suffices to show that a minimum cut has capacity 1. Such a cut is of course obtained by choosing all edges out of the source. So it remains to show that no cut has capacity less than 1. This is where we use the fact that $\mu_X(A) \geq \mu_Y(A)$ for all increasing $A$. Because the edges between $\mathcal{X} \times \{1\}$ and $\mathcal{X} \times \{2\}$ have infinite capacity, they cannot be used in a minimum cut. So we can restrict our attention to those cuts containing edges from $a$ to $A_* \times \{1\}$ and from $Z_* \times \{2\}$ to $z$ for subsets $A_*, Z_* \subseteq \mathcal{X}$. We must have

$$A_* \supseteq \{x \in \mathcal{X} : \exists y \in Z_*^c, x \geq y\},$$

to block all paths of the form $a \sim (x, 1) \sim (y, 2) \sim z$ with $x$ and $y$ as above. In fact, for a minimum cut, we further have

$$A_* = \{x \in \mathcal{X} : \exists y \in Z_*^c, x \geq y\},$$

205
Figure 4.3: Construction of a monotone coupling through the max-flow representation for independent Bernoulli pairs with parameters \( r \) (on the left) and \( q < r \) (on the right). Edge labels indicate capacity. Edges without labels have infinite capacity. The colored edges depict a suboptimal cut. The blue and orange vertices correspond respectively to the sets \( A_\ast \) and \( Z_\ast \) for this cut. The capacity of the cut is \( r^2 + r(1-r) + (1-q)^2 + (1-q)q = r + (1-q) > r + (1-r) = 1 \).
as adding an \( x \) not satisfying this property is redundant. See Figure 4.3. In particular \( A_s \) is increasing: if \( x_1 \in A_s \) and \( x_2 \geq x_1 \), then \( \exists y \in Z_c^x \) such that \( x_1 \geq y \) and, since \( x_2 \geq x_1 \geq y \), the same \( y \) works for \( x_2 \). Observe that, because \( y \geq y \), the set \( A_s \) also includes \( Z_c^x \). If it were the case that \( A_s \neq Z_c^x \), then we could construct a cut with lower or equal capacity by fixing \( A_s \) and setting \( Z_c^x := A_c^x \); because \( A_s \) is increasing, any \( y \in A_s \cap Z_c^x \) is such that paths of the form \( a \sim (x, 1) \sim (y, 2) \sim z \) with \( x \geq y \) are cut by \( x \in A_s \). Hence, for a minimum cut, we can assume that in fact \( A_s = Z_c^x \). The capacity of the cut is

\[
\mu_X(A_s) + \mu_Y(Z_s) = \mu_X(A_s) + 1 - \mu_Y(A_s) = 1 + (\mu_X(A_s) - \mu_Y(A_s)) \geq 1,
\]

where the term in parenthesis is nonnegative by assumption and the fact that \( A_s \) is increasing. That concludes the proof.

**Remark 4.32.** Strassen’s theorem holds more generally on Polish spaces with a closed partial order. See e.g. [Lin02, Section IV.1.2] for the details.

The proof of Corollary 4.26 immediately extends to:

**Corollary 4.33.** Let \( X \) and \( Y \) be \( \mathcal{X} \)-valued random variables with \( X \succeq Y \) and let \( f : \mathcal{X} \to \mathbb{R} \) be an increasing function. Then \( f(X) \succeq f(Y) \) and furthermore, provided \( \mathbb{E}|f(X)|, \mathbb{E}|f(Y)| < +\infty \), we have that

\[
\mathbb{E}[f(X)] \geq \mathbb{E}[f(Y)].
\]

**Ordering of Markov chains** Stochastic domination also arises in the context of Markov chains. We begin with an example.

**Example 4.34 (Lazier chain).** Consider a random walk \((X_t)\) on the network \( \mathcal{N} = ((V, E), c) \) where \( V = \{0, 1, \ldots, n\} \) and \( i \sim j \) if and only if \( |i - j| \leq 1 \) (including self-loops). Let \( \mathcal{N}' = ((V, E), c') \) be a modified version of \( \mathcal{N} \) on the same graph where for all \( i \) \( c(i, i') \leq c'(i, i) \). That is, if \((X'_t)\) is random walk on \( \mathcal{N}' \), then \((X'_t)\) is “lazier” than \((X_t)\) in that it is more likely to stay put. Assume that both \((X_t)\) and \((X'_t)\) start at \( i_0 \) and define \( M_s := \max_{t \leq s} X_t \) and \( M'_s := \max_{t \leq s} X'_t \). Since \((X'_t)\) “travels less” than \((X_t)\) the following claim is intuitively obvious:

**Claim 4.35.**

\[
M_s \succeq M'_s.
\]

We prove this by producing a monotone coupling. First set \((\hat{X}_t) := (X_t)\). We then generate \((\hat{X}'_t)\) as a “sticky” version of \((\hat{X}_t)\). That is, \((\hat{X}'_t)\) follows exactly the same
transitions as \((\hat{X}_t)\) (including the self-loops), but at each time it opts to stay where it currently is, say \(j\), for an extra time step with probability
\[
\frac{c'(j, j) - c(j, j)}{\sum_{i \sim j} c'(i, j)},
\]
which is in \([0, 1]\) by assumption. Marginally, \((\hat{X}'_t)\) is a random walk on \(\mathcal{X}'\) because by construction
\[
c'(j, j) = \sum_{i \sim j} c(i, j) c'(i, j),
\]
and for \(i \neq j\) with \(i \sim j\)
\[
\frac{c'(i, j)}{\sum_{i \sim j} c'(i, j)} = \left(\frac{\sum_{i \sim j} c(i, j)}{\sum_{i \sim j} c'(i, j)}\right) \frac{c(i, j)}{\sum_{i \sim j} c(i, j)},
\]
since \(c'(i, j) = c(i, j)\). This coupling satisfies
\[
\hat{M}_s := \max_{t \leq s} \hat{X}_t \geq \max_{t \leq s} \hat{X}'_t =: \hat{M}'_s, \quad \text{a.s.}
\]
because \((\hat{X}'_t)_{t \leq s}\) visits a subset of the states visited by \((\hat{X}_t)_{t \leq s}\). In other words \((\hat{M}_s, \hat{M}'_s)\) is a monotone coupling of \((M_s, M'_s)\) and this proves the claim. \(\blacksquare\)

The previous example involved an asynchronous coupling of the chains. Often, a simpler step-by-step approach is possible.

**Definition 4.36** (Stochastic domination for Markov kernels). Let \(P\) and \(Q\) be transition matrices on a finite or countable poset \((\mathcal{X}, \leq)\). The transition matrix \(Q\) is said to stochastically dominate the transition matrix \(P\) if
\[
x \leq y \implies P(x, \cdot) \preceq Q(y, \cdot). \tag{4.5}
\]
If the above condition is satisfied for \(P = Q\), we say that \(P\) is stochastically monotone.

The equivalent of Strassen’s theorem in this case is the following theorem, which we prove in the finite case only again.

**Theorem 4.37** (Strassen’s theorem for Markov kernels). Let \((X_t)\) and \((Y_t)\) be Markov chains on a finite poset \((\mathcal{X}, \leq)\) with transition matrices \(P\) and \(Q\) respectively. Assume that \(Q\) stochastically dominates \(P\). Then for all \(x_0 \leq y_0\) there is a coupling \((\hat{X}_t, \hat{Y}_t)\) of \((X_t)\) started at \(x_0\) and \((Y_t)\) started at \(y_0\) such that a.s.
\[
\hat{X}_t \leq \hat{Y}_t, \quad \forall t.
\]
Furthermore, if the chains are irreducible and have stationary distributions \( \pi \) and \( \mu \) respectively, then \( \pi \preceq \mu \).

Observe that, for a step-by-step monotone coupling to exist, it is not generally enough for the weaker condition \( P(x, \cdot) \preceq Q(x, \cdot) \) to hold for all \( x \), as should be clear from the proof. Also you should convince yourself that the chains in Example 4.34 do not in general satisfy (4.5). (Which pairs \( x, y \) cause problems?)

**Proof of Theorem 4.37.** Let

\[
W := \{(x, y) \in \mathcal{X} \times \mathcal{X} : x \leq y\}.
\]

For all \((x, y) \in W\), let \( R((x, y), \cdot) \) be the joint law of a monotone coupling of \( P(x, \cdot) \) and \( Q(y, \cdot) \). Such a coupling exists by Strassen’s theorem and Condition (4.5). Let \( (\hat{X}_t, \hat{Y}_t) \) be a Markov chain on \( W \) with transition matrix \( R \) started at \((x_0, y_0)\). By construction, \( \hat{X}_t \leq \hat{Y}_t \) for all \( t \) a.s. That proves the first half of the theorem.

For the second half, let \( A \) be increasing in \( \mathcal{X} \). Then, by the ergodic theorem for Markov chains (e.g. [Dur10, Exercise 6.6.4]),

\[
\pi(A) = \frac{1}{t} \sum_{s \leq t} 1_{\hat{X}_s \in A} \leq \frac{1}{t} \sum_{s \leq t} 1_{\hat{Y}_s \in A} \to \mu(A), \quad \text{a.s.}
\]

where we used that \( \hat{X}_s \in A \) implies \( \hat{Y}_s \in A \) because \( \hat{X}_s \leq \hat{Y}_s \) and \( A \) is increasing. This proves the claim by definition of stochastic domination.

An example of application of this theorem is given in the next subsection.

### 4.3.2 Ising model on \( \mathbb{Z}^d \): extremal measures

Consider the \( d \)-dimensional lattice \( \mathbb{Z}^d \). Let \( \Lambda \) be a finite subset of vertices in \( \mathbb{Z}^d \) and define \( \mathcal{X} := \{-1, +1\}^\Lambda \), which is a poset when equipped with the relation \( \sigma \leq \sigma' \) if and only if \( \sigma_i \leq \sigma'_i \) for all \( i \in \Lambda \). For shorthand, we occasionally write + and − instead of \(+1\) and \(-1\). For \( \xi \in \{-1, +1\}^{\mathbb{Z}^d} \), recall that the (ferromagnetic) Ising model on \( \Lambda \) with boundary conditions \( \xi \) and inverse temperature \( \beta \) is the probability distribution over spin configurations \( \sigma \in \mathcal{X} \) given by

\[
\mu_{\beta, \Lambda}^\xi(\sigma) := \frac{1}{Z_{\Lambda, \xi}(\beta)} e^{-\beta \mathcal{H}_{\Lambda, \xi}(\sigma)},
\]

where

\[
\mathcal{H}_{\Lambda, \xi}(\sigma) := -\sum_{i \sim j} \sigma_i \xi_j - \sum_{i \in \Lambda, j \notin \Lambda} \sigma_i \xi_j.
\]
is the Hamiltonian and

$$Z_{\Lambda, \xi}(\beta) := \sum_{\sigma \in \mathcal{X}} e^{-\beta \mathcal{H}_{\Lambda, \xi}(\sigma)},$$

is the partition function. (Warning: it is easy to get confused with the $-$ signs that cancel out in the exponent.) For the all-$(+1)$ and all-$(-1)$ boundary conditions we write respectively $\mu_{\beta, \Lambda}^+(\sigma)$ and $\mu_{\beta, \Lambda}^-(\sigma)$.

In this section, we show that these two measures are extremal in the following sense. For all boundary conditions $\xi \in \{-1, +1\}^{1^d}$:

**Claim 4.38.**

$$\mu_{\beta, \Lambda}^+ \succeq \mu_{\beta, \Lambda}^\xi \succeq \mu_{\beta, \Lambda}^-.$$  

In words, because the ferromagnetic Ising model favors spin agreement, the all-$(+1)$ boundary condition tends to produce more $+1$s which in turn makes increasing events more likely.

The idea of the proof is to use Theorem 4.37 with a suitable Markov chain.

**Stochastic domination** In this context, vertices are often referred to as sites. Recall that the single-site Glauber dynamics of the Ising model is the Markov chain on $\mathcal{X}$ which, at each time, selects a site $i \in \Lambda$ uniformly at random and updates the spin $\sigma_i$ according to $\mu_{\beta, \Lambda}^\xi(\sigma)$ conditioned on agreeing with $\sigma$ at all sites in $\Lambda \setminus \{i\}$.

Specifically, for $\gamma \in \{-1, +1\}$, $i \in \Lambda$, and $\sigma \in \mathcal{X}$, let $\sigma^{i, \gamma}$ be the configuration $\sigma$ with the state at $i$ being set to $\gamma$. Then, letting $n = |\Lambda|$, because the Ising measure factorizes, the transition matrix of the Glauber dynamics is simply

$$Q_{\beta, \Lambda}^\xi(\sigma, \sigma^{i, \gamma}) := \frac{1}{n} \cdot \frac{e^{\gamma \beta S_i^\xi(\sigma)}}{e^{-\beta S_i^\xi(\sigma)} + e^{\beta S_i^\xi(\sigma)}},$$

where

$$S_i^\xi(\sigma) := \sum_{j \sim i} \sigma_j + \sum_{j \not\sim i} \xi_j.$$  

All other transitions have probability 0.

This chain is clearly irreducible. It is also reversible with respect to $\mu_{\beta, \Lambda}^\xi$. Indeed, for all $\sigma \in \mathcal{X}$ and $i \in \Lambda$, letting

$$S_{\neq i}^\xi(\sigma) := \mathcal{H}_{\Lambda, \xi}(\sigma^{i, +}) + S_i^\xi(\sigma) = \mathcal{H}_{\Lambda, \xi}(\sigma^{i, -}) - S_i^\xi(\sigma),$$

210
we have

$$
\mu_{\beta,\Lambda}(\sigma^{i,-}) Q_{\beta,\Lambda}(\sigma^{i,-}, \sigma^{i,+}) = \frac{e^{-\beta S_{\varphi^{i}}(\sigma)}e^{-\beta S_{\xi}(\sigma)}}{Z_{\Lambda,\xi}(\beta)} \cdot \frac{e^{\beta S_{\xi}(\sigma)}}{n[e^{-\beta S_{\xi}(\sigma)} + e^{\beta S_{\xi}(\sigma)}]}
$$

$$
= \frac{e^{-\beta S_{\varphi^{i}}(\sigma)}}{nZ_{\Lambda,\xi}(\beta)[e^{-\beta S_{\xi}(\sigma)} + e^{\beta S_{\xi}(\sigma)}]}
$$

$$
= \frac{e^{-\beta S_{\varphi^{i}}(\sigma)}e^{\beta S_{\xi}(\sigma)}}{Z_{\Lambda,\xi}(\beta)} \cdot \frac{e^{-\beta S_{\xi}(\sigma)}}{n[e^{-\beta S_{\xi}(\sigma)} + e^{\beta S_{\xi}(\sigma)}]}
$$

$$
= \mu_{\beta,\Lambda}(\sigma^{i,+}) Q_{\beta,\Lambda}(\sigma^{i,+}, \sigma^{i,-}).
$$

In particular $\mu_{\beta,\Lambda}$ is the stationary distribution of $Q_{\beta,\Lambda}$.

Claim 4.39.

$$
\xi' \geq \xi \implies Q_{\beta,\Lambda}^{\xi'} stochastically dominates Q_{\beta,\Lambda}^{\xi}.
$$

**Proof.** Because the Glauber dynamics updates a single site at a time, establishing stochastic domination reduces to checking simple one-site inequalities:

**Lemma 4.40.** To establish (4.6), it suffices to show that, for all $\sigma \leq \tau$,

$$
Q_{\beta,\Lambda}^{\xi}(\sigma, \sigma^{i,+}) \leq Q_{\beta,\Lambda}^{\xi'}(\tau, \tau^{i,+}).
$$

**Proof.** Assume (4.7) holds. Let $A$ be increasing in $\mathcal{X}$ and let $\sigma \leq \tau$. Then, for the single-site Glauber dynamics, we have

$$
Q_{\beta,\Lambda}^{\xi}(\sigma, A) = Q_{\beta,\Lambda}^{\xi}(\sigma, A \cap B_{\sigma}),
$$

where

$$
B_{\sigma} := \{\sigma^{i,\gamma} : i \in \Lambda, \gamma \in \{-1, +1\}\},
$$

and similarly for $\tau, \xi'$. Moreover, because $A$ is increasing and $\tau \geq \sigma$,

$$
\sigma^{i,\gamma} \in A \implies \tau^{i,\gamma} \in A,
$$

and

$$
\sigma^{i,-} \in A \implies \sigma^{i,+} \in A.
$$

Letting

$$
I_{\sigma,A}^{\pm} := \{i \in \Lambda : \sigma^{i,\pm} \in A\},
$$

$$
I_{\sigma,A}^{\mp} := \{i \in \Lambda : \sigma^{i,\mp} \in A\},
$$

211
and similarly for $\tau$, we have by (4.7), (4.8), (4.9), and (4.10),

\[
Q^\xi_{\beta,\Lambda}(\sigma, A) = Q^\xi_{\beta,\Lambda}(\sigma, A \cap B_\sigma)
\]

\[
= \sum_{i \in I_{\sigma,A}^+} Q^\xi_{\beta,\Lambda}(\sigma, \sigma_{i,+}) + \sum_{i \in I_{\sigma,A}^-} \left[ Q^\xi_{\beta,\Lambda}(\sigma, \sigma_{i,-}) + Q^\xi_{\beta,\Lambda}(\sigma, \sigma_{i,+}) \right]
\]

\[
\leq \sum_{i \in I_{\sigma,A}^+} Q^{\xi'}_{\beta,\Lambda}(\tau, \tau_{i,+}) + \sum_{i \in I_{\sigma,A}^-} \frac{1}{n}
\]

\[
\leq \sum_{i \in I_{\tau,A}^+} Q^{\xi'}_{\beta,\Lambda}(\tau, \tau_{i,+}) + \sum_{i \in I_{\tau,A}^-} \left[ Q^{\xi'}_{\beta,\Lambda}(\tau, \tau_{i,-}) + Q^{\xi'}_{\beta,\Lambda}(\tau, \tau_{i,+}) \right]
\]

\[
= Q^{\xi'}_{\beta,\Lambda}(\tau, A),
\]

as claimed.

Returning to the proof of Claim 4.39, observe that

\[
Q^\xi_{\beta,\Lambda}(\sigma, \sigma_{i,+}) = \frac{1}{n} \cdot \frac{e^{2\beta S^\xi_i(\sigma)}}{e^{2\beta S^\xi_i(\sigma)} + 1}
\]

which is increasing in $S^\xi_i(\sigma)$. Now $\sigma \leq \tau$ and $\xi \leq \xi'$ imply that $S^\xi_i(\sigma) \leq S^\xi_i(\tau)$. That proves the claim.

Finally:


Observe that we have not used any special property of the $d$-dimensional lattice. Indeed Claim 4.38 in fact holds for any countable, locally finite graph with positive coupling constants.

**Thermodynamic limit** To be written. See [RAS, Theorem 9.13].

**4.3.3 Random walk on trees: speed**

To be written. See [LP, Proposition 13.3 and Exercise 13.1].\(^\dagger\)

\(^\dagger\)Requires: Section 2.3.3.
4.3.4 Correlation inequalities: FKG and Holley’s inequalities

A special case of stochastic domination is positive associations. In this section, we restrict ourselves to posets of the form \( \{0, 1\}^F \) for \( F \) finite. We begin with an example.

**Example 4.41** (Erdős-Rényi graphs: positive associations). Consider an Erdős-Rényi graph \( G \sim \mathbb{G}_{n,p} \). Let \( E = \{\{x, y\} : x, y \in [n], x \neq y\} \). Think of \( G \) as taking values in the poset \( (\{0, 1\}^E, \leq) \) where a 1 indicates that the corresponding edge is present. In fact observe that the law of \( G \), which we denote as usual by \( \mathbb{P}_{n,p} \), is a product measure on \( \{0, 1\}^E \). The event \( A \) that \( G \) is connected is increasing because adding edges cannot disconnect a connected graph. So is the event \( B \) of having a chromatic number larger than 4. Intuitively then, conditioning on \( A \) makes \( B \) more likely. Indeed the occurrence of \( A \) tends to be accompanied with a larger number of edges which in turn makes \( B \) more probable. This is a more general phenomenon. That is, for any non-empty increasing events \( A \) and \( B \), we have:

\[
\mathbb{P}_{n,p}[B | A] \geq \mathbb{P}_{n,p}[B].
\] (4.11)

Or, put differently, the conditional measure \( \mathbb{P}_{n,p}[\cdot | A] \) stochastically dominates the unconditional measure \( \mathbb{P}_{n,p}[\cdot] \). This is a special case of what is known as Harris’ inequality (proved below). Note that (4.11) is equivalent to \( \mathbb{P}_{n,p}[A \cap B] \geq \mathbb{P}_{n,p}[A] \mathbb{P}_{n,p}[B] \), i.e., to the fact that \( A \) and \( B \) are positively correlated.

More generally:

**Definition 4.43** (Positive associations). Let \( \mu \) be a probability measure on \( \{0, 1\}^F \) where \( F \) is finite. Then \( \mu \) is said to have positive associations, or is positively associated, if for all increasing functions \( f, g : \{0, 1\}^F \rightarrow \mathbb{R} \)

\[
\mu(fg) \geq \mu(f)\mu(g),
\]

where

\[
\mu(h) := \sum_{\omega \in \{0,1\}^F} \mu(\omega)h(\omega).
\]

In particular, for any increasing events \( A \) and \( B \) it holds that

\[
\mu(A \cap B) \geq \mu(A)\mu(B),
\]

i.e., \( A \) and \( B \) are positively correlated.
Remark 4.44. Note that positive associations is concerned only with increasing events. See Remark 4.63.

Remark 4.45. A notion of negative associations, which is a somewhat more delicate concept, was defined in Remark 3.121. See also [Pem00].

Let \( \mu \) be positively associated. Note that if \( A \) and \( B \) are decreasing, i.e. their complements are increasing, then
\[
\mu(A \cap B) = 1 - \mu(A^c \cup B^c) \\
= 1 - \mu(A^c) - \mu(B^c) + \mu(A^c \cap B^c) \\
\geq 1 - \mu(A^c) - \mu(B^c) + \mu(A^c)\mu(B^c) \\
= \mu(A)\mu(B).
\]

Similarly, if \( A \) is increasing and \( B \) is decreasing, we have
\[
\mu(A \cap B) \leq \mu(A)\mu(B).
\]

Harris' inequality states that product measures on \( \{0, 1\}^F \) have positive associations. We prove a more general result known as the FKG inequality. For two configurations \( \omega, \omega' \in \{0, 1\}^F \), we let \( \omega \wedge \omega' \) and \( \omega \vee \omega' \) be the coordinatewise minimum and maximum of \( \omega \) and \( \omega' \).

Theorem 4.46 (FKG inequality). Let \( \mathcal{X} = \{0, 1\}^F \) where \( F \) is finite. Suppose \( \mu \) is a positive probability measure on \( \mathcal{X} \) satisfying the FKG condition
\[
\mu(\omega \vee \omega') \mu(\omega \wedge \omega') \geq \mu(\omega) \mu(\omega'), \quad \forall \omega, \omega' \in \mathcal{X}. \tag{4.12}
\]

This property is also known as log-convexity or log-supermodularity. We call such a measure an FKG measure. Then \( \mu \) has positive associations.

Remark 4.47. Strict positivity is not in fact needed [FKG71]. The FKG condition is equivalent to a strong form of positive associations. See Exercise 4.4.

Note that product measures satisfy the FKG condition with equality. Indeed if \( \mu(\omega) \) is of the form \( \prod_{f \in F} \mu_f(\omega_f) \) then
\[
\mu(\omega \vee \omega') \mu(\omega \wedge \omega') = \prod_{f} \mu_f(\omega_f \vee \omega'_f) \mu_f(\omega_f \wedge \omega'_f) \\
= \prod_{f: \omega_f = \omega'_f} \mu_f(\omega_f)^2 \prod_{f: \omega_f \neq \omega'_f} \mu_f(\omega_f)\mu_f(\omega'_f) \\
= \prod_{f: \omega_f = \omega'_f} \mu_f(\omega_f)\mu_f(\omega'_f) \prod_{f: \omega_f \neq \omega'_f} \mu_f(\omega_f)\mu_f(\omega'_f) \\
= \mu(\omega) \mu(\omega').
\]
So the FKG inequality applies, for instance, to bond percolation and Erdős-Rényi graphs. The pointwise nature of the FKG condition also makes it relatively easy to check it for measures which are defined explicitly up to a normalizing constant, such as the Ising model.

**Example 4.48** (Ising model on $\mathbb{Z}^d$: checking FKG). Consider again the setting of Section 4.3.2. Of course we work on the space $\mathcal{X} := \{-1, +1\}^\Lambda$ rather than $\{0, 1\}^F$. Fix a finite $\Lambda \subseteq \mathbb{Z}^d$, $\xi \in \{-1, +1\}^{\mathbb{Z}^d}$ and $\beta > 0$.

**Claim 4.49.** The measure $\mu_{\beta, \Lambda}$ satisfies the FKG condition and therefore has positive associations.

Intuitively, taking the maximum or minimum of two configurations tends to increase spin agreement and therefore leads to a higher likelihood. By taking logarithms in the FKG condition, one sees that proving the claim boils down to checking an inequality for each term in the Hamiltonian. For $\sigma, \sigma' \in \mathcal{X}$, let $\tau = \sigma \lor \sigma'$ and $\tau = \sigma \land \sigma'$. When $i \in \Lambda$ and $j \notin \Lambda$ such that $i \sim j$, we have

$$
\tau_i \xi_j + \tau_i \xi_j = (\tau_i + \tau_i)\xi_j = (\sigma_i + \sigma'_i)\xi_j = \sigma_i \xi_j + \sigma'_i \xi_j.
$$

(4.13)

For $i, j \in \Lambda$ with $i \sim j$, note first that the case $\sigma_j = \sigma'_j$ reduces to the previous calculation, so we assume $\sigma_i \neq \sigma'_i$ and $\sigma_j \neq \sigma'_j$.

Then

$$
\tau_i \tau_j + \tau_i \tau_j = (+1)(+1) + (-1)(-1) = 2 \geq \sigma_i \sigma_j + \sigma'_i \sigma'_j,
$$

since 2 is the largest value the rightmost expression ever takes. We have shown that

$$
\mathcal{H}_{\Lambda, \xi}(\tau) + \mathcal{H}_{\Lambda, \xi}(\tau) \leq \mathcal{H}_{\Lambda, \xi}(\sigma) + \mathcal{H}_{\Lambda, \xi}(\sigma'),
$$

which implies the claim.

Again, we have not used any special property of the lattice and the same result holds for countable, locally finite graphs with positive coupling constants. Note however that in the anti-ferromagnetic case, i.e., if we multiply the Hamiltonian by $-1$, the above argument does not work. Indeed there is no reason to expect positive associations in that case.

The FKG inequality in turn follows from a more general result known as Holley’s inequality.

**Theorem 4.50** (Holley’s inequality). Let $\mathcal{X} = \{0, 1\}^F$ where $F$ is finite. Suppose $\mu_1$ and $\mu_2$ are positive probability measures on $\mathcal{X}$ satisfying

$$
\mu_2(\omega \lor \omega') \mu_1(\omega \land \omega') \geq \mu_2(\omega) \mu_1(\omega'), \quad \forall \omega, \omega' \in \mathcal{X}.
$$

(4.14)

Then $\mu_1 \preceq \mu_2$. 

\[215\]
Before proving Holley’s inequality, we check that it indeed implies the FKG inequality. See Exercise 4.1 for an elementary proof in the independent case, i.e., of Harris’ inequality.

**Proof of Theorem 4.46.** Assume that $\mu$ satisfies the FKG condition and let $f, g$ be increasing functions. Because of our restriction to positive measures in Holley’s inequality, we will work with positive functions. This is done without loss of generality. Indeed, letting $\mathbf{0}$ be the all-$0$ vector, note that $f$ and $g$ are increasing if and only if $f' := f - f(\mathbf{0}) + 1 > 0$ and $g' := g - g(\mathbf{0}) + 1 > 0$ are increasing and that, moreover,

$$
\mu(f'g') - \mu(f')\mu(g') = \mu([f' - \mu(f)] [g' - \mu(g'])
$$

$$
= \mu([f - \mu(f)] [g - \mu(g])
$$

$$
= \mu(fg) - \mu(f)\mu(g).
$$

In Holley’s inequality, we let $\mu_1 := \mu$ and define the positive probability measure

$$
\mu_2(\omega) := \frac{g(\omega)\mu(\omega)}{\mu(g)}.
$$

We check that $\mu_1$ and $\mu_2$ satisfy the conditions of Holley’s inequality. Note that $\omega' \leq \omega \lor \omega'$ for any $\omega$ so that, because $g$ is increasing, we have $g(\omega') \leq g(\omega \lor \omega')$. Hence, for any $\omega, \omega'$,

$$
\mu_1(\omega)\mu_2(\omega') = \mu(\omega)\frac{g(\omega')\mu(\omega')}{\mu(g)}
$$

$$
= \mu(\omega)\frac{g(\omega')}{\mu(g)}
$$

$$
\leq \mu(\omega \land \omega')\mu(\omega \lor \omega')\frac{g(\omega \lor \omega')}{\mu(g)}
$$

$$
= \mu_1(\omega \land \omega')\mu_2(\omega \lor \omega'),
$$

where on the third line we used the FKG condition satisfied by $\mu$.

So Holley’s inequality implies that $\mu_2 \succeq \mu_1$. Hence, since $f$ is increasing, by Corollary 4.33

$$
\mu(f) = \mu_1(f) \leq \mu_2(f) = \frac{\mu(fg)}{\mu(g)},
$$

and the theorem is proved.

**Proof of Theorem 4.50.** We use Theorem 4.37. This is similar to what was done in Section 4.3.2. This time we couple Metropolis-like chains. For $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$,
{0, 1}, we let $x^i\gamma$ be $x$ with coordinate $i$ set to $\gamma$. We write $x \sim y$ if $\|x - y\|_1 = 1$. Let $n = |F|$.

For $\alpha, \beta > 0$ small enough, the following transition matrix over $\mathcal{X}$ is irreducible and reversible w.r.t. its stationary distribution $\mu_2$: for all $i \in F$, $y \in \mathcal{X}$,

$$Q(y_{i,0}, y_{i,1}) = \frac{\alpha}{n} \beta,$$

$$Q(y_{i,1}, y_{i,0}) = \frac{\alpha}{n} \left\{ \beta \frac{\mu_2(y_{i,0})}{\mu_2(y_{i,1})} \right\},$$

$$Q(y, y) = 1 - \sum_{z \sim y} Q(y, z).$$

Let $P$ be similarly defined for $\mu_1$ with the same values of $\alpha$ and $\beta$. For reasons that will be clear below, the value of $\beta$ is chosen so that the sum of the two expressions in brackets above is smaller than 1 for all $y, i$. The value of $\alpha$ is then chosen so that $P(x, x), Q(y, y) \geq 0$ for all $x, y$. Reversibility follows immediately from the first two equations. We call the first transition above an upward transition and the second one a downward transition.

By Theorem 4.37, it remains to show that $Q$ stochastically dominates $P$. That is, for any $x \leq y$, we want to show that $P(x, \cdot) \preceq Q(y, \cdot)$. We produce a monotone coupling $(\hat{X}, \hat{Y})$ of these two distributions. Because $x \leq y$, our goal is never to perform an upward transition in $x$ simultaneously with a downward transition in $y$.

Observe that

$$\frac{\mu_1(x_{i,0})}{\mu_1(x_{i,1})} \geq \frac{\mu_2(y_{i,0})}{\mu_2(y_{i,1})} \quad (4.15)$$

by taking $\omega = y_{i,0}$ and $\omega' = x_{i,1}$ in Condition (4.14).

The coupling works as follows. Fix $x \leq y$. With probability $1 - \alpha$, set $(\hat{X}, \hat{Y}) := (x, y)$. Otherwise, pick a coordinate $i \in F$ uniformly at random. There are several cases to consider depending on the values of $x_i, y_i$ (with $x_i \leq y_i$ by assumption):

- $(x_i, y_i) = (0, 0)$: With probability $\beta$, perform an upward transition in both, i.e., set $\hat{X} := x_{i,1}$ and $\hat{Y} := y_{i,1}$. With probability $1 - \beta$, set $(\hat{X}, \hat{Y}) := (x, y)$ instead.

- $(x_i, y_i) = (1, 1)$: With probability $\beta \frac{\mu_2(y_{i,0})}{\mu_2(y_{i,1})}$, perform a downward transition in both, i.e., set $\hat{X} := x_{i,0}$ and $\hat{Y} := y_{i,0}$. With probability

$$\beta \left( \frac{\mu_1(x_{i,0})}{\mu_1(x_{i,1})} - \frac{\mu_2(y_{i,0})}{\mu_2(y_{i,1})} \right),$$
perform a downward transition in \( x \) only, i.e., set \( \hat{X} := x^{i,0} \) and \( \hat{Y} := y \). Note that (4.15) guarantees that the previous step is well-defined. With the remaining probability, set \( (\hat{X}, \hat{Y}) := (x, y) \) instead.

\[ (x_i, y_i) = (0, 1): \text{ With probability } \beta, \text{ perform an upward transition in } x \text{ only, i.e., set } \hat{X} := x^{i,1} \text{ and } \hat{Y} := y. \text{ With probability } \beta \mu^2(y_i, 0) \mu^2(y_i, 1), \text{ perform a downward transition in } y \text{ only, i.e., set } \hat{X} := x \text{ and } \hat{Y} := y^{i,0}. \text{ With the remaining probability, set } (\hat{X}, \hat{Y}) := (x, y) \text{ instead. (This is where we use the odd choice of } \beta.) \]

By construction, this coupling satisfies \( \hat{X} \leq \hat{Y} \) a.s. An application of Theorem 4.37 concludes the proof.

**Example 4.51** (Ising model: extremality revisited). Holley’s inequality gives another proof of Claim 4.38. To see this, just repeat the calculations of Example 4.48, where now (4.13) is replaced with an inequality. See Exercise 4.2.

4.3.5 ▶ **Erdős-Rényi graph: Janson’s inequality and application to the containment problem**

Let \( G = (V, E) \sim \mathcal{G}_{n,p} \) be an Erdős-Rényi graph. Repeating the computations of Section 2.3.2 (or see Claim 2.23), we see that the property of being triangle-free has threshold \( n^{-1} \). That is, the probability that \( G \) contains a triangle goes to 0 or 1 as \( n \to +\infty \) depending on whether \( p \ll n^{-1} \) or \( p \gg n^{-1} \) respectively. In this section, we investigate what happens at the threshold. From now on, we assume that \( p = \lambda/n \) for some \( \lambda > 0 \) not depending on \( n \).

For any subset \( S \) of three distinct vertices of \( G \), let \( B_S \) be the event that \( S \) forms a triangle in \( G \). So

\[
\varepsilon := P_{n,p}[B_S] = p^3 \to 0.
\]

(4.16)

Let \( X_n = \sum_{S \in \binom{V}{3}} 1_{B_S} \) be the number of triangles in \( G \). By the linearity of expectation, the expected number of triangles is

\[
\mathbb{E}_{n,p}X_n = \binom{n}{3} p^3 = \frac{n(n-1)(n-2)}{6} \left( \frac{\lambda}{n} \right)^3 \to \frac{\lambda^3}{6},
\]

as \( n \to +\infty \). If the events \( \{B_S\}_S \) were mutually independent, \( X_n \) would be binomially-distributed and the event that \( G \) is triangle-free would have probability

\[
\prod_{S \in \binom{V}{3}} P_{n,p}[B^c_S] = (1 - p^3)^{\binom{n}{3}} \to e^{-\lambda^3/6}.
\]

(4.17)

218
In fact, by the Poisson approximation to the binomial (e.g. [Dur10, Theorem 3.6.1]), we would have that the number of triangles converges weakly to $\text{Poi}(\lambda^3/6)$.

In reality, of course, the events $\{B_S\}$ are not mutually independent. Observe however that, for most pairs $S, S'$, the events $B_S$ and $B_{S'}$ are in fact pairwise independent. That is the case whenever $|S \cap S'| \leq 1$, i.e., whenever the edges connecting $S$ are disjoint from those connecting $S'$. Write $S \sim S'$ if $S \neq S'$ are not independent, i.e. if $|S \cap S'| = 2$. The expected number of (unordered) pairs $S \sim S'$ both forming a triangle is

$$\Delta := \frac{1}{2} \sum_{S,S' \in \binom{V_3}{2}} \mathbb{P}_{n,p}[B_S \cap B_{S'}] = \frac{1}{2} \binom{n}{3} (n-3)p^5 = \Theta(n^4 p^5) \to 0. \quad (4.18)$$

Given that the events $\{B_S\}$ are “mostly” independent, it is natural to expect that $X_n$ behaves asymptotically as it does in the independent case. Indeed we prove:

**Claim 4.52.**

$$\mathbb{P}_{n,p}[X_n = 0] \to e^{-\lambda^3/6}.$$

**Remark 4.53.** In fact, $X_n \xrightarrow{d} \text{Poi}(\lambda^3/6)$. See Exercises 2.13 and 4.5.

The FKG inequality immediately gives one direction. Recall that $\mathbb{P}_{n,p}$, as a product measure over edge sets, satisfies the FKG condition and therefore has positive associations by the FKG inequality. Moreover the events $B_S^c$ are decreasing for all $S$. Hence

$$\mathbb{P}_{n,p} \left[ \bigcap_{S \in \binom{V_3}{2}} B_S^c \right] = \prod_{S \in \binom{V_3}{2}} \mathbb{P}_{n,p}[B_S^c] \to e^{-\lambda^3/6},$$

by (4.17). As it turns out, the FKG inequality also gives a bound in the other direction. This is known as *Janson’s inequality*, which we state in a more general context.

**Janson’s inequality** Let $\mathcal{X} := \{0,1\}^F$ where $F$ is finite. Let $B_i, i \in I$, be a finite collection of events of the form $B_i := \{\omega \in \mathcal{X} : \omega \geq \beta^{(i)}\}$ for some $\beta^{(i)} \in \mathcal{X}$. Think of these as “bad events” corresponding to a certain subset of coordinates being set to 1. By definition, the $B_i$s are increasing. Assume $\mathbb{P}$ is a positive product measure on $\mathcal{X}$. Write $i \sim j$ if $\beta_r^{(i)} = \beta_r^{(j)} = 1$ for some $r$ and note that $B_i$ is independent of $B_j$ if $i \sim j$. Set

$$\Delta := \sum_{\{i,j\} \sim i \sim j} \mathbb{P}[B_i \cap B_j].$$
Theorem 4.54 (Janson’s inequality). Let $X, P, \{B_i\}_{i \in I}$ and $\Delta$ be as above. Assume further that there is $\varepsilon > 0$ such that $P[B_i] \leq \varepsilon$ for all $i \in I$. Then

$$\prod_{i \in I} P[B_i] \leq P[\cap_{i \in I} B_i^c] \leq e^{1/\varepsilon^2} \prod_{i \in I} P[B_i^c].$$

Before proving the theorem, we show that it implies Claim 4.52. We have already shown in (4.16) and (4.18) that $\varepsilon \to 0$ and $\Delta \to 0$. Janson’s inequality immediately implies the claim by (4.17).

**Proof of Theorem 4.54.** The lower bound follows from the FKG inequality.

In the other direction, the first step is somewhat clear. We apply the chain rule to obtain

$$P[\cap_{i \in I} B_i^c] = \prod_{i = 1}^m P[B_i^c \mid \cap_{j \in [i-1]} B_j^c].$$

The rest is clever manipulation. W.l.o.g. assume $I = [m]$. For $i \in [m]$, let $N(i) := \{\ell \in [m] : \ell \sim i\}$ and $N_<(i) := N(i) \cap [i-1]$. Note that $B_i$ is independent of $\{B_\ell : \ell \in [i-1] \setminus N_<(i)\}$. Then

$$P[B_i \mid \cap_{j \in [i-1]} B_j^c] = \frac{P[B_i \cap (\cap_{j \in [i-1]} B_j^c)]}{P[\cap_{j \in [i-1]} B_j^c]} \geq \frac{P[B_i \cap (\cap_{j \in N_<(i)} B_j^c)]}{P[\cap_{j \in [i-1] \setminus N_<(i)} B_j^c]} = \frac{P[B_i \cap (\cap_{j \in [i-1] \setminus N_<(i)} B_j^c)]}{P[B_i] \times P[B_i \mid \cap_{j \in [i-1] \setminus N_<(i)} B_j^c]} \times P[B_i \mid \cap_{j \in [i-1] \setminus N_<(i)} B_j^c],$$

By a union bound the second term on the last line is

$$\begin{align*}
P[\cap_{j \in N_<(i)} B_j^c \mid B_i \cap (\cap_{j \in [i-1] \setminus N_<(i)} B_j^c)] &\geq 1 - \sum_{j \in N_<(i)} P[B_j \mid B_i \cap (\cap_{j \in [i-1] \setminus N_<(i)} B_j^c)] \\
&\geq 1 - \sum_{j \in N_<(i)} P[B_j \mid B_i].
\end{align*}$$

220
where the last line follows from the FKG inequality applied to the product measure $\mathbb{P}[-|B_i]$ (on $\{0, 1\}^F$ with $F := \{\ell \in [m] : \beta^{(i)}_\ell = 0\}$). Combining the last three displays and using $1 + z \leq e^z$, we get

$$\mathbb{P}\left[\cap_{i \in I} B_i^c\right] \leq \prod_{i=1}^m \left[ \mathbb{P}[B_i^c] + \sum_{j \in N_{<}(i)} \mathbb{P}[B_i \cap B_j] \right] \leq \prod_{i=1}^m \mathbb{P}[B_i^c] \exp \left( \frac{1}{1 - \varepsilon} \sum_{j \in N_{<}(i)} \mathbb{P}[B_i \cap B_j] \right).$$

By the definition of $\Delta$, we are done. 

\section{Percolation on $\mathbb{Z}^2$: RSW theory and a proof of Harris’ theorem}

Consider bond percolation on the two-dimensional lattice $\mathbb{L}^2$. Recall that the percolation function is given by

$$\theta(p) := \mathbb{P}_p[|C_0| = +\infty],$$

where $C_0$ is the open cluster of the origin. We known from Example 4.30 that $\theta(p)$ is non-decreasing. Let

$$p_c(\mathbb{L}^2) := \sup\{p \geq 0 : \theta(p) = 0\},$$

be the critical value. We proved in Section 2.2.4 that there is a non-trivial transition, i.e., $p_c(\mathbb{L}^2) \in (0, 1)$. In fact we showed that $p_c(\mathbb{L}^2) \in [1/3, 2/3]$ (see Exercise 2.2).

Our goal in this section is to use the FKG inequality to improve this further to:

\textbf{Theorem 4.55} (Harris’ theorem).

$$\theta(1/2) = 0.$$

Or, put differently, $p_c(\mathbb{L}^2) \geq 1/2$.

\textbf{Remark 4.56.} This bound is tight, i.e., in fact $p_c(\mathbb{L}^2) = 1/2$. The other direction, known as Kesten’s theorem, is postponed to Section ?? where an additional ingredient is introduced, Russo’s formula.

Several proofs of Harris’ theorem are known. A particularly elegant one is sketched in Exercise 5.1. Here we present a proof that uses an important tool in percolation theory, the RSW lemma, an application of the FKG inequality.
Harris’ theorem  To motivate the RSW lemma, we start with the proof of Harris’ theorem.

Proof of Theorem 4.55. Fix \( p = 1/2 \). We use duality. Consider the \( \mathbb{L}^2 \) annulus

\[
\text{Ann}(\ell) := [-3\ell, 3\ell]^2 \setminus [-\ell, \ell].
\]

The existence of a closed dual cycle inside \( \text{Ann}(\ell) \), which we denote by \( O_d(\ell) \), prevents the possibility of an infinite open self-avoiding path from the origin in the primal lattice \( \mathbb{L}^2 \). See Figure 4.4. That is,

\[
P_{1/2}[|C_0| = +\infty] \leq \prod_{k=0}^{K} \{1 - P_{1/2}[O_d(3^k)]\},
\]

for all \( K \), where we took powers of 3 to make the annuli disjoint and therefore independent.

To prove the theorem, it suffices to show that there is a constant \( c^* > 0 \) such that, for all \( \ell \), \( P_{1/2}[O_d(\ell)] \geq c^* \). Then the r.h.s. of (4.19) tends to 0 as \( K \to +\infty \).

To simplify further, thinking of \( \text{Ann}(\ell) \) as a union of four rectangles \([-3\ell, -\ell] \times [-3\ell, 3\ell], [-3\ell, 3\ell] \times (\ell, 3\ell], -\ell \times \{\ell, 3\ell\}, \text{etc.} \), it suffices to consider the event \( O_d^#(\ell) \) that each one of these rectangles contains a closed dual self-avoiding path connecting its two shorter sides. (More precisely, for the first rectangle above, the path connects \([-3\ell + 1/2, -\ell - 1/2] \times \{3\ell - 1/2\} \) to \([-3\ell + 1/2, -\ell - 1/2] \times \{-3\ell + 1/2\} \) and stays inside the rectangle, etc.) See Figure 4.4. By symmetry the probability that such a path exists is the same for all four rectangles. Denote it by \( \rho_{\ell} \). Moreover the event that such a path exists is increasing so, although the four events are not independent, we can apply the FKG inequality. Hence, because \( O_d^#(\ell) \subseteq O_d(\ell) \), we finally get the bound

\[
P_{1/2}[O_d(\ell)] \geq \rho_{\ell}^4.
\]

The RSW lemma and some symmetry arguments, both of which are detailed below, imply that there is some \( c > 0 \) such that, for all \( \ell \):

Lemma 4.57.

\[
\rho_{\ell} \geq c.
\]

That concludes the proof.

It remains to prove Lemma 4.57. We first state the RSW lemma.
Figure 4.4: Top: the event $O_d(\ell)$. Bottom: the event $O_d^\#(\ell)$. 
RSW theory  We have reduced the proof of Harris’ theorem to bounding the probability that certain closed paths exist in the dual lattice. To be consistent with the standard RSW notation, we switch to the primal lattice and consider open paths. We also let $p$ take any value in $(0, 1)$.

Let $R_{n, \alpha}(p)$ be the probability that the rectangle

$$B(\alpha n, n) := [-n, (2\alpha - 1)n] \times [-n, n],$$

has an open self-avoiding path connecting its left and right sides with the path remaining inside the rectangle. Such a path is called an (open) left-right crossing. The event that a left-right crossing exists in a rectangle $B$ is denoted by $LR(B)$. We similarly define the event, $TB(B)$, that a top-bottom crossing exists in $B$. In essence, the RSW lemma says this: if there is a significant probability that a left-right crossing exists in the square $B(n, n)$, then there is a significant probability that a left-right crossing exists in the rectangle $B(3n, n)$. More precisely, here is a version of the theorem that will be enough for our purposes. (See Exercise 4.6 for a generalization.)

**Lemma 4.58** (RSW lemma). For all $n \geq 2$ (divisible by 4) and $p \in (0, 1)$,

$$R_{n, 3}(p) \geq \frac{1}{4} R_{n, 1}(p)^{11} R_{n/2, 1}(p)^{12}. \quad (4.20)$$

The r.h.s. of (4.20) depends only on the probability of crossing a square from left to right. By a duality argument, at $p = 1/2$, it turns out that this probability is at least $1/2$ independently of $n$. Before presenting a proof of the RSW lemma, we detail this argument and finish the proof of Harris’ theorem.

**Proof of Lemma 4.57.** The point of (4.20) is that, if $R_{n, 1}(p)$ is bounded away from 0 uniformly in $n$, so is the l.h.s. By the argument in the proof of Harris’ theorem, this then implies that an open self-avoiding cycle exists in $Ann(n)$ with a probability bounded away from 0 as well. Hence to prove Lemma 4.57 it suffices to give a lower bound on $R_{n, 1}(1/2)$. It is crucial that this bound not depend on the “scale,” $n$. As it turns out, a simple duality-based symmetry argument does the trick. The following fact about $\mathbb{L}^2$ is a variant of the contour lemma (Lemma 2.17). Its proof is similar and Exercise 4.7 asks for the details (the “if” direction being the non-trivial implication).

**Lemma 4.59.** There is an open left-right crossing in the primal rectangle $[0, n + 1] \times [0, n]$ if and only if there is no closed top-bottom crossing in the dual rectangle $[1/2, n + 1/2] \times [-1/2, n + 1/2]$. 

224
Figure 4.5: Illustration of the implication \( \text{LR}(B_1') \cap \text{TB}(B_1' \cap B_2') \cap \text{LR}(B_2') \subseteq \text{LR}(B(3n, n)) \).

By symmetry, when \( p = 1/2 \), the two events in Lemma 4.59 have equal probability. So they must have probability 1/2 because they form a partition of the space of outcomes. By monotonicity, that implies \( R_{n,1}(1/2) \geq 1/2 \) for all \( n \). The RSW lemma then implies the required bound.

The proof of the RSW lemma involves a clever choice of event that relates the existence of crossings in squares and rectangles. (Combining crossings of squares into crossings of rectangles is not as trivial as it might look. Try it before reading this proof.)

**Proof of Lemma 4.58.** There are several steps in the proof.

**Step 1: it suffices to bound** \( R_{n,3/2}(p) \). We first reduce the proof to finding a bound on \( R_{n,3/2}(p) \). Let \( B_1' := B(2n, n) \) and \( B_2' := [n, 5n] \times [-n, n] \). Note that \( B_1' \cup B_2' = B(3n, n) \) and \( B_1' \cap B_2' = [n, 3n] \times [-n, n] \). Then we have the implication

\[
\text{LR}(B_1') \cap \text{TB}(B_1' \cap B_2') \cap \text{LR}(B_2') \subseteq \text{LR}(B(3n, n)).
\]

See Figure 4.5. Each event on the l.h.s. is increasing so the FKG inequality gives
$$R_{n,3}(p) \geq R_{n,2}(p)^2 R_{n,1}(p).$$

A similar argument over $B(2n, n)$ gives

$$R_{n,2}(p) \geq R_{n,3/2}(p)^2 R_{n,1}(p).$$

Combining the two, we have proved:

**Lemma 4.60** (Proof of RSW: step 1).

$$R_{n,3}(p) \geq R_{n,3/2}(p)^4 R_{n,1}(p)^3. \quad (4.21)$$

**Step 2: bounding $R_{n,3/2}(p)$** The heart of the proof is to bound $R_{n,3/2}(p)$ using an event involving crossings of squares. Let

\begin{align*}
B_1 & := B(n, n) = [-n, n] \times [-n, n], \\
B_2 & := [0, 2n] \times [-n, n], \\
B_{12} & := B_1 \cap B_2 = [0, n] \times [-n, n], \\
S & := [0, n] \times [0, n].
\end{align*}

Let $\Gamma_1$ be the event that there are paths $P_1, P_2$, where $P_1$ is a top-bottom crossing of $S$ and $P_2$ is an open self-avoiding path connecting the left side of $B_1$ to $P_1$ and stays inside $B_1$. Similarly let $\Gamma'_2$ be the event that there are paths $P'_1, P'_2$, where $P'_1$ is a top-bottom crossing of $S$ and $P'_2$ is an open self-avoiding path connecting the right side of $B_2$ to $P'_1$ and stays inside $B_2$. Then we have the implication

$$\Gamma_1 \cap LR(S) \cap \Gamma'_2 \subseteq LR(B(3n/2, n)).$$

See Figure 4.6. By symmetry $\mathbb{P}_p[\Gamma_1] = \mathbb{P}_p[\Gamma'_2]$. Moreover, the events on the l.h.s. are increasing so by the FKG inequality:

**Lemma 4.61** (Proof of RSW: step 2).

$$R_{n,3/2}(p) \geq \mathbb{P}_p[\Gamma_1]^2 R_{n,2}(p). \quad (4.22)$$

**Step 3: bounding $\mathbb{P}_p[\Gamma_1]$** It remains to bound $\mathbb{P}_p[\Gamma_1]$. That requires several additional definitions. Let $P_1$ and $P_2$ be top-bottom crossings of $S$. There is a natural partial order over such crossings. The path $P_1$ divides $S$ into two subgraphs: $\{P_1\}$ which includes the left side of $S$ (including edges on the left incident with $P_1$ but not those edges on $P_1$ itself) and $\{P_1\}$ which includes the right side of $S$ (and $P_1$ itself). Then we write $P_1 \preceq P_2$ if $\{P_1\} \subseteq \{P_2\}$. Assuming TB($S$) holds, one also gets the existence of a unique rightmost crossing. Roughly speaking, take the rightmost crossing

226
Figure 4.6: Top: illustration of the implication $\Gamma_1 \cap \text{LR}(S) \cap \Gamma_2' \subseteq \text{LR}(B(3n/2, n))$. Bottom: the event $\text{LR}^+([P^*]) \cap \{P = P_s^*\}$; the dashed path is the mirror image of the rightmost top-bottom crossing in $S$; the pink region is the complement in $B_1$ of the set $[P^*]$. Note that, because on the bottom figure the left-right path must stay within $[P^*]$, the configuration shown in the top figure where the purple left-right path “travels behind” the top-bottom crossing of $S$ cannot occur.
union of all top-bottom crossings of $S$ as sets of edges; then the “right boundary”
of this set is a top-bottom crossing $P^*_S$ such that $P^*_S \preceq P$ for all top-bottom crossings $P$ of $S$. (We accept as a fact the existence of a unique rightmost crossing. See Exercise 4.7 for a related construction.)

Let $I_S$ be the set of self-avoiding (not necessarily open) paths connecting the top and bottom of $S$ and stay inside $S$. For $P \in I_S$, we let $P'$ be the reflection of $P$ through the $x$-axis in $B_{12}\backslash S$ and we let $P^*_{PR}$ be the union of $P$ and $P'$. Define $[P^*_{PR}]$ to be the subgraph of $B_1$ to the left of $P^*_{PR}$ (including edges on the left incident with $P^*_{PR}$ but not those edges on $P^*_{PR}$ itself). Let $LR^+([P^*_{PR}])$ be the event that there is a left-right crossing of $[P^*_{PR}]$ ending on $P$, i.e., that there is an open self-avoiding path connecting the left side of $B_1$ and $P$ that stays within $[P^*_{PR}]$. See Figure 4.6. Note that the existence of a left-right crossing of $B_1$ implies the existence of an open self-avoiding path connecting the left side of $B_1$ to $P^*_{PR}$. By symmetry we then get

$$P_p[LR^+([P^*_{PR}])] \geq \frac{1}{2} P_p[LR(B_1)] = \frac{1}{2} R_{n,1}(p). \quad (4.23)$$

Now comes a subtle point. We turn to the rightmost crossing of $S$—for two reasons:

- First, by uniqueness of the rightmost crossing, $\{P^*_S = P\}_{P \in I_S}$ forms a partition of $TB(S)$. Recall that we are looking to bound a probability from below, and therefore we have to be careful not to “double count.”

- Second, the rightmost crossing has a Markov-like property. Observe that, for $P \in I_S$, the event that $\{P^*_S = P\}$ depends only the bonds in $\{P\}$. In particular it is independent of the bonds in $[P^*_{PR}]$, e.g. of the event $LR^+([P^*_{PR}])$. Hence

$$P_p[LR^+([P^*_{PR}]) \mid P^*_S = P] = P_p[LR^+([P^*_{PR}])]. \quad (4.24)$$

Note that the event $\{P^*_S = P\}$ is not increasing, as adding more open bonds can shift the rightmost crossing rightward. Therefore, we cannot use the FKG inequality here.

Combining (4.23) and (4.24), we get

$$P_p[\Gamma_1] \geq \sum_{P \in I_S} P_p[P^*_S = P] P_p[LR^+([P^*_{PR}]) \mid P^*_S = P]$$

$$\geq \frac{1}{2} R_{n,1}(p) \sum_{P \in I_S} P_p[P^*_S = P]$$

$$= \frac{1}{2} R_{n,1}(p) P_p[TB(S)]$$

$$= \frac{1}{2} R_{n,1}(p) R_{n/2,1}(p).$$
We have proved:

**Lemma 4.62** (Proof of RSW: step 3).

\[ P_p[\Gamma_1] \geq \frac{1}{2} R_{n,1}(p) R_{n/2,1}(p). \]  

(4.25)

**Step 4: putting everything together** Combining (4.21), (4.22) and (4.25) gives that

\[
R_{n,3}(p) \geq R_{n,3/2}(p)^4 R_{n,1}(p)^3 \\
\geq [P_p[\Gamma_1]^2 R_{n/2,1}(p)]^4 R_{n,1}(p)^3 \\
\geq \left[ \left( \frac{1}{2} R_{n,1}(p) R_{n/2,1}(p) \right)^2 R_{n/2,1}(p) \right]^4 R_{n,1}(p)^3.
\]

Collecting the terms concludes the proof of the RSW lemma.

**Remark 4.63.** This argument is quite subtle. It is instructive to read the remark after [Gri97, Theorem 9.3].

### 4.4 Couplings of Markov chains

As we have seen, coupling is useful to bound total variation distance. A natural application is mixing as we show in this section.

#### 4.4.1 Bounding the mixing time via coupling

Let \( P \) be an irreducible, aperiodic Markov transition matrix on the finite state space \( V \) with stationary distribution \( \pi \). Recall that, for a fixed \( 0 < \varepsilon < 1/2 \), the mixing time of \( P \) is

\[ t_{\text{mix}}(\varepsilon) := \min \{ t : d(t) \leq \varepsilon \}, \]

where

\[ d(t) := \max_{x \in V} \| P^t(x, \cdot) - \pi \|_{TV}. \]

It will be easier to work with

\[ \bar{d}(t) := \max_{x,y \in V} \| P^t(x, \cdot) - P^t(y, \cdot) \|_{TV}. \]

The quantities \( d(t) \) and \( \bar{d}(t) \) are related in the following way.
Lemma 4.64.

\[ d(t) \leq \bar{d}(t) \leq 2d(t), \quad \forall t. \]

Proof. The second inequality follows from an application of the triangle inequality. For the first inequality, note that by definition of the total variation distance

\[ \| P^t(x, \cdot) - \pi \|_{TV} = \sup_{A \subseteq V} | P^t(x, A) - \pi(A) | \]

\[ = \sup_{A \subseteq V} \left| \sum_{y \in V} \pi(y) [ P^t(x, A) - P^t(y, A) ] \right| \]

\[ \leq \sup_{A \subseteq V} \sum_{y \in V} \pi(y) | P^t(x, A) - P^t(y, A) | \]

\[ \leq \sum_{y \in V} \pi(y) \left\{ \sup_{A \subseteq V} | P^t(x, A) - P^t(y, A) | \right\} \]

\[ \leq \sum_{y \in V} \pi(y) \| P^t(x, \cdot) - P^t(y, \cdot) \|_{TV} \]

\[ \leq \max_{x,y \in V} \| P^t(x, \cdot) - P^t(y, \cdot) \|_{TV}. \]

A coupling of Markov chains with transition matrix P is a Markov chain \((X_t, Y_t)\) on \(V \times V\) such that both \((X_t)\) and \((Y_t)\) are Markov chains with transition matrix \(P\). For our purposes, the following special type of coupling will suffice.

Definition 4.65 (Markovian coupling). A Markovian coupling of \(P\) is a Markov chain \((X_t, Y_t)\) on \(V \times V\) with transition matrix \(Q\) satisfying:

- (Markovian coupling) For all \(x, y, x', y' \in V\),

\[ \sum_{z'} Q((x, y), (x', z')) = P(x, x'), \]

\[ \sum_{z'} Q((x, y), (z', y')) = P(y, y'). \]

We say that a Markovian coupling is coalescing if further:

- (Coalescing) For all \(z \in V\),

\[ x' \neq y' \implies Q((z, z), (x', y')) = 0. \]
Let \((X_t, Y_t)\) be a coalescing Markovian coupling of \(P\). By the coalescing condition, if \(X_s = Y_s\) then \(X_t = Y_t\) for all \(t \geq s\). That is, once \((X_t)\) and \((Y_t)\) meet, they remain equal. Let \(\tau_{\text{coal}}\) be the coalescence time (also called coupling time), i.e.,
\[
\tau_{\text{coal}} := \inf\{t \geq 0 : X_t = Y_t\}.
\]
The key to the coupling approach to mixing times is the following immediate consequence of the coupling inequality (Lemma 4.9). For any starting point \((x, y)\),
\[
\|P^t(x, \cdot) - P^t(y, \cdot)\|_{\text{TV}} \leq P(x, y)[X_t \neq Y_t] = P(x, y)[\tau_{\text{coal}} > t]. \tag{4.26}
\]
Combining (4.26) and Lemma 4.64, we get the main result of this section.

**Theorem 4.66** (Bounding the mixing time: coupling method). Let \((X_t, Y_t)\) be a coalescing Markovian coupling of an irreducible transition matrix \(P\) on a finite state space \(V\) with stationary distribution \(\pi\). Then
\[
d(t) \leq \max_{x, y \in V} P(x, y)[\tau_{\text{coal}} > t].
\]
In particular
\[
t_{\text{mix}}(\varepsilon) \leq \inf \{ t \geq 0 : P(x, y)[\tau_{\text{coal}} > t] \leq \varepsilon, \forall x, y \}.
\]
We give a few simple examples in the next subsection.

### 4.4.2 Markov chains: mixing on cycles, hypercubes, and trees

In this section, we consider lazy simple random walk on various graphs. By this we mean that the walk stays put with probability 1/2 and otherwise picks an adjacent vertex uniformly at random. In each case, we construct a coupling to bound the mixing time. As a reference, we compare our upper bounds to the diameter-based lower bound derived in Section 2.4.7. Recall that, by Claim 2.74, for a finite, reversible Markov chain with stationary distribution \(\pi\) and diameter \(\Delta\) we have the lower bound
\[
t_{\text{mix}}(\varepsilon) = \Omega \left( \frac{\Delta^2}{\log(n \lor \pi_{\text{min}}^{-1})} \right),
\]
for \(\varepsilon > 0\), where \(\pi_{\text{min}}\) is the smallest value taken by \(\pi\).
**Cycle**  Let \((Z_t)\) be lazy simple random walk on the cycle of size \(n\), \(Z_n := \{0, 1, \ldots, n - 1\}\), where \(i \sim j\) if \(|j - i| = 1 \pmod n\). For any starting points \(x, y\), we construct a coalescing Markovian coupling \((X_t, Y_t)\) of this chain. Set \((X_0, Y_0) := (x, y)\). At each time, flip a fair coin. On heads, \(Y_t\) stays put and \(X_t\) moves one step, the direction of which is uniform at random. On tails, proceed similarly with the roles of \(X_t\) and \(Y_t\) reversed. Let \(D_t\) be the clockwise distance between \(X_t\) and \(Y_t\). Observe that, by construction, \((D_t)\) is simple random walk on \(\{0, 1, \ldots, n\}\) and \(\tau_{\text{coal}} = \tau_{\{0,n\}}\), the first time \((D_t)\) hits \(\{0, n\}\).

We use Markov’s inequality, Theorem 2.1, to bound \(\mathbb{P}_{(x,y)}[\tau_{\{0,n\}} > t]\). Denote by \(D_0 = d_{x,y}\) the starting distance. By Wald’s second equation (e.g. [Dur10, Example 4.1.6]),

\[
\mathbb{E}_{(x,y)}[\tau_{\{0,n\}}] = d_{x,y}(n - d_{x,y}).
\]

Applying Theorem 4.66 and Markov’s inequality we get

\[
d(t) \leq \max_{x,y\in V} \mathbb{P}_{(x,y)}[\tau_{\text{coal}} > t] \\
\leq \max_{x,y\in V} \frac{\mathbb{E}_{(x,y)}[\tau_{\{0,n\}}]}{t} \\
= \max_{x,y\in V} \frac{d_{x,y}(n - d_{x,y})}{t} \\
\leq \frac{n^2}{4t},
\]

or:

**Claim 4.67.**

\[
t_{\text{mix}}(\varepsilon) \leq \frac{n^2}{4\varepsilon}.
\]

By our diameter-based lower bound on mixing in Section 2.4.7, this bound gives the correct order of magnitude in \(n\) up to logarithmic factors. Indeed, the diameter is \(\Delta = n/2\) and \(\pi_{\text{min}} = 1/n\) so that Claim 2.74 gives

\[
t_{\text{mix}}(\varepsilon) \geq \frac{n^2}{64 \log n},
\]

for \(n\) large enough. Exercise 4.9 sketches a tighter lower bound.

**Hypercube**  Let \((Z_t)\) be lazy simple random walk on the \(n\)-dimensional hypercube \(Z_n^2 := \{0, 1\}^n\) where \(i \sim j\) if \(||i - j||_1 = 1\). We denote the coordinates of \(Z_t\) by \((Z_t^{(1)}, \ldots, Z_t^{(n)})\). The coupling \((X_t, Y_t)\) started at \((x, y)\) is the following.
At each time \( t \), pick a coordinate \( i \) uniformly at random in \( [n] \), pick a bit value \( b \) in \( \{0, 1\} \) uniformly at random independently of the coordinate choice. Set both \( i \) coordinates to \( b \), i.e., \( X_t^{(i)} = Y_t^{(i)} = b \). This is equivalent to performing the Glauber dynamics chain on an empty graph. Because of the way the updating is done, the chains stay put with probability \( 1/2 \) at each time as required. Clearly the chains coalesce when all coordinates have been updated at least once. The following standard bound on the coupon collector problem is what is needed to conclude.

**Lemma 4.68 (Coupon collecting).** Let \( \tau_{\text{coll}} \) be the time it takes to update each coordinate at least once. Then, for any \( c > 0 \),

\[
P[\tau_{\text{coll}} > \lceil n \log n + cn \rceil] \leq e^{-c}.
\]

**Proof.** Let \( B_i \) be the event that the \( i \)-th coordinate has not been updated by time \( \lceil n \log n + cn \rceil \). Then

\[
P[\tau_{\text{coll}} > \lceil n \log n + cn \rceil] \leq \sum_i P[B_i] = \sum_i \left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil} \leq n \exp\left(-\frac{n \log n + cn}{n}\right) = e^{-c}.
\]

Applying Theorem 4.66, we get

\[
d(\lceil n \log n + cn \rceil) \leq \max_{x,y \in V} P(x,y[\tau_{\text{coal}} > \lceil n \log n + cn \rceil] \leq P[\tau_{\text{coll}} > \lceil n \log n + cn \rceil] \leq e^{-c}.
\]

Hence for \( c := c_\varepsilon > 0 \) large enough:

**Claim 4.69.**

\[
t_{\text{mix}}(\varepsilon) \leq \lceil n \log n + c_\varepsilon n \rceil.
\]

Again we get a quick lower bound using our diameter-based result from Section 2.4.7. Here \( \Delta = n \) and \( \pi_{\text{min}} = 1/2^n \) so that Claim 2.74 gives

\[
t_{\text{mix}}(\varepsilon) \geq \frac{n^2}{12 \log n + (4 \log 2)n} = \Omega(n),
\]

233
for $n$ large enough. So the upper bound we derived above is off at most by a logarithmic factor in $n$. In fact:

**Claim 4.70.**

$$t_{\text{mix}}(\epsilon) \geq \frac{1}{2} n \log n - O(n).$$

**Proof.** For simplicity, we assume that $n$ is odd. Let $W_t$ be the number of 1s, or Hamming weight, at time $t$. Let $A$ be the event that the Hamming weight is $\leq n/2$. To bound the mixing time, we use the fact that for any $z_0$

$$d(t) \geq \|P^t(z_0, \cdot) - \pi\|_{\text{TV}} \geq |P^t(z_0, A) - \pi(A)|.$$  (4.27)

Under the stationary distribution, the Hamming weight is equal in distribution to a $\text{Bin}(n, 1/2)$. In particular the probability that a majority of coordinates are 0 is $1/2$. That is, $\pi(A) = 1/2$.

On the other hand, let $(Z_t)$ start at $z_0 = 1$, where 1 is the all-1 vector. Let $U_t$ be the number of updated coordinates up to time $t$ in the Glauber dynamics representation of the chain discussed above. We use Chebyshev’s inequality (Theorem 2.2) to bound the probability of event $A$ at time $t$. We first need to compute the expectation and variance of $W_t$. Observe that, conditioned on $U_t$, the Hamming weight $W_t$ is equal in distribution to $\text{Bin}(U_t, 1/2) + (n - U_t)$ as the updated coordinates are uniform and the other ones are 1. Thus we have

$$\mathbb{E}[W_t] = \mathbb{E}[\mathbb{E}[W_t \mid U_t]]$$  
$$= \mathbb{E} \left[ \frac{1}{2} U_t + (n - U_t) \right]$$  
$$= n - \frac{1}{2} n \left[ 1 - \left( 1 - \frac{1}{n} \right)^t \right]$$  
$$= \frac{n}{2} \left[ 1 + \left( 1 - \frac{1}{n} \right)^t \right],$$  (4.28)

where on the third line we used that $\mathbb{E}[U_t] = n \left[ 1 - \left( 1 - \frac{1}{n} \right)^t \right]$ by linearity of expectation, and

$$\text{Var}[W_t] = \mathbb{E}[\text{Var}[W_t \mid U_t]] + \text{Var}[\mathbb{E}[W_t \mid U_t]]$$  
$$= \frac{1}{4} \mathbb{E}[U_t] + \frac{1}{4} \text{Var}[U_t].$$  (4.29)
It remains to compute $\text{Var}[U_t]$. Let $I_{t}^{(i)}$ be 1 if coordinate $i$ has not been updated up to time $t$ and 0 otherwise. Note that for $i \neq j$

$$\text{Cov}[I_{t}^{(i)}, I_{t}^{(j)}] = \mathbb{E}[I_{t}^{(i)} I_{t}^{(j)}] - \mathbb{E}[I_{t}^{(i)}] \mathbb{E}[I_{t}^{(j)}]$$

$$= \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{1}{n}\right)^{2t}$$

$$= \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{2}{n} + \frac{1}{n^2}\right)^t$$

$$\leq 0,$$

that is, $I_{t}^{(i)}$ and $I_{t}^{(j)}$ are negatively correlated, while

$$\text{Var}[I_{t}^{(i)}] = \mathbb{E}[(I_{t}^{(i)})^2] - (\mathbb{E}[I_{t}^{(i)}])^2 \leq \mathbb{E}[I_{t}^{(i)}] = \left(1 - \frac{1}{n}\right)^t.$$

Then, writing $n - U_t$ as the sum of these indicators, we have

$$\text{Var}[U_t] = \text{Var}[n - U_t] = \sum_{i=1}^{n} \text{Var}[I_{t}^{(i)}] + 2 \sum_{i<j} \text{Cov}[I_{t}^{(i)}, I_{t}^{(j)}] \leq n \left(1 - \frac{1}{n}\right)^t.$$

Plugging this back into (4.29), we get

$$\text{Var}[W_t] \leq \frac{n}{4} \left[1 - \left(1 - \frac{1}{n}\right)^t\right] + \frac{n}{4} \left(1 - \frac{1}{n}\right)^t = \frac{n}{4}.$$

For $t_\alpha = \frac{1}{2} n \log n - (\log 2\alpha)n$ with $\alpha > 0$, by (4.28),

$$\mathbb{E}[W_{t_\alpha}] = \frac{n}{2} + e^{t_\alpha(-1/n+\Theta(1/n^2))} = \frac{n}{2} + \alpha \sqrt{n} + o(1),$$

where we used that by a Taylor expansion, for $|z| \leq 1/2$, $\log (1 - z) = -z + \Theta(z^2)$. Fix $0 < \varepsilon < 1/2$. By Chebyshev’s inequality (Theorem 2.2), for $t_\alpha = \frac{1}{2} n \log n - (\log 2\alpha)n$ and $n$ large enough,

$$\mathbb{P}[W_{t_\alpha} \leq n/2] \leq \mathbb{P}[|W_{t_\alpha} - \mathbb{E}[W_{t_\alpha}]| \geq (\alpha/2)\sqrt{n}] \leq \frac{n/4}{(\alpha/2)^2 n} \leq \frac{1}{2} - \varepsilon,$$

for $\alpha$ large enough. By (4.27), that implies $d(t_\alpha) \geq \varepsilon$ and we are done. \hfill \blacksquare

The proof of Claim 4.70 relies on a “distinguishing statistic.” Recall from Lemma 4.17 that for any random variables $X, Y$ and map $h$ it holds that $\|\mu_h(X) - \mu_h(Y)\|_{TV} \leq \frac{n}{4}$.
\[\|\mu_X - \mu_Y\|_{TV}\], where \(\mu_Z\) is the law of \(Z\). The map used in the proof of the claim is the Hamming weight. In essence, we gave a lower bound on the total variation distance between the laws of the Hamming weight at stationarity and under \(P^t(z_0, \cdot)\). See Exercise 4.10 for a more general instantiation of the distinguishing statistic approach.

**Remark 4.71.** The upper bound in Claim 4.69 is indeed off by a factor of 2. See [LPW06, Theorem 18.3] for an improved upper bound and a discussion of the so-called cutoff phenomenon. The latter refers to the fact that for all \(0 < \varepsilon < 1/2\) it can be shown that

\[
\lim_{n \to +\infty} \frac{t_{\text{mix}}^n(\varepsilon)}{t_{\text{mix}}^n(1 - \varepsilon)} = 1,
\]

where \(t_{\text{mix}}^n(\varepsilon)\) is the mixing time on the \(n\)-dimensional hypercube. In words, for large \(n\), the total variation distance drops from 1 to 0 in a short time window. See Exercise 5.2 for a necessary condition for cutoff.

**b-ary tree**  Let \((Z_t)\) be lazy simple random walk on the \(\ell\)-level rooted \(b\)-ary tree, \(\hat{T}_{b,\ell}\). The root, 0, is on level 0 and the leaves, \(L\), are on level \(\ell\). All vertices have degree \(b + 1\), except for the root which has degree \(b\) and the leaves which have degree 1. Hence the stationary distribution is

\[\pi(x) := \frac{\delta(x)}{2(n - 1)},\]

where \(n\) is the number of vertices and \(\delta(x)\) is the degree of \(x\). We construct a coupling \((X_t, Y_t)\) of this chain started at \((x, y)\). Assume w.l.o.g. that \(x\) is no further from the root than \(y\), which we denote by \(x \preceq y\). The coupling has two stages:

- In the first stage, at each time, flip a fair coin. On heads, \(Y_t\) stays put and \(X_t\) moves one step chosen uniformly at random among its neighbors. Similarly, on tails, reverse the roles of \(X_t\) and \(Y_t\) and proceed in the same manner. Do this until \(X_t\) and \(Y_t\) are on the same level.

- In the second stage, i.e., once the two chains are on the same level, at each time first let \(X_t\) move as a lazy random walk on \(\hat{T}_{b,\ell}\). Then let \(Y_t\) move in the same direction as \(X_t\), i.e., if \(X_t\) moves closer to the root, so does \(Y_t\) and so on.

By construction, \(X_t \preceq Y_t\) for all \(t\). The key observation is the following. Let \(\tau^*\) be the first time \((X_t)\) visits the root after visiting the leaves. By time \(\tau^*\), the two chains have necessarily met: because \(X_t \preceq Y_t\), when \(X_t\) reaches the leaves, so does
Y_t; after that time, the coupling is in the second stage so X_t and Y_t remain on the same level; in particular, when X_t reaches the root, so does Y_t. Hence τ_{coal} ≤ τ^*. Intuitively, the mixing time is indeed dominated by the time it takes to reach the root from the worst starting point, a leaf. See Figure 4.7 and the corresponding lower bound argument.

To estimate P_{(x,y)}[τ^* > t], we use Markov’s inequality (Theorem 2.1), for which we need a bound on E_{(x,y)}[τ^*]. We note that E_{(x,y)}[τ^*] is less than the mean time for the walk to go from the root to the leaves and back. Let L_t be the level of X_t and let N be the corresponding network (where the conductances are equal to the number of edges on each level of the tree). In terms of L_t, the quantity we seek to bound is the mean of τ_{0,ℓ}, the commute time of the chain (L_t) between the states 0 and ℓ. By the commute time identity (Theorem 3.115),

\[ E_0[τ_{0,ℓ}] = c_N \hat{R}(0 \leftrightarrow ℓ), \tag{4.30} \]

where

\[ c_N = 2 \sum_{e=(x,y) \in N} c(e) = 4(n - 1), \]

where we simply counted the number of edges in \( \hat{T}_{b,ℓ} \) and the factor of 4 accounts for self-loops. Using network reduction techniques, we computed the effective resistance \( \hat{R}(0 \leftrightarrow ℓ) \) in Examples 3.96 and 3.97—without self-loops. Of course adding self-loops does not affect the effective resistance as we can use the same voltage and current. So, ignoring them, we get

\[ \hat{R}(0 \leftrightarrow ℓ) = \sum_{j=0}^{ℓ-1} b^{-(j+1)} = \sum_{j=0}^{ℓ-1} b^{-j+1} = \frac{1}{b^ℓ} \cdot \frac{b^{-ℓ} - 1}{b-1}, \]

which implies

\[ \frac{1}{b} \leq \hat{R}(0 \leftrightarrow ℓ) \leq \frac{1}{b-1} \leq 1. \]

Finally, applying Theorem 4.66 and Markov’s inequality and using (4.30), we get

\[ d(t) \leq \max_{x,y \in V} P_{(x,y)}[τ^* > t] \]

\[ \leq \max_{x,y \in V} \frac{E_{(x,y)}[τ^*]}{t} \]

\[ \leq \frac{E_0[τ_{0,ℓ}]}{t} \]

\[ \leq \frac{4n}{t}, \]

or:

237
Figure 4.7: Setup for the lower bound on the mixing time on a $b$-ary tree. (Here $b = 2$.)

Claim 4.72.

\[ t_{\text{mix}}(\varepsilon) \leq \frac{4n}{\varepsilon}. \]

This time the diameter-based bound is far off. We have $\Delta = 2\ell = \Theta(\log n)$ and $\pi_{\text{min}} = 1/2(n - 1)$ so that Claim 2.74 gives

\[ t_{\text{mix}}(\varepsilon) \geq \frac{(2\ell)^2}{12 \log n + 4 \log(2(n - 1))} = \Omega(\log n), \]

for $n$ large enough. Here is a better lower bound. Intuitively the mixing time is significantly greater than the squared diameter because the chain tends to be pushed away from the root: going from the leaves on one side of the root to the leaves on the other typically takes time linear in $n$. Formally let $x_0$ be a leaf of $T_{b,\ell}$ and let $A$ be the set of vertices “on the other side of root (inclusively),” i.e., vertices whose graph distance from $x_0$ is at least $\ell$. See Figure 4.7. Then $\pi(A) \geq 1/2$ by symmetry. We use the fact that

\[ \|P^t(x_0, \cdot) - \pi\|_{TV} \geq |P^t(x_0, A) - \pi(A)|, \]

to bound the mixing time. We claim that, started at $x_0$, the walk typically takes time linear in $n$ to reach $A$. Consider again the level $L_\ell$ of $X_\ell$. Using the expression for
the effective resistance above, we have
\[ P_\ell [\tau_0 < \tau_+^\ell] = \frac{1}{c(\ell)} \mathbb{P}(0 \leftrightarrow \ell) = \frac{1}{2b^\ell} \cdot \frac{b - 1}{1 - b^{-\ell}} = \frac{b - 1}{2b^\ell - 2} = O \left( \frac{1}{n} \right). \]
Hence, started from the leaves, the number of excursions back to the leaves needed to reach the root for the first time is geometric with success probability \( O(n^{-1}) \). Each such excursion takes time at least \( 2 \). So \( P^t(x_0, A) \) is bounded above by the probability that at least one such excursion was successful among the first \( t/2 \) attempts. That is,
\[ P^t(x_0, A) \leq 1 - \left( 1 - O\left( n^{-1} \right) \right)^{t/2} < \frac{1}{2} - \varepsilon, \]
for all \( t \leq \alpha \varepsilon n \) with \( \alpha \varepsilon > 0 \) small enough and
\[ \| P^{\alpha \varepsilon n}(x_0, \cdot) - \pi \|_{TV} \geq | P^{\alpha \varepsilon n}(x_0, A) - \pi(A) | > \varepsilon. \]
We have proved that \( t_{\text{mix}}(\varepsilon) \geq \alpha \varepsilon n. \)

### 4.4.3 Path coupling

**Path coupling** is a method for constructing Markovian couplings from “simpler” couplings. The building blocks are one-step couplings starting from pairs of initial states that are close in some dissimilarity graph.

Let \((X_t)\) be an irreducible Markov chain on a finite state space \( V \) with transition matrix \( P \) and stationary distribution \( \pi \). Assume that we are given a dissimilarity graph \( H_0 = (V_0, E_0) \) on \( V_0 := V \) with edge weights \( w_0 : E_0 \to \mathbb{R}_+ \). This graph need not have the same edges as the transition graph of \((X_t)\). We extend \( w_0 \) to the path metric
\[ w_0(x, y) := \inf \left\{ \sum_{i=0}^{m-1} w_0(x_i, x_{i+1}) : x = x_0, \ldots, x_m = y \text{ is a path in } H_0 \right\}, \]
where the infimum is over all paths connecting \( x \) and \( y \) in \( H_0 \). We call a path achieving the infimum a **minimum-weight path**. Let
\[ \Delta_0 := \max_{x, y} w_0(x, y), \]
be the weighted diameter of \( H_0 \).

**Theorem 4.73** (Path coupling method). Assume that
\[ w_0(u, v) \geq 1, \]
for all \( \{u, v\} \in E_0 \). Assume further that there exists \( \kappa \in (0, 1) \) such that:
For all \(x, y\) with \(\{x, y\} \in E_0\), there is a coupling \((X^*, Y^*)\)
of \(P(x, \cdot)\) and \(P(y, \cdot)\) satisfying the following contraction property
\[
E[w_0(X^*, Y^*)] \leq \kappa w_0(x, y).
\] (4.31)

Then
\[
d(t) \leq \Delta_0 \kappa^t,
\]
or
\[
t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{\log \Delta_0 + \log \varepsilon^{-1}}{\log \kappa} \right\rceil.
\]

Proof. The crux of the proof is to extend (4.31) to arbitrary pairs of vertices.

Claim 4.74 (Global coupling). For all \(x, y \in V\), there is a coupling \((X^*, Y^*)\) of \(P(x, \cdot)\) and \(P(y, \cdot)\) such that (4.31) holds.

Iterating the coupling in this claim immediately implies the existence of a coalescing Markovian coupling \((X_t, Y_t)\) of \(P\) such that
\[
E_{(x, y)}[w_0(X_t, Y_t)] = E_{(x, y)}[E[w_0(X_t, Y_t) | X_{t-1}, Y_{t-1}]] \\
\leq E_{(x, y)}[\kappa w_0(X_{t-1}, Y_{t-1})] \\
\leq \cdots \\
\leq \kappa^t E_{(x, y)}[w_0(X_0, Y_0)] \\
= \kappa^t w_0(x, y) \\
\leq \kappa^t \Delta_0.
\]

By assumption, \(1_{\{x \not= y\}} \leq w_0(x, y)\) so that by the coupling inequality and Lemma 4.64, we have
\[
d(t) \leq \overline{d}(t) \leq \max_{x,y} P_{(x, y)}[X_t \not= Y_t] \leq \max_{x,y} E_{(x, y)}[w_0(X_t, Y_t)] \leq \kappa^t \Delta_0,
\]
which implies the theorem. It remains to prove Claim 4.74.

Proof of Claim 4.74. Fix \(x', y' \in V\) such that \(\{x', y'\}\) is not an edge in the dissimilarity graph \(H_0\). The idea is to combine the local couplings on a minimum-weight path between \(x'\) and \(y'\) in \(H_0\). Let \(x' = x_0 \sim \cdots \sim x_m = y'\) be such a path. For all \(i = 0, \ldots, m-1\), let \((Z_{i,0}^*, Z_{i,1}^*)\) be a coupling of \(P(x_i, \cdot)\) and \(P(x_{i+1}, \cdot)\) satisfying the contraction property (4.31). Set \(Z^{(0)} := Z_{0,0}^*\) and \(Z^{(1)} := Z_{1,0}^*\). Then iteratively pick \(Z^{(i+1)}\) according to the law \(P[Z_{i,1}^* \in \cdot | Z_{i,0}^* = Z^{(i)}]\). By induction on
Figure 4.8: Coupling of \( P(x', \cdot) \) and \( P(y', \cdot) \) constructed from a sequence of local couplings \((Z_{0,0}, Z_{0,1}), \ldots, (Z_{0,m-1}, Z_{0,m-1})\).

\( i, (X^*, Y^*) := (Z^{(0)}, Z^{(m)}) \) is then a coupling of \( P(x', \cdot) \) and \( P(y', \cdot) \). Formally, define the transition matrix

\[
R_i(z^{(i)}, z^{(i+1)}) := \mathbb{P}[Z_{i,1}^* = z^{(i+1)} | Z_{i,0}^* = z^{(i)}],
\]

and observe that

\[
\sum_{z^{(i+1)}} R_i(z^{(i)}, z^{(i+1)}) = 1, \tag{4.32}
\]

and

\[
\sum_{z^{(i)}} P(x_{i}, z^{(i)}) R_i(z^{(i)}, z^{(i+1)}) = P(x_{i+1}, z^{(i+1)}), \tag{4.33}
\]

by construction of the coupling \((Z_{i,0}^*, Z_{i,1}^*)\). See Figure 4.8. The law of the full coupling \((Z^{(0)}, \ldots, Z^{(m)})\) is

\[
\mathbb{P}[(Z^{(0)}, \ldots, Z^{(m)}) = (z^{(0)}, \ldots, z^{(m)})] = P(x_0, z^{(0)}) R_0(z^{(0)}, z^{(1)}) \cdots R_{m-1}(z^{(m-1)}, z^{(m)}).
\]

Using (4.32) and (4.33) inductively gives

\[
\mathbb{P}[X^* = z^{(0)}] = \mathbb{P}[Z^{(0)} = z^{(0)}] = P(x_0, z^{(0)}),
\]

\[
\mathbb{P}[Y^* = z^{(m)}] = \mathbb{P}[Z^{(m)} = z^{(m)}] = P(x_m, z^{(m)}),
\]

241
as required.

By the triangle inequality for \( w_0 \), the coupling \((X^*, Y^*)\) satisfies

\[
E[w_0(X^*, Y^*)] = E\left[w_0(Z^{(0)}, Z^{(m)})\right] \\
\leq \sum_{i=0}^{m-1} E\left[w_0(Z^{(i)}, Z^{(i+1)})\right] \\
\leq \sum_{i=0}^{m-1} \kappa w_0(x_i, x_{i+1}) \\
= \kappa w_0(x', y'),
\]

where, on the third line, we used (4.31) for adjacent pairs and the last line follows from the fact that we chose a minimum-weight path. 

That concludes the proof of the theorem.

We illustrate the path coupling method in the next two subsections. See Exercise 4.11 for an optimal transport perspective on the path coupling method.

### 4.4.4 Ising model: Glauber dynamics at high temperature

Let \( G = (V, E) \) be a finite, connected graph with maximal degree \( \bar{\delta} \). Define \( \mathcal{X} := \{-1, +1\}^V \). Recall that the (ferromagnetic) Ising model on \( V \) with inverse temperature \( \beta \) is the probability distribution over spin configurations \( \sigma \in \mathcal{X} \) given by

\[
\mu_\beta(\sigma) := \frac{1}{Z(\beta)} e^{-\beta \mathcal{H}(\sigma)},
\]

where

\[
\mathcal{H}(\sigma) := -\sum_{i\sim j} \sigma_i \sigma_j,
\]

is the Hamiltonian and

\[
Z(\beta) := \sum_{\sigma \in \mathcal{X}} e^{-\beta \mathcal{H}(\sigma)},
\]

is the partition function. In this context, recall that vertices are often referred to as sites. The single-site Glauber dynamics of the Ising model is the Markov chain on \( \mathcal{X} \) which, at each time, selects a site \( i \in V \) uniformly at random and updates the spin \( \sigma_i \) according to \( \mu_\beta(\sigma) \) conditioned on agreeing with \( \sigma \) at all sites in \( V \setminus \{i\} \). Specifically, for \( \gamma \in \{-1, +1\}, i \in V, \) and \( \sigma \in \mathcal{X}, \) let \( \sigma_i^{\gamma} \) be the configuration \( \sigma \)
with the state at $i$ being set to $\gamma$. Then, letting $n = |V|$, the transition matrix of the Glauber dynamics is

$$Q_\beta(\sigma, \sigma^{i, \gamma}) := \frac{1}{n} \cdot \frac{e^{\gamma S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}} = \frac{1}{n} \left\{ 1 + \frac{1}{2} \tanh(\gamma \beta S_i(\sigma)) \right\}, \quad (4.34)$$

where

$$S_i(\sigma) := \sum_{j \sim i} \sigma_j.$$

All other transitions have probability 0. Recall that this chain is irreducible and reversible with respect to $\mu_\beta$. In particular $\mu_\beta$ is the stationary distribution of $Q_\beta$.

In this section, we give an upper bound on the mixing time, $t_{\text{mix}}(\epsilon)$, of $Q_\beta$ using path coupling. We say that the Glauber dynamics is fast mixing if $t_{\text{mix}}(\epsilon) = O(n \log n)$. We first make a simple observation:

**Claim 4.75 (Glauber dynamics: lower bound on mixing).**

$$t_{\text{mix}}(\epsilon) = \Omega(n), \quad \forall \beta > 0.$$  

**Proof.** We use a coupon collecting argument. Let $\bar{\sigma}$ be the all-$(-1)$ configuration and let $A$ be the set of configurations where at least half of the sites are $+1$. Then, by symmetry, $\mu_\beta(A) = \mu_\beta(A^c) = 1/2$ where we assumed for simplicity that $n$ is odd. By definition of the total variation distance

$$d(t) \geq \|Q_\beta^t(\bar{\sigma}, \cdot) - \mu_\beta(\cdot)\|_{TV} \geq |Q_\beta^t(\bar{\sigma}, A) - \mu_\beta(A)| = |Q_\beta^t(\bar{\sigma}, A) - 1/2|.$$  

So it remains to show that by time $cn$, for $c > 0$ small, the chain is unlikely to have reached $A$. That happens if, say, fewer than a third of the sites have been updated. Using the notation of Example 2.4, we are seeking a bound on $T_{n,n/3}$, i.e., the time to collect $n/3$ coupons out of $n$. We can write this random variable as a sum of $n/3$ independent geometric variables $T_{n,n/3} = \sum_{i=1}^{n/3} \tau_{n,i}$, where $E[\tau_{n,i}] = (1 - \frac{i-1}{n})^{-1}$ and $\text{Var}[\tau_{n,i}] = (1 - \frac{i-1}{n})^{-2}$. Hence

$$E[T_{n,n/3}] = \sum_{i=1}^{n/3} \left( 1 - \frac{i-1}{n} \right)^{-1} = n \sum_{j=2n/3+1}^{n} j^{-1} = \Theta(n), \quad (4.35)$$

and

$$\text{Var}[T_{n,n/3}] \leq \sum_{i=1}^{n/3} \left( 1 - \frac{i-1}{n} \right)^{-2} = n^2 \sum_{j=2n/3+1}^{n} j^{-2} = \Theta(n). \quad (4.36)$$
So by Chebyshev’s inequality (Theorem 2.2)
\[
P[|T_{n,n/3} - E[T_{n,n/3}]| \geq \varepsilon n] \leq \frac{\text{Var}[T_{n,n/3}]}{(\varepsilon n)^2} \to 0,
\]
by (4.35) and (4.36). Taking \( \varepsilon > 0 \) small enough and \( n \) large enough, we have shown that, for \( t \leq c_\varepsilon n \) for some \( c_\varepsilon > 0 \),
\[
Q_\beta^t(\bar{\sigma}, A) \leq 1/3,
\]
which proves the claim.

\[\blacksquare\]

**Remark 4.76.** In fact, Ding and Peres proved that \( t_{\text{mix}}(\varepsilon) = \Omega(n \log n) \) for any graph on \( n \) vertices [DP11]. In Claim 4.70, we treated the special case of the empty graph, which is equivalent to lazy random walk on the hypercube.

In our main result, we show that the Glauber dynamics of the Ising model is fast mixing when the inverse temperature \( \beta \) is small enough as a function of the maximum degree.

**Claim 4.77** (Glauber dynamics: fast mixing at high temperature).

\[
\beta < \bar{\delta}^{-1} \implies t_{\text{mix}}(\varepsilon) = O(n \log n).
\]

**Proof.** We use path coupling. Let \( H_0 = (V_0, E_0) \) where \( V_0 := \mathcal{X} \) and \( \{\sigma, \omega\} \in E_0 \) if \( \frac{1}{2}||\sigma - \omega||_1 = 1 \) with unit \( w_0 \)-weights on all edges. (To avoid confusion, we reserve the notation \( \sim \) for adjacency in \( G \).) Let \( \{\sigma, \omega\} \in E_0 \) differ at coordinate \( i \).

We construct a coupling \( (X^*, Y^*) \) of \( Q_\beta(\sigma, \cdot) \) and \( Q_\beta(\omega, \cdot) \). We first pick the same coordinate \( i_* \) to update. If \( i_* \) is such that all its neighbors in \( G \) have the same state in \( \sigma \) and \( \omega \), i.e., if \( \sigma_j = \omega_j \) for all \( j \sim i_* \), we update \( X^* \) from \( \sigma \) according to the Glauber rule and set \( Y^* := X^* \). Note that this includes the case \( i_* = i \). Otherwise, i.e. if \( i_* \sim i \), we proceed as follows. From the state \( \sigma \), the probability of updating site \( i_* \) to state \( \gamma \in \{-1, +1\} \) is given by the expression in brackets in (5.6), and similarly for \( \omega \). Unlike the previous case, we cannot guarantee that the update is identical in both chains. However, in order to minimize the chances of increasing the distance between the two chains, we pick a uniform \([-1, 1]\) variable \( U \) and set

\[
X_{i_*}^* := \begin{cases} 
+1, & \text{if } U \leq \tanh(\beta S_{i_*}(\sigma)), \\
-1, & \text{o.w.}
\end{cases}
\]

and

\[
Y_{i_*}^* := \begin{cases} 
+1, & \text{if } U \leq \tanh(\beta S_{i_*}(\omega)), \\
-1, & \text{o.w.}
\end{cases}
\]
We set \( X_j^* := \sigma_j \) and \( Y_j^* := \omega_j \) for all \( j \neq i^* \). The expected distance between \( X^* \) and \( Y^* \) is then

\[
E[w_0(X^*, Y^*)] = 1 - \frac{1}{n} + \frac{1}{n} \sum_{j \sim i} \frac{1}{2} |\tanh(\beta S_j(\sigma)) - \tanh(\beta S_j(\omega))|, \tag{4.37}
\]

where (a) corresponds to \( i_* = i \) in which case \( w_0(X^*, Y^*) = 0 \) and (b) corresponds to \( i_* \sim i \) in which case \( w_0(X^*, Y^*) = 2 \) with probability \( \frac{1}{2} |\tanh(\beta S_i(\sigma)) - \tanh(\beta S_i(\omega))| \) by our coupling, and \( w_0(X^*, Y^*) = 1 \) otherwise. To bound (b), we note that for \( j \sim i \)

\[
|\tanh(\beta S_j(\sigma)) - \tanh(\beta S_j(\omega))| = \tanh(\beta(s + 2)) - \tanh(\beta s), \tag{4.38}
\]

where

\[
s := S_j(\sigma) \land S_j(\omega).
\]

The derivative of \( \tanh \) is maximized at 0 where it is equal to 1. So the r.h.s. of (4.38) is \( \leq 2\beta \). Plugging this back into (4.37), we get

\[
E[w_0(X^*, Y^*)] \leq 1 - \frac{1}{n} + \frac{\delta \beta}{n} \leq \exp \left( - \frac{1 - \delta \beta}{n} \right) = \kappa w_0(\sigma, \omega),
\]

where

\[
\kappa := \exp \left( - \frac{1 - \delta \beta}{n} \right) < 1,
\]

by assumption. The diameter of \( H_0 \) is \( \Delta_0 = n \). By Theorem 4.73,

\[
t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{\log \Delta_0 + \log \varepsilon^{-1}}{\log \kappa^{-1}} \right\rceil = \left\lceil \frac{n(\log n + \log \varepsilon^{-1})}{1 - \delta \beta} \right\rceil,
\]

which implies the claim.

\[\blacksquare\]

\textbf{Remark 4.78.} A slightly more careful analysis shows that the condition \( \bar{\delta} \tanh(\beta) < 1 \) is enough for the claim to hold. See [LPW06, Theorem 15.1].

\subsection*{4.4.5 From approximate sampling to approximate counting}

To be written. See [MU05, Section 10.3] and [LPW06, Sections 14.3 and 14.4]. See also [JS97].

245
4.5 Chen-Stein method

The Chen-Stein method serves to establish Poisson approximation results with quantitative bounds in certain settings with dependent variables that are common, for instance, in random graphs and string statistics. The basic setup is a sum of \{0, 1\}-valued random variables \( \{X_i\}_{i=1}^n \)

\[
W = \sum_{i=1}^{n} X_i \tag{4.39}
\]

where the \(X_i\)'s are not assumed independent or identically distributed. Define

\[
p_i = \Pr[X_i = 1] \tag{4.40}
\]

and

\[
\mathbb{E}[W] = \lambda := \sum_{i=1}^{n} p_i. \tag{4.41}
\]

Letting \(\mu\) denote the law of \(W\) and \(\pi\) be the Poisson distribution with mean \(\lambda\), our goal is to bound \(\|\mu - \pi\|_{TV}\).

We first state the main bounds and give some examples of its use. We then motivate and prove the result, and return to further applications.

4.5.1 Main bounds and examples

We begin with a definition.

**Definition 4.79 (Stein coupling).** Using the notation in (4.39), (4.40) and (4.41), a **Stein coupling** is a pair \((U_i, V_i)\), for each \(i = 1, \ldots, n\), such that

\(U_i \sim W\), \hspace{1cm} \(V_i \sim W - 1|X_i = 1\).

Each pair \((U_i, V_i)\) is defined on a joint probability space, but different pairs are not.

The key bound is the following. It is proved in Section 4.5.2.

**Theorem 4.80 (Chen-Stein method).** Using the notation in (4.39), (4.40) and (4.41), let \((U_i, V_i)\), \(i = 1, \ldots, n\), be a Stein coupling. Then

\[
\|\mu - \pi\|_{TV} \leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E}|U_i - V_i|, \tag{4.42}
\]

where, as before, \(\mu\) denotes the law of \(W\) and \(\pi\) is the Poisson distribution with mean \(\lambda\).
As a first example, we derive a Poisson approximation result in the independent case. Compare to Theorem 4.16.

**Example 4.81** (Independent $X_i$’s). Assume the $X_i$’s are independent. We prove the following:

**Claim 4.82.**

$$\|\mu - \pi\|_{TV} \leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i^2.$$

We use the following Stein coupling. For each $i = 1, \ldots, n$, we let

$$U_i = W$$

and

$$V_i = \sum_{j \neq i} X_j.$$

By independence,

$$V_i = W - X_i \sim W - 1|X_i = 1,$$

as desired. Plugging into (4.42), we obtain the bound

$$\|\mu - \pi\|_{TV} \leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E} |U_i - V_i|$$

$$\leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E} |W - \sum_{j \neq i} X_j|$$

$$\leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E} |X_i|$$

$$\leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i^2.$$

Here is a less straightforward example.

**Example 4.83** (Balls in boxes). Suppose we throw $k$ balls uniformly at random in $n$ boxes independently. Let

$$X_i = 1\{\text{box } i \text{ is empty}\},$$
and let $W = \sum_{i=1}^{n} X_i$ be the number of empty boxes. Note that the $X_i$'s are not independent. (Think of what happens with one ball.) Note that

$$p_i = \left(1 - \frac{1}{n}\right)^k,$$

for all $i$ and, hence,

$$\lambda = n \left(1 - \frac{1}{n}\right)^k.$$

For each $i = 1, \ldots, n$, we generate the $(U_i, V_i)$ in the following way. We let $U_i = W$. If box $i$ is empty then $V_i = W - 1$. Otherwise, we re-distribute all balls in box $i$ among the remaining boxes and let $V_i$ count the number of empty boxes $\neq i$. By construction, both conditions of the Stein coupling are satisfied. Moreover we have almost surely $V_i \leq U_i$ so that

$$\mathbb{E}[U_i - V_i] = \mathbb{E}[U_i - V_i] = \lambda - \sum_{i=1}^{n} p_i \mathbb{E}[V_i].$$

By the fact that $V_i \sim U_i - 1 | X_i = 1$ and Bayes’ rule,

$$\sum_{i=1}^{n} p_i \mathbb{E}[V_i] = \sum_{i=1}^{n} \mathbb{P}[X_i = 1] \sum_{k=1}^{n} (k - 1) \mathbb{P}[V_i = k - 1]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{n} (k - 1) \mathbb{P}[U_i = k | X_i = 1] \mathbb{P}[X_i = 1]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{n} (k - 1) \mathbb{P}[X_i = 1 | U_i = k] \mathbb{P}[U_i = k]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{n} (k - 1) \mathbb{E}[X_i | U_i = k] \mathbb{P}[U_i = k]$$

$$= \sum_{k=1}^{n} (k - 1) k \mathbb{P}[U_i = k]$$

$$= \mathbb{E}[U_i^2] - \mathbb{E}[U_i].$$

It remains to compute $\mathbb{E}[U_i^2] = \mathbb{E}[W^2]$. We have by symmetry

$$\mathbb{E}[W^2] = n \mathbb{E}[X_i^2] + n(n - 1) \mathbb{E}[X_1 X_2]$$

$$= \lambda + n(n - 1) \left(1 - \frac{2}{n}\right)^k,$$
so
\[ \|\mu - \pi\|_{TV} \leq (1 \wedge \lambda^{-1}) \left\{ \lambda^2 - n(n - 1) \left( 1 - \frac{2}{n} \right)^k \right\}. \]

This last example is generalized in Exercise 4.15.

In special settings, one can give useful general bounds by providing an appropriate Stein coupling. We give an important example next. Recall that \([n] = \{1, \ldots, n\}\).

**Theorem 4.84** (Chen-Stein: Dissociated Case). *Using the notation in (4.39), (4.40) and (4.41), suppose that for each \(i\) there is a neighborhood \(\mathcal{N}_i \subseteq [n] \setminus \{i\}\) such that \(X_i \) is independent of \(\{X_j : j \notin \mathcal{N}_i \cup \{i\}\}\).*

Then
\[ \|\mu - \pi\|_{TV} \leq (1 \wedge \lambda^{-1}) \sum_{i=1}^{n} \left\{ p_i^2 + \sum_{j \in \mathcal{N}_i} (p_i p_j + E[X_i X_j]) \right\}. \]

**Proof:** We use the following Stein coupling. Let
\[ U_i = W. \]

Then generate
\[ (Y_j^{(i)})_{j \in \mathcal{N}_i} \sim (X_j)_{j \in \mathcal{N}_i} \mid \{X_k : k \notin \mathcal{N}_i \cup \{i\}\}, X_i = 1, \]
and set
\[ V_i = \sum_{k \notin \mathcal{N}_i \cup \{i\}} X_k + \sum_{j \in \mathcal{N}_i} Y_j^{(i)}. \]

Because the law of \(\{X_k : k \notin \mathcal{N}_i \cup \{i\}\}\) is independent of the event \(\{X_i = 1\}\), the above scheme satisfies the conditions of the Stein coupling.
Therefore we can apply Theorem 4.80

$$\|\mu - \pi\|_{TV} \leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E}[U_i - V_i]$$

$$= (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E}\left[\sum_{j=1}^{n} X_j - \sum_{k \notin N_i \cup \{i\}} X_k - \sum_{j \in N_i} Y_j^{(i)}\right]$$

$$= (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E}\left[X_i + \sum_{j \in N_i} (X_j - Y_j^{(i)})\right]$$

$$\leq (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \left(\mathbb{E}|X_i| + \sum_{j \in N_i} \mathbb{E}|X_j| + \mathbb{E}|Y_j^{(i)}|\right)$$

$$= (1 \land \lambda^{-1}) \sum_{i=1}^{n} p_i \left(p_i + \sum_{j \in N_i} [p_j + \mathbb{E}|X_j||X_i = 1]\right)$$

$$= (1 \land \lambda^{-1}) \sum_{i=1}^{n} \left\{p_i^2 + \sum_{j \in N_i} (p_i p_j + p_i \mathbb{E}[X_j|X_i = 1])\right\}$$

$$= (1 \land \lambda^{-1}) \sum_{i=1}^{n} \left\{p_i^2 + \sum_{j \in N_i} (p_i p_j + \mathbb{E}[X_i X_j])\right\}.$$ 

That concludes the proof. 

We give an example.

**Example 4.85** (Longest head run). Let $0 < q < 1$ and let $Z_1, Z_2, \ldots$ be i.i.d. Bernoulli random variables with $q = \mathbb{P}[Z_i = 1]$. We are interested in the distribution of $R$, the length of the longest run of 1’s starting in the first $n$ tosses. For a positive integer $t$, define

$$X_1^{(t)} = Z_1 \cdots Z_t,$$

$$X_i^{(t)} = (1 - Z_{i-1})Z_i \cdots Z_{i+t-1},$$

and

$$W^{(t)} = \sum_{i=1}^{n} X_i^{(t)}.$$
The event \{X_i^{(t)} = 1\} indicates that a head run of length at least \( t \) starts at the \( i \)-th toss. The key observation is that

\[
\{ R < t \} = \{ W^{(t)} = 0 \}. \tag{4.43}
\]

Notice that, for fixed \( t \), the \( X_i^{(t)} \)'s are neither independent nor identically distributed. However, they exhibit a natural neighborhood structure as in Theorem 4.84. Indeed let

\[
\mathcal{N}_i^{(t)} = \{ \alpha \in [n] : |\alpha - i| \leq t \} \setminus \{ i \}.
\]

Then, \( X_i^{(t)} \) is independent of \( \{ X_j^{(t)} : j \notin \mathcal{N}_i \cup \{ i \} \} \). For example,

\[
X_i^{(t)} = (1 - Z_{i-1})Z_i \cdots Z_{i+t-1},
\]

and

\[
X_{i+t}^{(t)} = (1 - Z_{i+t})Z_{i+t+1} \cdots Z_{i+2t},
\]

do not depend on any common \( Z_j \), while \( X_i^{(t)} \) and

\[
X_{i+t}^{(t)} = (1 - Z_{i+t-1})Z_{i+t} \cdots Z_{i+2t-1},
\]

both depend on \( Z_{i+t-1} \).

We compute the quantities needed to apply Theorem 4.84. We have

\[
p_1^{(t)} = \mathbb{E}[Z_1 \cdots Z_t] = \prod_{j=1}^{t} \mathbb{E}[Z_j] = q^t,
\]

and, for \( i \geq 2 \),

\[
p_i^{(t)} = \mathbb{E}[(1 - Z_{i-1})Z_i \cdots Z_{i+t-1}]
= \mathbb{E}[1 - Z_{i-1}] \prod_{j=i}^{i+t-1} \mathbb{E}[Z_j]
= (1 - q)q^t
\leq q^t.
\]

For \( i \geq 1 \) and \( j \in \mathcal{N}_i^{(t)} \), observe that a head run of length at least \( t \) cannot start simultaneously at \( i \) and \( j \), where by definition of the neighborhoods \( j \) is within \( t \) of \( i \). Hence, \( \mathbb{E}[X_i^{(t)} X_j^{(t)}] = 0 \) in that case. We also have

\[
\lambda^{(t)} = \mathbb{E}[W^{(t)}] = q^t + (n - 1)(1 - q)q^t \in [n(1 - q)q^t, nq^t],
\]

251
and \[ |X_i^{(t)}| \leq 2t. \]

We are ready to apply Theorem 4.84. We get
\[
\|\mu - \pi\|_{TV} \leq (1 \wedge (\lambda(t))^{-1}) \sum_{i=1}^{n} \left\{ \left( p_i^{(t)} \right)^2 + \sum_{j \in N_i} \left( p_i^{(t)} p_j^{(t)} + \mathbb{E}[X_i^{(t)} X_j^{(t)}] \right) \right\}
\leq (1 \wedge (n(1 - q)q^t)^{-1}) \left[ nq^{2t} + 2tnq^{2t} \right]
\leq \frac{1}{(1 - q)n} (1 \wedge (nq^t)^{-1})(2t + 1)(nq^t)^2.
\]

This bound is non-asymptotic—it holds for any \( q, n, t \). One special regime of note is \( t = \log_{1/q} n + C \) with large \( n \). In that case, we have \( nq^t \to C' \) as \( n \to +\infty \) for some \( 0 < C' < +\infty \) and the total variation above is of the order to \( n^{-1} \log n \).

Going back to (4.43), we finally obtain when \( t = \log_{1/q} n + C \) that
\[
\left| P[R_n < t] - e^{-\lambda(t)} \right| = O \left( \frac{\log n}{n} \right).
\]

4.5.2 Motivation and proof

The idea behind the Chen-Stein method is to interpolate between \( \mu \) and \( \pi \) in Theorem 4.80 by constructing a Markov chain with initial distribution \( \mu \) and stationary distribution \( \pi \). There are many ways to define such a chain. Here we use a discrete-time, finite Markov chain for simplicity.

Proof of Theorem 4.80. We seek a bound on
\[
\|\mu - \pi\|_{TV} = \sup_{A \subseteq \mathbb{Z}_+} |\mu(A) - \pi(A)|
= \mu(A^*) - \pi(A^*)
= \sum_{z \in A^*} (\mu(z) - \pi(z)),
\]
where \( A^* = \{ z \in \mathbb{Z}_+ : \mu(z) > \pi(z) \} \), by Lemma 4.13. Since \( W \leq n \) almost surely, \( \mu(z) = 0 \) for all \( z > n \) which implies that \( A^* \subseteq \{0, 1, \ldots, n\} \). In particular, it will suffice to bound \( \mu(z) - \pi(z) \) for \( 0 \leq z \leq n \).
Constructing the Markov chain

For this purpose, it will be convenient to truncate \( \pi \) at \( n \), that is, we define

\[
\bar{\pi}(z) = \begin{cases} 
\pi(z), & 0 \leq z \leq n, \\
1 - \Pi(n), & z = n + 1, \\
0, & \text{otherwise},
\end{cases}
\]

where \( \Pi(z) = \sum_{w \leq z} \pi(w) \) is the cumulative distribution function of the Poisson distribution with mean \( \lambda \). We construct a Markov chain with stationary distribution \( \bar{\pi} \). We will also need the chain to be aperiodic and irreducible over \( \{0, 1, \ldots, n + 1\} \). We choose the transition matrix \( (P(x,y)) \) to be that of a birth-death chain reversible with respect to \( \bar{\pi} \), that is, we require

\[
P(x,y) = 0 \quad \text{unless} \quad |x - y| \leq 1
\]

and

\[
P(x,x+1) P(x+1,x) = \frac{\bar{\pi}(x+1)}{\bar{\pi}(x)}, \quad \forall x \in [n].
\]

For \( x < n \),

\[
\frac{\bar{\pi}(x+1)}{\bar{\pi}(x)} = \frac{\pi(x+1)}{\pi(x)} = \frac{e^{-\lambda} \lambda^{x+1}/(x+1)!}{e^{-\lambda} \lambda^x/x!} = \frac{\lambda}{x+1}.
\]

In view of this, we let the nonzero transition probabilities take values

\[
P(x,y) = \begin{cases} 
C\lambda, & \text{if } y = x + 1, 0 \leq x \leq n, \\
Cx, & \text{if } y = x - 1, 1 \leq x \leq n, \\
1 - Cx - C\lambda, & \text{if } y = x, 1 \leq x \leq n, \\
1 - C\lambda, & \text{if } y = x = 0, \\
C\lambda \frac{\pi(n)}{1 - \Pi(n)}, & \text{if } y = x - 1, x = n + 1, \\
1 - C\lambda \frac{\pi(n)}{1 - \Pi(n)}, & \text{if } y = x = n + 1,
\end{cases}
\]

for a constant \( C \in (0, +\infty) \) small enough that all probabilities above are strictly positive. Then, \( P \) is aperiodic, irreducible and satisfies the detailed balance conditions (4.45).

Interpolating between \( \mu \) and \( \pi \)

By the convergence theorem for Markov chains, Theorem 1.22,

\[
P^t(y, z) \to \pi(z)
\]

for all \( 0 \leq y \leq n + 1 \) and \( 0 \leq z \leq n \) as \( t \to +\infty \). Hence, letting \( \delta_z(x) = 1_{\{x = z\}} \),

\[
\delta_z(y) - \pi(z) = \lim_{t \to +\infty} \mathbb{E}_y[\delta_z(X_0) - \delta_z(X_t)]
\]

\[
= \lim_{t \to +\infty} \sum_{s=0}^{t-1} \mathbb{E}_y[\delta_z(X_s) - \delta_z(X_{s+1})], \quad (4.47)
\]
and, taking expectation over \( \mu \),

\[
\mu(z) - \pi(z) = \lim_{t \to +\infty} E_{\mu}[\delta_z(X_0) - \delta_z(X_t)] \\
= \lim_{t \to +\infty} \sum_{s=0}^{t-1} E_{\mu}[\delta_z(X_s) - \delta_z(X_{s+1})],
\]

(4.48)

where the subscript of \( E \) indicates the initial distribution or state, and we used the fact that \( \pi(z) = \bar{\pi}(z) \) on \( \{0, 1 \ldots, n\} \).

**Markov chain calculations** We use standard Markov chains facts to re-write the expression above in a more useful form. By Chapman-Kolmogorov, Theorem 1.13, applied to the first step of the chain,

\[
E_y[\delta_z(X_{s+1})] \\
= P(y, y + 1)E_{y+1}[\delta_z(X_s)] + P(y, y)E_y[\delta_z(X_s)] + P(y, y - 1)E_{y-1}[\delta_z(X_s)].
\]

Using that \( P(y, y + 1) + P(y, y) + P(y, y - 1) = 1 \) and rearranging we get for \( 0 \leq y \leq n \) and \( 0 \leq z \leq n \)

\[
\sum_{s=0}^{t-1} E_y[\delta_z(X_s) - \delta_z(X_{s+1})] \\
= \sum_{s=0}^{t-1} E_y[\delta_z(X_s) - \delta_z(X_{s+1})] \\
= \sum_{s=0}^{t-1} \left\{ -P(y, y + 1)(E_{y+1}[\delta_z(X_s)] - E_y[\delta_z(X_s)]) \\
+ P(y, y - 1)(E_y[\delta_z(X_s)] - E_{y-1}[\delta_z(X_s)]) \right\} \\
= -P(y, y + 1)C^{-1}g_z^t(y + 1) + P(y, y - 1)C^{-1}g_z^t(y),
\]

(4.49)

where we defined

\[
g_z^t(y) = C \sum_{s=0}^{t-1} (E_y[\delta_z(X_s)] - E_{y-1}[\delta_z(X_s)]).
\]

(4.50)

The function \( g_z^t(y) \) is the difference between the expected number of visits to \( z \) when started at \( y \) and \( y - 1 \) respectively. We establish after the proof of the theorem that it has a well-defined limit. That fact is not immediately obvious as the limit of \( g_z^t(y) \) is difference of two infinities.
Lemma 4.86. Let \( g_z : \{0, 1, \ldots, n+1\} \to \mathbb{R} \) be defined in (4.50). Then there exists a bounded function \( g_z^\infty : \{0, 1, \ldots, n+1\} \to \infty \) such that for all \( 0 \leq z \leq n \) and \( 0 \leq y \leq n+1 \),
\[
g_z^\infty(y) = \lim_{t \to +\infty} g_z^t(y).
\]

Chen’s equation and a recursion

For \( A \subseteq \{0, 1, \ldots, n\} \), let
\[
g_A^\infty(y) = \sum_{z \in A} g_z^\infty(y).
\]

We obtain the following key bound.

Lemma 4.87 (Chen’s equation). Let \( W \sim \mu \) and \( \pi \sim \text{Poi}(\lambda) \). Then,
\[
\|\mu - \pi\|_{TV} = \mathbb{E}[ -\lambda g_A^\infty(W + 1) + W g_A^\infty(W)] \tag{4.51}
\]
where \( A^* = \{z \in \mathbb{Z}_+: \mu(z) > \pi(z)\} \).

Proof. Combine (4.44), (4.46), (4.48), (4.49) and Lemma 4.86.

The expectation on the right-hand side of (4.51) can actually be used to characterize the Poisson distribution and is the starting point of a simpler proof of the Chen-Stein method. See Exercise 4.12.

In fact, an explicit expression for \( g_z^\infty \) can be derived via the following recursion. That expression in turn will be helpful to bound the right-hand side of (4.51). For notational convenience, define
\[
g_z^\infty(0) = g_z^\infty(n+2) = 0.
\]
for all \( 0 \leq z \leq n \).

Lemma 4.88. For all \( 0 \leq y \leq n \) and \( 0 \leq z \leq n \),
\[
\delta_z(y) - \pi(z) = -\lambda g_z^\infty(y+1) + yg_z^\infty(y).
\]

Proof. Combine (4.47), (4.46), (4.49) and Lemma 4.86.

Lemma 4.88 leads to the following formula for \( g_z^\infty \), which we establish after the proof of the theorem.

Lemma 4.89. For \( 0 \leq y \leq n+1 \) and \( 0 \leq z \leq n \),
\[
g_z^\infty(y) = \begin{cases} \frac{\Pi(y-1)}{y\Pi(y)} \pi(z), & \text{if } z \geq y, \\
\frac{1-\Pi(y-1)}{y\Pi(y)} \pi(z), & \text{if } z < y. \end{cases} \tag{4.52}
\]
Lemma 4.89 in turn can be used to derive a Lipschitz constant for $g_A^\infty$. That lemma is also established after the proof of the theorem.

**Lemma 4.90.** For $A \subseteq \{0, 1, \ldots, n\}$ and $y, y' \in \{0, 1, \ldots, n + 1\}$,

$$|g_A^\infty(y') - g_A^\infty(y)| \leq (1 \wedge \lambda^{-1})|y' - y|.$$

**Using the Stein coupling** We now use Lemma 4.87 and Definition 4.79 to derive the final bound in the claim. By (4.51), using the facts that $\lambda = \sum_{i=1}^{n} p_i$ and $W = \sum_{i=1}^{n} X_i$, we get

$$\|\mu - \pi\|_{TV} = \mathbb{E}[-\lambda g_A^\infty(W + 1) + W g_A^\infty(W)]$$

$$= \mathbb{E} \left[ - \left( \sum_{i=1}^{n} p_i \right) g_A^\infty(W + 1) + \left( \sum_{i=1}^{n} X_i \right) g_A^\infty(W) \right]$$

$$= \sum_{i=1}^{n} (-p_i \mathbb{E} [g_A^\infty(W + 1)] + \mathbb{E} [ X_i g_A^\infty(W)])

= \sum_{i=1}^{n} (-p_i \mathbb{E} [g_A^\infty(W + 1)] + \mathbb{E} [ g_A^\infty(W) | X_i = 1] \ P[X_i = 1])$$

$$= \sum_{i=1}^{n} p_i (-\mathbb{E} [g_A^\infty(W + 1)] + \mathbb{E} [ g_A^\infty(W) | X_i = 1]). \quad (4.54)$$

Let $(U_i, V_i), i = 1, \ldots, n$, be a Stein coupling. Then, we can re-write this last expression as

$$\sum_{i=1}^{n} p_i (-\mathbb{E} [g_A^\infty(U_i + 1)] + \mathbb{E} [ g_A^\infty(V_i + 1)])$$

$$= \sum_{i=1}^{n} p_i (-\mathbb{E} [g_A^\infty(U_i + 1)] + \mathbb{E} [ g_A^\infty(V_i + 1)]). \quad (4.55)$$

By Lemma 4.90, combining (4.54) and (4.55) gives

$$\|\mu - \pi\|_{TV} \leq (1 \wedge \lambda^{-1}) \sum_{i=1}^{n} p_i \mathbb{E} |U_i - V_i|,$$

which concludes the proof.
**Additional proofs**  It remains to prove Lemmas 4.86, 4.89 and 4.90.

*Proof of Lemma 4.86.* We use a coupling argument. Let \((Y_s, \tilde{Y}_s)_{s=0}^{+\infty}\) be an independent coupling of \((Y_s)\), the chain started at \(y - 1\), and \((\tilde{Y}_s)\), the chain started at \(y\). Let \(\tau\) be the first time \(s\) that \(Y_s = \tilde{Y}_s\). Because \(Y_s\) and \(\tilde{Y}_s\) are independent and \(P\) is a birth-death chain with strictly positive transition probabilities, the coupled chain \((Y_s, \tilde{Y}_s)_{s=0}^{+\infty}\) is aperiodic and irreducible over \(\{0, 1, \ldots, n+1\}^2\). By the exponential tail of hitting times, Lemma 3.26, it holds that \(E[\tau] < +\infty\).

Modify the coupling \((Y_s, \tilde{Y}_s)\) to enforce \(\tilde{Y}_s = Y_s\) for all \(s \geq \tau\) (while not changing \((Y_s)\)). By the Strong Markov property, Theorem 3.11, the resulting chain \((Y^*_s, \tilde{Y}^*_s)\) is also a coupling of the chain started at \(y - 1\) and \(y\) respectively. Using this coupling, we re-write

\[
g^t(z) = C \sum_{s=0}^{t-1} \left( E_y[\delta_z(X_s)] - E_{y-1}[\delta_z(X_s)] \right) 
\]

\[
= C \sum_{s=0}^{t-1} E[\delta_z(\tilde{Y}^*_s) - \delta_z(Y^*_s)] 
\]

\[
= CE \left[ \sum_{s=0}^{t-1} (\delta_z(\tilde{Y}^*_s) - \delta_z(Y^*_s)) \right] . 
\]

The random variable inside the expectation is bounded in absolute value by

\[
\left| \sum_{s=0}^{t-1} (\delta_z(\tilde{Y}^*_s) - \delta_z(Y^*_s)) \right| \leq \tau, 
\]

uniformly in \(t\). Indeed, after \(s = \tau\), the terms in the sum are 0, while before \(s = \tau\) the terms are bounded by 1 in absolute value. By the integrability of \(\tau\), the dominated convergence theorem allows to take the limit, leading to

\[
g^\infty(z) = \lim_{t \to +\infty} C E \left[ \sum_{s=0}^{+\infty} (\delta_z(\tilde{Y}^*_s) - \delta_z(Y^*_s)) \right] 
\]

\[
= C E \left[ \sum_{s=0}^{+\infty} (\delta_z(\tilde{Y}^*_s) - \delta_z(Y^*_s)) \right] < +\infty . 
\]

That concludes the proof.  

257
Proof of Lemma 4.89. Our starting point is Lemma 4.88, from which we deduce the recursive formula

\[ g_z^\infty(y + 1) = \frac{1}{\lambda} \{ yg_z^\infty(y) + \pi(z) - \delta_z(y) \}, \quad (4.56) \]

for \( 0 \leq y \leq n \) and \( 0 \leq z \leq n \). Hence,

\[ g_z^\infty(1) = \frac{1}{\lambda} \{ \pi(z) - \delta_z(0) \}, \quad (4.57) \]

\[ g_z^\infty(2) = \frac{1}{\lambda} \{ g_z^\infty(1) + \pi(z) - \delta_z(1) \} \]
\[ = \frac{1}{\lambda} \left\{ \frac{1}{\lambda} \{ \pi(z) - \delta_z(0) \} + \pi(z) - \delta_z(1) \right\} \]
\[ = \frac{1}{\lambda^2} \{ \pi(z) - \delta_z(0) \} + \frac{1}{\lambda} \{ \pi(z) - \delta_z(1) \}, \]

\[ g_z^\infty(3) = \frac{1}{\lambda} \{ 2g_z^\infty(2) + \pi(z) - \delta_z(2) \} \]
\[ = \frac{1}{\lambda} \left\{ 2 \frac{1}{\lambda^2} \{ \pi(z) - \delta_z(0) \} + 2 \frac{1}{\lambda} \{ \pi(z) - \delta_z(1) \} + \pi(z) - \delta_z(2) \right\} \]
\[ = \frac{2}{\lambda^3} \{ \pi(z) - \delta_z(0) \} + \frac{2}{\lambda^2} \{ \pi(z) - \delta_z(1) \} + \frac{1}{\lambda} \{ \pi(z) - \delta_z(2) \}, \]

and so forth. We posit the general formula

\[ g_z^\infty(y) = \frac{(y - 1)!}{\lambda^y} \sum_{k=0}^{y-1} \frac{\lambda^k}{k!} \{ \pi(z) - \delta_z(k) \}, \quad (4.58) \]

for \( 1 \leq y \leq n + 1 \) and \( 0 \leq z \leq n \). The formula is straightforward to confirm by induction. Indeed it holds for \( y = 1 \) as can be seen in (4.57) (and recalling that \( 0! = 1 \) by convention) and, assuming it holds for \( y \), we have by (4.56)

\[ g_z^\infty(y + 1) = \frac{1}{\lambda} \{ yg_z^\infty(y) + \pi(z) - \delta_z(y) \} \]
\[ = \frac{1}{\lambda} \left\{ y \frac{(y - 1)!}{\lambda^y} \sum_{k=0}^{y-1} \frac{\lambda^k}{k!} \{ \pi(z) - \delta_z(k) \} + \pi(z) - \delta_z(y) \right\} \]
\[ = \frac{y!}{\lambda^{y+1}} \sum_{k=0}^{y-1} \frac{\lambda^k}{k!} \{ \pi(z) - \delta_z(k) \} + \frac{1}{\lambda} \{ \pi(z) - \delta_z(y) \} \]
\[ = \frac{y!}{\lambda^{y+1}} \sum_{k=0}^{y} \frac{\lambda^k}{k!} \{ \pi(z) - \delta_z(k) \}, \]
as desired.

We re-write (4.58) depending on whether the term $\delta_z(y) = 1\{z = y\}$ plays a role in the equation. For $z \geq y$, the equation simplifies to

$$g^\infty_z(y) = \frac{(y - 1)!}{\lambda^y} \sum_{k=0}^{y-1} \frac{\lambda^k}{k!} \pi(z)$$

$$= \frac{1}{y} y! e^{-\lambda \lambda y} \sum_{k=0}^{y-1} \frac{e^{-\lambda} \lambda^k}{k!} \pi(z)$$

$$= \frac{\Pi(y - 1)}{y \pi(y)} \pi(z).$$

For $z < y$, we get instead

$$g^\infty_z(y) = \frac{(y - 1)!}{\lambda^y} \left\{ \frac{\sum_{k=0}^{y-1} \lambda^k}{k!} \pi(z) - \frac{\lambda^y}{y!} \right\}$$

$$= \frac{1}{y} y! e^{-\lambda \lambda y} \left\{ \sum_{k=0}^{y-1} \frac{e^{-\lambda} \lambda^k}{k!} \pi(z) - \pi(z) \right\}$$

$$= \frac{\Pi(y - 1)}{y \pi(y)} - 1 \pi(z).$$

That concludes the proof.

**Proof of Lemma 4.90.** It suffices to prove that, for $A \subseteq \{0, 1, \ldots, n\}$ and $y \in \{0, 1, \ldots, n\}$,

$$|g^\infty_A(y + 1) - g^\infty_A(y)| \leq (1 \wedge \lambda^{-1}),$$

and then use the triangle inequality.

We use the expression derived in Lemma 4.89. For $y < z$,

$$g^\infty_z(y + 1) - g^\infty_z(y) = \frac{\Pi(y)}{(y + 1) \pi(y + 1)} \pi(z) - \frac{\Pi(y - 1)}{y \pi(y)} \pi(z)$$

$$= \pi(z) \frac{1}{y \pi(y)} \left\{ \frac{y}{\lambda} \Pi(y) - \Pi(y - 1) \right\},$$

where we used that $\pi(y + 1)/\pi(y) = \lambda/(y + 1)$. We show that the expression in
curly brackets is non-negative. Indeed,

\[
\begin{align*}
\frac{y}{\lambda} \sum_{k'=0}^{y} e^{-\lambda k'} \left( \frac{y}{\lambda} \right)^{k'} - \sum_{k=0}^{y-1} e^{-\lambda k} \frac{k}{k!} & = \frac{y}{\lambda} e^{-\lambda} e^{\frac{y}{\lambda}} + \sum_{k=0}^{y-1} e^{-\lambda (k+1)} \frac{1}{y} \frac{k}{k!} - \sum_{k=0}^{y-1} e^{-\lambda} \frac{k}{k!} \\
& \geq \frac{y}{\lambda} e^{-\lambda} + \sum_{k=0}^{y-1} e^{-\lambda} \frac{k}{k!} - \sum_{k=0}^{y-1} e^{-\lambda} \frac{k}{k!} \\
& \geq 0.
\end{align*}
\]

So \( g_z^\infty (y + 1) - g_z^\infty (y) \geq 0 \) for \( 0 \leq y < z \). A similar calculation, which we omit, shows that the same inequality holds for \( z < y \leq n \). For \( y = n + 1 \), we get

\[
g_z^\infty (n + 2) - g_z^\infty (n + 1) = 0 + \frac{1 - \Pi(n)}{(n+1)\pi(n+1)} \pi(z) \geq 0.
\]

Moreover,

\[
0 = g_z^\infty (n + 2) - g_z^\infty (0) = \sum_{y=1}^{n+1} \{ g_z^\infty (y + 1) - g_z^\infty (y) \}.
\]

We have shown that all terms in the last sum are non-negative, with the exception of the term \( y = z \). Hence, for \( 0 \leq y \leq n \), it must be that the maximum of \( |g_z^\infty (y + 1) - g_z^\infty (y)| \) is achieved at \( z = y \). In fact, for \( 0 \leq y \leq n \), it must be that the maximum of \( |g_A^\infty (y + 1) - g_A^\infty (y)| \) over \( A \subseteq \{0,1,\ldots,n\} \) is achieved at \( A = \{y\} \). It remains to bound that last case.
We have

\[
|g^\infty_y(y + 1) - g^\infty_y(y)| = \left| -\frac{1 - \Pi(y)}{(y + 1)\pi(y + 1)} \pi(y) - \frac{\Pi(y - 1)}{y\pi(y)} \pi(y) \right|
\]

\[
= \frac{1}{\lambda} \sum_{k \geq y + 1} e^{-\lambda} \frac{\lambda^k}{k!} + \frac{1}{y} \sum_{k = 0}^{y - 1} e^{-\lambda} \frac{\lambda^k}{k!}
\]

\[
= e^{-\lambda} \left\{ \sum_{k = 1}^{y} \frac{\lambda^k}{k'!} \frac{y}{y} + \sum_{k \geq y + 1} \frac{\lambda^k}{k!} + \sum_{k \geq y + 1} \frac{\lambda^k}{k!} \right\}
\]

\[
\leq e^{-\lambda} \left\{ \sum_{k \geq 1} \frac{\lambda^k}{k!} \right\}
\]

\[
= e^{-\lambda} \left\{ e^{\lambda} - 1 \right\}
\]

\[
= 1 - e^{-\lambda}.
\]

For \( \lambda \geq 1 \), we have \( \frac{1 - e^{-\lambda}}{\lambda} \leq \frac{1}{\lambda} = (1 \wedge \lambda^{-1}) \), while for \( 0 < \lambda < 1 \) we have \( \frac{1 - e^{-\lambda}}{\lambda} \leq \frac{1}{\lambda} = (1 \wedge \lambda^{-1}) \). 

\[ \blacksquare \]

4.5.3 Erdős-Rényi: subgraph containment at the threshold

We revisit the subgraph containment problem of Section 2.3.2 (and Section 4.3.5). Let \( G_n \sim \mathcal{G}_{n,p_n} \) be an Erdős-Rényi graph with \( n \) vertices and density \( p_n \). Let \( \omega(G) \) be the clique number of a graph \( G \), i.e., the size of its largest clique. We showed previously that the property \( \omega(G) \geq 4 \) has threshold function \( n^{2/3} \). Here we consider what happens when \( p_n = C n^{-2/3} \). We use the Chen-Stein method.

For an enumeration \( S_1, \ldots, S_m \) of the 4-tuples of vertices in \( G_n \), let \( A_1, \ldots, A_m \) be the events that the corresponding 4-cliques are present and define \( Z_i = 1_{A_i} \). Then \( W = \sum_{i=1}^{m} Z_i \) is the number of 4-cliques in \( G_n \). We argued previously that

\[
q_i := \mathbb{E}[Z_i] = p_n^6,
\]

and

\[
\lambda := \mathbb{E}[W] = \binom{n}{4} p_n^6.
\]

In our regime of interest, \( \lambda \) is of constant order.
Observe that the $Z_i$'s are not independent because the 4-tuples may share potential edges. However they admit a neighborhood structure as in Theorem 4.84. Specifically, for $i = 1, \ldots, m$, define

$$N_i = \{ j : S_i \text{ and } S_j \text{ share at least two vertices} \} \setminus \{ i \}.$$ 

Then the conditions of Theorem 4.84 are satisfied, that is, $X_i$ is independent of $\{ Z_j : j \notin N_i \cup \{ i \} \}$. We argued previously that

$$|N_i| = \left( \begin{array}{c} 4 \\ 3 \end{array} \right) (n - 4) + \left( \begin{array}{c} 4 \\ 2 \end{array} \right) \frac{(n - 4)}{2} = \Theta(n^2),$$

where the first term counts the number of $S_j$'s sharing exactly three vertices with $S_i$, in which case $E[Z_i Z_j] = p_{12}^2 n$, and the second term counts those sharing two, in which case $E[Z_i Z_j] = p_{11}^2 n$.

We are ready to apply the bound in Theorem 4.84. Using the formulas above, we get

$$\|\mu - \pi\|_{TV} \leq (1 \wedge \lambda^{-1}) \sum_{i=1}^{n} \left\{ \frac{n}{4} q_i^2 + \sum_{j \in N_i} (q_i q_j + E[Z_i Z_j]) \right\}$$

$$\leq (1 \wedge \lambda^{-1}) \left( \begin{array}{c} n \\ 4 \end{array} \right) \left[ p_{12}^2 + \left( \begin{array}{c} 4 \\ 3 \end{array} \right) (n - 4)(p_{12}^2 + p_{11}^2) + \left( \begin{array}{c} 4 \\ 2 \end{array} \right) \frac{(n - 4)}{2} (p_{12}^2 + p_{11}^2) \right]$$

$$= (1 \wedge \lambda^{-1}) \Theta(n^4 p_{12}^4 + n^5 p_{12}^2 + n^6 p_{11}^2)$$

$$= (1 \wedge \lambda^{-1}) \Theta(n^{4} n^{-8} + n^{5} n^{-6} + n^{6} n^{-22/3})$$

$$= (1 \wedge \lambda^{-1}) \Theta(n^{-1}),$$

which goes to 0 as $n \to +\infty$ when $p_n = Cn^{-2/3}$.

See Exercise 4.14 for an improved bound.

**Exercises**

**Exercise 4.1** (Harris’ inequality: alternative proof). We say that $f : \mathbb{R}^n \to \mathbb{R}$ is nondecreasing if it is nondecreasing in each variable while keeping the other variables fixed.
• (Chebyshev’s association inequality) Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be nondecreasing and let $X$ be a real random variable. Show that

$$
\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)].
$$

[Hint: Consider the quantity $(f(X) - f(X'))(g(X) - g(X'))$ where $X'$ is an independent copy of $X$.]

• (Harris’ inequality) Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be nondecreasing and let $X = (X_1, \ldots, X_n)$ be independent real random variables. Show by induction on $n$ that

$$
\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)].
$$

Exercise 4.2. Provide the details for Example 4.51.

Exercise 4.3 (FKG: sufficient conditions). Let $\mathcal{X} := \{0, 1\}^F$ where $F$ is finite and let $\mu$ be a positive probability measure on $\mathcal{X}$. We use the notation introduced in the proof of Holley’s inequality (Theorem 4.50).

a) To check the FKG condition, show that it suffices to check that, for all $x \leq y \in \mathcal{X}$ and $i \in F$,

$$
\frac{\mu(y^i,1)}{\mu(y^i,0)} \geq \frac{\mu(x^i,1)}{\mu(x^i,0)}.
$$

[Hint: Write $\mu(\omega \vee \omega')/\mu(\omega)$ as a telescoping product.]

b) To check the FKG condition, show that it suffices to check (4.12) only for those $\omega, \omega' \in \mathcal{X}$ such that $||\omega - \omega'||_1 = 2$ and neither $\omega \leq \omega'$ nor $\omega' \leq \omega$.

[Hint: Use a).]

Exercise 4.4 (FKG and strong positive association). Let $\mathcal{X} := \{0, 1\}^F$ where $F$ is finite and let $\mu$ be a positive probability measure on $\mathcal{X}$. For $\Lambda \subseteq F$ and $\xi \in \mathcal{X}$, let

$$
\mathcal{X}_\Lambda^\xi := \{\omega_\Lambda \times \xi_{\Lambda^c} : \omega_\Lambda \in \{0, 1\}^\Lambda\},
$$

where $\omega_\Lambda \times \xi_{\Lambda^c}$ agrees with $\omega$ on coordinates in $\Lambda$ and with $\xi$ on coordinates in $F \setminus \Lambda$. Define the measure $\mu^\xi_\Lambda$ over $\{0, 1\}^\Lambda$ as

$$
\mu^\xi_\Lambda(\omega_\Lambda) := \frac{\mu(\omega_\Lambda \times \xi_{\Lambda^c})}{\mu(\mathcal{X}_\Lambda^\xi)}.
$$

That is, $\mu^\xi_\Lambda$ is $\mu$ conditioned on agreeing with $\xi$ on $F \setminus \Lambda$. The measure $\mu$ is said to be strongly positively associated if $\mu^\xi_\Lambda(\omega_\Lambda)$ is positively associated for all $\Lambda$ and $\xi$. Prove that the FKG condition is equivalent to strong positive association. [Hint: Use Exercise 4.3 as well as the FKG inequality.]
Exercise 4.5 (Triangle-freeness: a second proof). Consider again the setting of Section 4.3.5.

a) Let $e_t$ be the minimum number of edges in a $t$-vertex union of $k$ not mutually vertex-disjoint triangles. Show that, for any $k \geq 2$ and $k \leq t < 3k$, it holds that $e_t > t$.

b) Use Exercise 2.13 to give a second proof of the fact that $P[X_n = 0] \rightarrow e^{-\lambda^3/6}$.

Exercise 4.6 (RSW lemma: general $\alpha$). Let $R_{n,\alpha}(p)$ be as defined in Section 4.3.6. Show that for all $n \geq 2$ (divisible by 4) and $p \in (0, 1)$

$$R_{n,\alpha}(p) \geq \left(\frac{1}{2}\right)^{2\alpha-2} R_{n,1}(p)^{6\alpha-7} R_{n/2,1}(p)^{6\alpha-6}.$$

Exercise 4.7 (Primal and dual crossings). Modify the proof of Lemma 2.17 to prove Lemma 4.59.

Exercise 4.8 (Square-root trick). Let $\mu$ be an FKG measure on $\{0, 1\}^F$ where $F$ is finite. Let $A_1$ and $A_2$ be increasing events with $\mu(A_1) = \mu(A_2)$. Show that

$$\mu(A_1) \geq 1 - \sqrt{1 - \mu(A_1 \cup A_2)}.$$

Exercise 4.9 (Mixing on cycles: lower bound). Let $(Z_t)$ be lazy, simple random walk on the cycle of size $n$, $Z_n := \{0, 1, \ldots, n-1\}$, where $i \sim j$ if $|j - i| = 1$ (mod $n$).

a) Let $A = \{n/2, \ldots, n-1\}$. By coupling $(Z_t)$ with lazy, simple random walk on $Z$, show that

$$P^{\alpha n^2}(n/4, A) < \frac{1}{2} - \varepsilon,$$

for $\alpha \geq \alpha_c > 0$. [Hint: Use Kolmogorov’s maximal inequality (e.g. [Dur10, Theorem 2.5.2]).]

b) Deduce that

$$t_{\text{mix}}(\varepsilon) \geq \alpha_c n^2.$$

Exercise 4.10 (Lower bound on mixing: distinguishing statistic). Let $X$ and $Y$ be random variables on a finite state space $S$. Let $h : S \rightarrow \mathbb{R}$ be a measurable real-valued map. Assume that

$$\mathbb{E}[h(Y)] - \mathbb{E}[h(X)] \geq r\sigma,$$
where \( r > 0 \) and \( \sigma^2 := \max\{\text{Var}[h(X)], \text{Var}[h(Y)]\} \). Show that

\[
\|\mu_X - \mu_Y\|_{TV} \geq 1 - \frac{8}{r^2}.
\]

[Hint: Consider the interval on one side of the midpoint between \( \mathbb{E}[h(X)] \) and \( \mathbb{E}[h(Y)] \).]

**Exercise 4.11** (Path coupling and optimal transport). Let \( V \) be a finite state space and let \( P \) be an irreducible transition matrix on \( V \) with stationary distribution \( \pi \). Let \( w_0 \) be a metric on \( V \). For probability measures \( \mu, \nu \) on \( V \), let

\[
W_0(\mu, \nu) := \inf \{ \mathbb{E}[w_0(X,Y)] : (X,Y) \text{ is a coupling of } \mu \text{ and } \nu \},
\]

be the so-called *transportation metric* or *Wasserstein distance* between \( \mu \) and \( \nu \).

a) Show that \( W_0 \) is a metric. [Hint: Proof of Claim 4.74.]

b) Assume that the conditions of Theorem 4.73 hold. Show that for any probability measures \( \mu, \nu \)

\[
W_0(\mu P, \nu P) \leq \kappa W_0(\mu, \nu).
\]

c) Use a) and b) to prove Theorem 4.73.

**Exercise 4.12** (Stein equation for the Poisson distribution). Let \( \lambda > 0 \). Show that a non-negative integer-valued random variable \( Z \) is \( \text{Poi}(\lambda) \) if and only if for all \( g \) bounded

\[
\mathbb{E}[\lambda g(Z + 1) - Z g(Z)] = 0.
\]

**Exercise 4.13** (Chen-Stein for positively related variables). Using the notation in (4.39), (4.40) and (4.41), suppose that for each \( i \) we can construct a coupling \( \{(X_i : i = 1, \ldots, n), (Y_j^{(i)} : j \neq i)\} \) such that

\[
(Y_j^{(i)}, j \neq i) \sim (X_j, j \neq i) | X_i = 1 \quad \text{and} \quad Y_j^{(i)} \geq X_j, \forall j \neq i.
\]

Show that

\[
\|\mu - \pi\|_{TV} \leq (1 \wedge \lambda^{-1}) \left\{ \text{Var}(W) - \lambda + 2 \sum_{i=1}^{n} \mu_i^2 \right\}.
\]

**Exercise 4.14** (Chen-Stein and 4-cliques). Use Exercise 4.13 to give an improved asymptotic bound in the setting of Section 4.5.3.
Exercise 4.15 (Chen-Stein for negatively related variables). Using the notation in (4.39), (4.40) and (4.41), suppose that for each $i$ we can construct a coupling $\{(X_i : i = 1, \ldots, n), (Y_j^{(i)} : j \neq i)\}$ such that

$$(Y_j^{(i)}, j \neq i) \sim (X_j, j \neq i)|X_i = 1 \quad \text{and} \quad Y_j^{(i)} \leq X_j, \forall j \neq i.$$ 

Show that

$$\|\mu - \pi\|_{TV} \leq (1 \wedge \lambda^{-1}) \{\lambda - \text{Var}(W)\}.$$ 

Bibliographic remarks

Section 4.1  The coupling method is generally attributed to Doeblin [Doe38]. The standard reference on coupling is [Lin02]. See that reference for a history of coupling and a facsimile of Doeblin’s paper. See also [dH]. Section 4.2.2 is based on [vdH14, Section 5.3] and Section 4.1.2 is based on [Per, Section 6].

Section 4.3  Strassen’s theorem is due to Strassen [Str65]. Harris’ inequality is due to Harris [Har60]. The FKG inequality is due to Fortuin, Kasteleyn, and Ginibre [FKG71]. A “four-function” version of Holley’s inequality, which also extends to distributive lattices, was proved by Ahlswede and Daykin [AD78]. See e.g. [AS11, Section 6.1]. An exposition of submodularity and its connections to convexity can be found in [Lov83]. For more on Markov random fields, see e.g. [RAS]. Section 4.3.5 follows [AS11, Sections 8.1, 8.2, 10.1]. Janson’s inequality is due to Janson [Jan90]. Boppana and Spencer [BS89] gave the proof presented here. For more on Janson’s inequality, see [JLR11, Section 2.2]. The presentation in Section 4.3.6 follows closely [BR06b, Sections 3 and 4]. See also [BR06a, Chapter 3]. Broadbent and Hammersley [BH57, Ham57] initiated the study of the critical value of percolation. Harris’ theorem was proved by Harris [Har60] and Kesten’s theorem was proved two decades later by Kesten [Kes80], confirming non-rigorous work of Sykes and Essam [SE64]. The RSW theorem was obtained independently to Russo [Rus78] and Seymour and Welsh [SW78]. The proof we gave here is due to Bollobas and Riordan [BR06b]. Another short proof of a version of the RSW theorem for critical site percolation on the triangular lattice was given by Smirnov. See e.g. [Ste]. The type of “scale invariance” seen in the RSW theorem plays a key role in the contemporary theory of critical two-dimensional percolation and of two-dimensional lattice models more generally. See e.g. [Law05, Gri10a].

Section 4.4  The material in Section 4.4 borrows heavily from [LPW06, Chapters 5, 14, 15] and [AF, Chapter 12]. Aldous [Ald83] was the first author to make
explicit use of coupling to bound total variation distance to stationarity of finite Markov chains. The link between couplings of Markov chains and total variation distance was also used by Griffeath [Gri75] and Pitman [Pit76]. The proof of Claim 4.70 is partly based on [LPW06, Proposition 7.13]. See also [DGG+00] and [HS07] for alternative proofs. Path coupling is due to Bubley and Dyer [BD97]. The optimal transport perspective on the path coupling method in Exercise 4.11 is from [LPW06, Chapter 14]. For more on optimal transport, see e.g. [Vil09]. The main result in Section 4.4.4 is taken from [LPW06, Theorem 15.1]. For more background on the so-called “critical slowdown” of the Glauber dynamics of Ising and Potts models on various graphs, see [CDL+12, LS12]. The connection between sampling and counting was first considered by Jerrum, L. Valiant and V. Vazirani [JVV86]. For more on this topic, see e.g. [Sin93].

**Section 4.5** The Chen-Stein method was introduced by Chen in [Che75] as an adaptation of the Stein method [Ste72] to the Poisson distribution. The presentation in Section 4.5 is inspired by [Dey] and [vH16]. Example 4.85 is from [AGG89]. Further applications of the Chen-Stein and Stein methods to random graphs can be found in [JLR11, Chapter 6].
Chapter 5

Spectral methods

In this chapter, we develop spectral methods to study random graphs and Markov chains.

5.1 Spectral techniques for reversible Markov chains

Recall that a Markov chain with transition matrix $P$ is reversible w.r.t. a measure $\eta$ if $\eta(x)P(x, y) = \eta(y)P(y, x)$ for all $x, y \in V$. The main example we will be concerned with is the following.

Definition 5.1 (Random walk on a network). Let $G = (V, E)$ be a finite or countable, locally finite graph. Let $c : E \to \mathbb{R}_+$ be a positive edge weight function on $G$. We call $\mathcal{N} = (G, c)$ a network. We assume that the sum of the $c$’s adjacent to each vertex is finite. Random walk on $\mathcal{N}$ is the Markov chain on $V$, started at an arbitrary vertex, which at each time picks a neighbor of the current state proportionally to the weight of the corresponding edge.

Let $(X_t)$ be random walk on a network $\mathcal{N} = (G, c)$. Then $(X_t)$ is reversible w.r.t. $\eta(v) := c(v)$, where $c(v) := \sum_{x \sim v} c(v, x)$.

5.1.1 Spectral gap

In this section, we restrict ourselves to a finite state space $V$. We bound the mixing time in terms of the eigenvalues of the transition matrix. Suppose the transition
matrix $P$ is irreducible, aperiodic and has stationary distribution $\pi$. Recall that by the convergence theorem (Theorem 1.22), for all $x, y$, $P^t(x, y) \to \pi(y)$ as $t \to +\infty$. The mixing time is

$$t_{\text{mix}}(\varepsilon) := \min\{t \geq 0 : d(t) \leq \varepsilon\},$$

where $d(t) := \max_{x \in V} \|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}}$. It will be convenient to work with a different notion of distance.

**Definition 5.2** (Separation distance). The separation distance is defined as

$$s_x(t) := \max_{y \in V} \left[1 - \frac{P^t(x, y)}{\pi(y)}\right],$$

and we let $s(t) := \max_{x \in V} s_x(t)$.

Because both $\{\pi(y)\}$ and $\{P^t(x, y)\}$ are non-negative and sum to 1, we have that $s_x(t) \geq 0$.

**Lemma 5.3** (Separation distance v. total variation distance).

$$d(t) \leq s(t).$$

**Proof.** Because $1 = \sum_y \pi(y) = \sum_y P^t(x, y)$, we have

$$\sum_{y : P^t(x, y) < \pi(y)} \pi(y) - P^t(x, y) = \sum_{y : P^t(x, y) \geq \pi(y)} [P^t(x, y) - \pi(y)].$$

Hence,

$$\|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}} = \frac{1}{2} \sum_y |\pi(y) - P^t(x, y)| = \sum_{y : P^t(x, y) \leq \pi(y)} [\pi(y) - P^t(x, y)] = \sum_{y : P^t(x, y) \leq \pi(y)} \pi(y) \left[1 - \frac{P^t(x, y)}{\pi(y)}\right] \leq s_x(t).$$
We let $n := |V| < +\infty$. Assume that $P$ is irreducible and reversible w.r.t. its stationary distribution $\pi > 0$. Define

$$\langle f, g \rangle_\pi := \sum_{x \in V} \pi(x) f(x) g(x), \quad \|f\|_\pi^2 := \langle f, f \rangle_\pi,$$

$$(Pf)(x) := \sum_y P(x, y) f(y).$$

We let $\ell^2(V, \pi)$ be the Hilbert space of real-valued functions on $V$ equipped with the inner product $\langle \cdot, \cdot \rangle_\pi$ (equivalent to the vector space $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_\pi)$). Think of $P$ as an operator on $\ell^2(V, \pi)$.

**Theorem 5.4** (Reversible chain: eigendecomposition). If $P$ is finite and reversible w.r.t. $\pi$, then there is an orthonormal basis of $\ell^2(V, \pi)$ formed of real eigenfunctions $\{f_j\}_{j=1}^n$ of $P$ with real eigenvalues $\{\lambda_j\}_{j=1}^n$. We can take $f_1 \equiv 1$ and $\lambda_1 = 1$.

**Proof.** We work over $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_\pi)$. Let $D_\pi$ be the diagonal matrix with $\pi$ on the diagonal. By reversibility,

$$M(x, y) := (D_\pi^{1/2} P D_\pi^{-1/2})_{xy} = \sqrt{\frac{\pi(x)}{\pi(y)}} P(x, y) = \sqrt{\frac{\pi(y)}{\pi(x)}} P(y, x) = (D_\pi^{1/2} P D_\pi^{-1/2})_{yx} =: M(y, x).$$

So $M = (M(x, y))_{x,y} = D_\pi^{1/2} P D_\pi^{-1/2}$, as a symmetric matrix, has real eigenvectors $\{\phi_j\}_{j=1}^n$ forming an orthonormal basis of $\mathbb{R}^n$ with corresponding real eigenvalues $\{\lambda_j\}_{j=1}^n$ (see e.g. [HJ13, Section 2.5]). Define $f_j := D_\pi^{-1/2} \phi_j$. Then

$$P f_j = D_\pi^{-1/2} \phi_j = D_\pi^{-1/2} D_\pi^{1/2} P D_\pi^{-1/2} \phi_j = D_\pi^{-1/2} M \phi_j = \lambda_j D_\pi^{-1/2} \phi_j = \lambda_j f_j,$$
and

\[
\langle f_i, f_j \rangle_\pi = \langle D_\pi^{-1/2} \phi_i, D_\pi^{-1/2} \phi_j \rangle_\pi
= \sum_x \pi(x)[\pi(x)^{-1/2}\phi_i(x)][\pi(x)^{-1/2}\phi_j(x)]
= \langle \phi_i, \phi_j \rangle_\pi.
\]

Because \(P\) is stochastic, the all-one vector is a right eigenvector of \(P\) with eigenvalue 1. Because \(\{\phi_j\}\) is an orthonormal basis of \(\mathbb{R}^n\), we have that \(\{f_j\}_j\) is an orthonormal basis of \((\mathbb{R}^n, \langle \cdot, \cdot \rangle_\pi)\).

We collect a few more basic facts about the eigenbasis.

**Lemma 5.5 (Expectation).** For all \(j \neq 1\),

\[
\sum_x \pi(x)f_j(x) = 0.
\]

**Proof.** By orthonormality, \(\langle f_1, f_j \rangle_\pi = 0\). Now use the fact that \(f_1 \equiv 1\).

**Lemma 5.6 (Orthogonality).** Let \(\delta_x(y) := 1_{\{x=y\}}\). For all \(x, y\),

\[
\sum_{j=1}^n f_j(x)f_j(y) = \pi(x)^{-1}\delta_x(y).
\]

**Proof.** Using the notation of the Theorem 5.4, the matrix \(\Phi\) whose columns are the \(\phi_j\)s is unitary so \(\Phi \Phi^\prime = I\). That is,

\[
\sum_{j=1}^n \phi_j(x)\phi_j(y) = \delta_x(y),
\]

or

\[
\sum_{j=1}^n \sqrt{\pi(x)\pi(y)}f_j(x)f_j(y) = \delta_x(y).
\]

Rearranging gives the result.

**Lemma 5.7 (Decomposition).** Let \(g \in \ell^2(V, \pi)\). Then

\[
g = \sum_{j=1}^n \langle g, f_j \rangle_\pi f_j.
\]

271
Proof. By Lemma 5.6, for all \(x\)

\[
\sum_{j=1}^{n} \langle g, f_j \rangle \pi f_j(x) = \sum_{j=1}^{n} \sum_y \pi(y) g(y) f_j(y) f_j(x) \\
= \sum_y \pi(y) g(y) [\pi(x)^{-1} \delta_x(y)] \\
= g(x).
\]

Lemma 5.8 (Norm). Let \(g \in \ell^2(V, \pi)\). Then

\[
\|g\|_\pi^2 = \sum_{j=1}^{n} \langle g, f_j \rangle^2 \pi.
\]

Proof. By the previous Lemma 5.7,

\[
\|g\|_\pi^2 = \left\| \sum_{j=1}^{n} \langle g, f_j \rangle \pi f_j \right\|_\pi^2 \\
= \left\langle \sum_{i=1}^{n} \langle g, f_i \rangle \pi f_i, \sum_{j=1}^{n} \langle g, f_j \rangle \pi f_j \right\rangle_\pi \\
= \sum_{i,j} \langle g, f_i \rangle \pi \langle g, f_j \rangle \pi \langle f_i, f_j \rangle_\pi,
\]

which gives the result by the orthonormality of the \(f_j\)s under \(\langle \cdot, \cdot \rangle_\pi\).

Eigenvalues: variational characterization  Let \(P\) be finite, irreducible and reversible. We first recall a basic fact about the eigenvalues of a transition matrix.

Lemma 5.9. Any eigenvalue \(\lambda\) of \(P\) (or of its transpose \(P'\)) satisfies \(|\lambda| \leq 1\).

Proof. The matrix \(P\) and its transpose have the same eigenvalues because the characteristic polynomial is unaffected by transposing. We have

\[
P f = \lambda f \implies |\lambda||f|_\infty = \|P f\|_\infty = \max_x \left| \sum_y P(x, y) f(y) \right| \leq \|f\|_\infty.
\]

272
We order the eigenvalues $1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq -1$. The second eigenvalue will play an important role below.

**Lemma 5.10.** We have $\lambda_1 = 1$ and $\lambda_2 < 1$. Also we can take $f_1 \equiv 1$.

**Proof.** Because $P$ is stochastic, the all-one vector is a right eigenvector with eigenvalue 1. Any eigenfunction with eigenvalue 1 is $P$-harmonic. By Corollary 3.84, for a finite, irreducible chain the only harmonic functions are the constant functions. So the eigenspace corresponding to 1 is one-dimensional. Since all eigenvalues are real, we must have $\lambda_2 < 1$.

The following variational characterization of the second eigenvalue will be useful.

**Theorem 5.11** (Rayleigh’s quotient). Let $P$ be finite, irreducible and reversible with respect to $\pi$. The second largest eigenvalue is characterized by

$$\lambda_2 = \sup \left\{ \frac{\langle f, Pf \rangle}{\langle f, f \rangle} : f \in \ell^2(V, \pi), \sum_{x} \pi(x) f(x) = 0 \right\}.$$ 

(Similarly, $\lambda_1 = \sup_{f \in \ell^2(V, \pi)} \frac{\langle f, Pf \rangle}{\langle f, f \rangle}$.)

**Proof.** Recalling that $f_1 \equiv 1$, the condition $\sum_x \pi(x) f(x) = 0$ is equivalent to $\langle f_1, f \rangle = 0$. For such an $f$, the eigendecomposition is

$$f = \sum_{j=1}^{n} \langle f, f_j \rangle \pi f_j = \sum_{j=2}^{n} \langle f, f_j \rangle \pi f_j,$$

and

$$P f = \sum_{j=2}^{n} \langle f, f_j \rangle \pi \lambda_j f_j,$$

so that

$$\frac{\langle f, Pf \rangle}{\langle f, f \rangle} = \frac{\sum_{i=2}^{n} \sum_{j=2}^{n} \langle f, f_i \rangle \pi \langle f, f_j \rangle \pi \lambda_j \langle f_i, f_j \rangle}{\sum_{j=2}^{n} \langle f, f_j \rangle^2 \pi} = \frac{\sum_{j=2}^{n} \langle f, f_j \rangle^2 \pi \lambda_j}{\sum_{j=2}^{n} \langle f, f_j \rangle^2 \pi} \leq \lambda_2.$$ 

Taking $f = f_2$ achieves the supremum. □
The Dirichlet form is defined as $E(f, g) := \langle f, (I - P)g \rangle_\pi$. Note that

\begin{align*}
2\langle f, (I - P)f \rangle_\pi &= 2\langle f, f \rangle_\pi - 2\langle f, Pf \rangle_\pi \\
&= \sum_x \pi(x)f(x)^2 + \sum_y \pi(y)f(y)^2 - 2\sum_x \pi(x)f(x)f(y)P(x, y) \\
&= \sum_{x, y} f(x)^2\pi(x)P(x, y) + \sum_{x, y} f(y)^2\pi(y)P(y, x) - 2\sum_x \pi(x)f(x)f(y)P(x, y) \\
&= \sum_{x, y} f(x)^2\pi(x)P(x, y) + \sum_{x, y} f(y)^2\pi(x)P(x, y) - 2\sum_x \pi(x)f(x)f(y)P(x, y) \\
&= \sum_{x, y} \pi(x)P(x, y)[f(x) - f(y)]^2 = 2E(f),
\end{align*}

where

\[ E(f) := \frac{1}{2} \sum_{x, y} c(x, y)[f(x) - f(y)]^2, \]

is the Dirichlet energy encountered previously.

We note further that if $\sum_x \pi(x)f(x) = 0$ then

\[ \langle f, f \rangle_\pi = \langle f - \langle 1, f \rangle_\pi, f - \langle 1, f \rangle_\pi \rangle_\pi = \text{Var}_\pi[f], \]

where the last expression denotes the variance under $\pi$. So the variational characterization of $\lambda_2$ translates into

\[ \text{Var}_\pi[f] \leq \gamma^{-1}E(f), \]

where $\gamma = 1 - \lambda_2$, for all $f$ such that $\sum_x \pi(x)f(x) = 0$ (in fact for any $f$ by considering $f - \langle 1, f \rangle_\pi$ and noticing that both sides are unaffected by adding a constant), which is known as a Poincaré inequality.

**Spectral decomposition** Using the eigendecomposition of $P$, we get the following expression for $P^t$.

**Theorem 5.12** (Spectral decomposition of $P^t$). Let $\{f_j\}_{j=1}^n$ be the eigenfunctions of a reversible and irreducible transition matrix $P$ with corresponding eigenvalues $\{\lambda_j\}_{j=1}^n$, as defined previously. Assume $\lambda_1 \geq \cdots \geq \lambda_n$. We have the decomposition

\[ \frac{P^t(x, y)}{\pi(y)} = 1 + \sum_{j=2}^n f_j(x)f_j(y)\lambda_j^t, \]

274
Proof. Let $F$ be the matrix whose columns are the eigenvectors $\{f_j\}_{j=1}^n$ and let $D_\lambda$ be the diagonal matrix with $\{\lambda_j\}_{j=1}^n$ on the diagonal. Using the notation of the proof of the eigenbasis theorem, Theorem 5.4,

$$D_\pi^{1/2} P^t D_\pi^{-1/2} = M^t = (D_\pi^{1/2} F) D_\lambda (D_\pi^{1/2} F)^t,$$

which after rearranging becomes

$$P^t D_\pi^{-1} = F D_\lambda F^t.$$

Expanding gives the result, where we used Lemma 5.6.

Example 5.13 (Two-state chain). Let $V := \{0, 1\}$ and, for $\alpha, \beta \in (0, 1),$

$$P := \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}.$$

Observe that $P$ is reversible w.r.t. to the stationary distribution

$$\pi := \begin{pmatrix} \frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta} \end{pmatrix}.$$

We know that $f_1 \equiv 1$ is an eigenfunction with eigenvalue 1. As can be checked by direct computation, the other eigenfunction (in vector form) is

$$f_2 := \left( \frac{\alpha}{\sqrt{\beta}}, -\frac{\sqrt{\alpha}}{\beta} \right)' ,$$

with eigenvalue $\lambda_2 := 1 - \alpha - \beta$. We normalized $f_2$ so that $\|f_2\|_2^2 = 1$.

By Theorem 5.12, the spectral decomposition is therefore

$$P^t D_\pi^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + (1 - \alpha - \beta)^t \begin{pmatrix} \frac{\alpha}{\beta} & -1 \\ -1 & \frac{\beta}{\alpha} \end{pmatrix}.$$

Or, rearranging,

$$P^t = \begin{pmatrix} \frac{\beta}{\alpha + \beta} & \frac{\alpha}{\alpha + \beta} \\ \frac{\alpha}{\alpha + \beta} & \frac{\beta}{\alpha + \beta} \end{pmatrix} + (1 - \alpha - \beta)^t \begin{pmatrix} \frac{\alpha + \beta}{\beta} & -\frac{\alpha}{\alpha + \beta} \\ -\frac{\alpha}{\alpha + \beta} & \frac{\alpha + \beta}{\beta} \end{pmatrix}.$$

(Note for instance that the case $\alpha + \beta = 1$ corresponds to a rank-one $P$, which immediately converges to stationarity.)
Assume $\beta \geq \alpha$. Then

$$d(t) = \max_x \frac{1}{2} \sum_y |P^t(x, y) - \pi(y)| = \frac{\beta}{\alpha + \beta} |1 - \alpha - \beta|^t.$$ 

As a result,

$$t_{\text{mix}}(\varepsilon) = \left[ \frac{\log \left( \frac{\varepsilon (\alpha + \beta)}{\beta} \right)}{\log |1 - \alpha - \beta|} \right] = \left[ \frac{\log \left( \varepsilon^{-1} - \log \left( \frac{\alpha + \beta}{\beta} \right) \right)}{\log |1 - \alpha - \beta|^{-1}} \right].$$

\[\boxed{}\]

**Spectral gap** From the spectral decomposition, the speed of convergence of $P^t(x, y)$ to $\pi(y)$ is dominated by the largest eigenvalue of $P$ not equal to 1.

**Definition 5.14** (Spectral gap). The absolute spectral gap is $\gamma_* := 1 - \lambda_*$ where $\lambda_* := \max \{ |\lambda_j| \}$. The spectral gap is $\gamma := 1 - \lambda_2$.

Note that the eigenvalues of the lazy version $\frac{1}{2}P + \frac{1}{2}I$ of $P$ are $\{\frac{1}{2}(\lambda_j + 1)\}_{j=1}^n$, which are all nonnegative. So, there, $\gamma_* = \gamma$.

**Definition 5.15** (Relaxation time). The relaxation time is defined as

$$t_{\text{rel}} := \gamma_*^{-1}.$$

**Example 5.16** (Two-state chain (continued)). Returning to Example 5.13, there are two cases:

- $\alpha + \beta \leq 1$: In that case the spectral gap is $\gamma = \gamma_* = \alpha + \beta$ and the relaxation time is $t_{\text{rel}} = 1/(\alpha + \beta)$.

- $\alpha + \beta > 1$: In that case the spectral gap is $\gamma = \gamma_* = 2 - \alpha - \beta$ and the relaxation time is $t_{\text{rel}} = 1/(2 - \alpha - \beta)$.

\[\boxed{}\]

The following result clarifies the relationship between the mixing and relaxation times in general.

**Theorem 5.17** (Mixing time v. relaxation time). Let $P$ be reversible, irreducible, and aperiodic with stationary distribution $\pi$. Let $\pi_{\text{min}} = \min_x \pi(x)$. For all $\varepsilon > 0$,

$$(t_{\text{rel}} - 1) \log \left( \frac{1}{2\varepsilon} \right) \leq t_{\text{mix}}(\varepsilon) \leq \log \left( \frac{1}{\varepsilon \pi_{\text{min}}} \right) t_{\text{rel}}.$$
Proof. We start with the upper bound. By Lemma 5.3, it suffices to find \( t \) such that 
\[ s(t) \leq \varepsilon. \]
By the spectral decomposition and Cauchy-Schwarz,
\[
\left| \frac{P^t(x,y)}{\pi(y)} - 1 \right| \leq \lambda_*^t \sum_{j=2}^n |f_j(x)f_j(y)| \leq \lambda_*^t \sqrt{\sum_{j=2}^n f_j(x)^2 \sum_{j=2}^n f_j(y)^2}.
\]
By Lemma 5.6, \( \sum_{j=2}^n f_j(x)^2 \leq \pi(x) \). Plugging this back above, we get
\[
\left| \frac{P^t(x,y)}{\pi(y)} - 1 \right| \leq \lambda_*^t \sqrt{\pi(x)^{-1} \pi(y)^{-1}} \leq \frac{\lambda_*^t}{\pi_{\text{min}}} \leq \frac{e^{-\gamma_* t}}{\pi_{\text{min}}}. 
\]
Observe that the r.h.s. is less than \( \varepsilon \) when \( t \geq \log \left( \frac{1}{\varepsilon \pi_{\text{min}}} \right) t_{\text{rel}}. \)

For the lower bound, let \( f_* \) be an eigenfunction associated with an eigenvalue achieving \( \lambda_* := |\lambda_2| \vee |\lambda_1|. \) Let \( z \) be such that \( |f_*(z)| = \|f_*\|_\infty. \) By our previous lemma, \( \sum_y \pi(y)f_*(y) = 0. \) Hence
\[
\lambda_*^t |f_*(z)| = |P^t f_*(z)| 
\]
\[
= \left| \sum_y \left[ P^t(z,y)f_*(y) - \pi(y)f_*(y) \right] \right| 
\]
\[
\leq \|f_*\|_\infty \sum_y |P^t(z,y) - \pi(y)| \leq \|f_*\|_\infty 2d(t),
\]
so \( d(t) \geq \frac{1}{2} \lambda_*^t. \) When \( t = t_{\text{mix}}(\varepsilon), \varepsilon \geq \frac{1}{2} \lambda_*^{t_{\text{mix}}(\varepsilon)}. \) Therefore
\[
t_{\text{mix}}(\varepsilon) \left( \frac{1}{\lambda_*} - 1 \right) \geq t_{\text{mix}}(\varepsilon) \log \left( \frac{1}{\lambda_*} \right) \geq \log \left( \frac{1}{2\varepsilon} \right).
\]
The result follows from \( \left( \frac{1}{\lambda_*} - 1 \right)^{-1} = \left( \frac{1}{1-\lambda_*} \right)^{-1} = \left( \gamma_* \frac{1}{\lambda_*} \right)^{-1} = t_{\text{rel}} - 1. \)

5.1.2 Markov chains: random walk on cycles and hypercubes revisited

We apply the previous results to random walks on cycles and hypercubes.

**Random walk on the cycle** Consider simple random walk on the \( n \)-cycle. That is, \( V := \{0, 1, \ldots, n-1\} \) and \( P(x,y) = 1/2 \) if and only if \( |x-y| = 1 \mod n. \) We showed before in Section 4.4.2 that in fact \( t_{\text{mix}}(\varepsilon) = \Theta(n^2). \) We re-derive this result using spectral techniques.

We first obtain the eigendecomposition, which in this case can be determined explicitly.
**Lemma 5.18** (Eigenbasis on the cycle). For \( j = 0, \ldots, n - 1 \), the function

\[
f_j(x) := \cos \left( \frac{2\pi j x}{n} \right), \quad x = 0, 1, \ldots, n - 1,
\]

is an eigenfunction of \( P \) with eigenvalue

\[
\lambda_j := \cos \left( \frac{2\pi j}{n} \right).
\]

**Proof.** Note that, for all \( i, x \),

\[
\sum_y P(x, y) f_j(y) = \frac{1}{2} \left[ \cos \left( \frac{2\pi j (x - 1)}{n} \right) + \cos \left( \frac{2\pi j (x + 1)}{n} \right) \right]
\]

\[
= \frac{1}{2} \left[ \frac{e^{i \frac{2\pi j (x - 1)}{n}} + e^{-i \frac{2\pi j (x - 1)}{n}}}{2} + \frac{e^{i \frac{2\pi j (x + 1)}{n}} + e^{-i \frac{2\pi j (x + 1)}{n}}}{2} \right]
\]

\[
= \left[ \cos \left( \frac{2\pi j x}{n} \right) \right] \left[ \cos \left( \frac{2\pi j}{n} \right) \right]
\]

\[
= \cos \left( \frac{2\pi j}{n} \right) f_j(x).
\]

From the eigenvalues, we derive the relaxation time.

**Theorem 5.19** (Relaxation time on the cycle). The relaxation time for lazy simple random walk on the cycle is

\[
t_{\text{rel}} = \frac{2}{1 - \cos \left( \frac{2\pi}{n} \right)} = \Theta(n^2).
\]

**Proof.** The eigenvalues are

\[
\frac{1}{2} \left[ \cos \left( \frac{2\pi j}{n} \right) + 1 \right].
\]

The spectral gap is therefore \( \frac{1}{2} \left( 1 - \cos \left( \frac{2\pi}{n} \right) \right) \). By a Taylor expansion,

\[
1 - \cos \left( \frac{2\pi}{n} \right) = \frac{4\pi^2}{n^2} + O(n^{-4}).
\]
Since \( \pi_{\min} = 1/n \), we get \( t_{\text{mix}}(\varepsilon) = O(n^2 \log n) \) and \( t_{\text{mix}}(\varepsilon) = \Omega(n^2) \).

Our upper bound is off by a logarithmic factor. In this case, a sharper bound on the mixing time can be obtained by working directly with the spectral decomposition. By Jensen’s inequality,

\[
4\|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}}^2 = \left\{ \sum_y \pi(y) \left| \frac{P^t(x, y)}{\pi(y)} - 1 \right| \right\}^2 \\
\leq \sum_y \pi(y) \left( \frac{P^t(x, y)}{\pi(y)} - 1 \right)^2 \\
= \left\| \sum_{j=2}^n \lambda_j^t f_j(x) f_j \right\|_{\pi}^2 \\
= \sum_{j=2}^n \lambda_j^{2t} f_j(x)^2.
\]

Here comes the trick. The last sum does not depend on \( x \) by symmetry. Summing over \( x \) and dividing by \( n \), which is the same as multiplying by \( \pi(x) \), gives

\[
4\|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}}^2 \leq \sum_x \pi(x) \sum_{j=2}^n \lambda_j^{2t} f_j(x)^2 \\
= \sum_{j=2}^n \lambda_j^{2t} \sum_x \pi(x) f_j(x)^2 \\
= \sum_{j=2}^n \lambda_j^{2t},
\]

where we used that \( \|f_j\|_{\pi}^2 = 1 \).

Consider the non-lazy chain with \( n \) odd. We get

\[
4d(t)^2 \leq \sum_{j=2}^n \cos \left( \frac{2\pi j}{n} \right)^{2t} = 2 \sum_{j=1}^{(n-1)/2} \cos \left( \frac{\pi j}{n} \right)^{2t}.
\]

For \( x \in [0, \pi/2) \), \( \cos x \leq e^{x^2/2} \). (Indeed, let \( h(x) = \log(e^{x^2/2} \cos x) \). Then \( h'(x) = x - \tan x \leq 0 \) since \( (\tan x)' = 1 + \tan^2 x \geq 1 \) for all \( x \) and \( \tan 0 = 0 \).
So $h(x) \leq h(0) = 0$.) Then

$$4d(t)^2 \leq 2 \sum_{j=1}^{(n-1)/2} \exp\left(-\frac{\pi^2 j^2 t}{n^2}\right) \leq 2 \exp\left(-\frac{\pi^2 t}{n^2}\right) \sum_{j=1}^{\infty} \exp\left(-\frac{\pi^2 (j^2 - 1)}{n^2} t\right)$$

$$\leq 2 \exp\left(-\frac{\pi^2 t}{n^2}\right) \sum_{\ell=0}^{\infty} \exp\left(-\frac{3\pi^2 t}{n^2} \ell\right) = \frac{2 \exp\left(-\frac{\pi^2 t}{n^2}\right)}{1 - \exp\left(-\frac{3\pi^2 t}{n^2}\right)},$$

where we used that $j^2 - 1 \geq 3(j-1)$ for all $j = 1, 2, 3, \ldots$ So $t_{\text{mix}}(\varepsilon) = O(n^2)$.

**Random walk on the hypercube** Consider simple random walk on the hypercube $V := \{-1, +1\}^n$ where $x \sim y$ if they differ at exactly one coordinate. We showed in Section 4.4.2 that $t_{\text{mix}}(\varepsilon) = \Theta(n \log n)$. We rederive this result using spectral techniques.

For $J \subseteq [n]$, we let $\chi_J(x) = \prod_{j \in J} x_j$, $x \in V$.

These are called *parity functions*. We show that the parity functions form an eigenbasis of the transition matrix.

**Lemma 5.20** (Eigenbasis on the hypercube). For all $J \subseteq [n]$, the function $\chi_J$ is an eigenfunction of $P$ with eigenvalue

$$\lambda_J := \frac{n - 2|J|}{n}.$$

**Proof.** For $x \in V$ and $i \in [n]$, let $x^{[i]}$ be $x$ where coordinate $i$ is flipped. Note that, for all $J, x$,

$$\sum_y P(x, y) \chi_J(y) = \sum_{i=1}^n \frac{1}{n} \chi_J(x^{[i]}) = \frac{n - |J|}{n} \chi_J(x) - \frac{|J|}{n} \chi_J(x) = \frac{n - 2|J|}{n} \chi_J(x).$$

From the eigenvalues, we obtain the relaxation time.

**Theorem 5.21** (Relaxation time on the hypercube). The relaxation time for lazy simple random walk on the hypercube is

$$t_{\text{rel}} = n.$$
Proof: The eigenvalues are $\frac{n-|J|}{n}$ for $J \subseteq [n]$. The spectral gap is $\gamma_* = \gamma = 1 - \frac{n-1}{n} = \frac{1}{n}$.

Because $|V| = 2^n$, $\pi_{\text{min}} = 1/2^n$. Hence we have $t_{\text{mix}}(\varepsilon) = O(n^2)$ and $t_{\text{mix}}(\varepsilon) = \Omega(n)$. These bounds are off by a logarithmic factor.

As we did for the cycle, we obtain a sharper bound by working directly with the spectral decomposition. By the same argument,

$$4d(t)^2 \leq \sum_{J \neq \emptyset} \lambda_J^2.$$ 

Consider the lazy chain again. Then

$$4d(t)^2 \leq \sum_{J \neq \emptyset} \left( \frac{n-|J|}{n} \right)^{2t} = \sum_{\ell=1}^{n} \left( \frac{n}{\ell} \right)^{2t} \leq \sum_{\ell=1}^{n} \left( \frac{n}{\ell} \right)^{2t} \exp \left( -\frac{2t\ell}{n} \right) = \left( 1 + \exp \left( -\frac{2t}{n} \right) \right)^n - 1.$$ 

So $t_{\text{mix}}(\varepsilon) \leq \frac{1}{2} n \log n + O(n)$.

5.1.3 Canonical paths and comparison

To be written. See [Ber14, Sections 3 and 4].

5.1.4 Spectral radius

The previous results cannot in general be extended to infinite networks. Suppose $P$ is irreducible, aperiodic and positive recurrent. Then it can be shown (see e.g. [LPW06, Theorem 21.14]) that, if $\pi$ is the stationary distribution, then for all $x$

$$\|P^t(x, \cdot) - \pi(\cdot)\|_{\text{TV}} \to 0,$$

as $t \to +\infty$. However, one needs stronger conditions on $P$ than reversibility for the spectral theorem to apply (in a form similar to what we used above), e.g., compactness (that is, $P$ maps bounded sets to relatively compact sets, i.e. sets whose closure is compact).

Example 5.22 (A positive recurrent chain whose $P$ is not compact). For $p < 1/2$, let $(X_t)$ be the birth-death chain with $V := \{0, 1, 2, \ldots\}$, $P(0, 0) := 1 - p$, $P(0, 1) = p$, $P(x, x+1) := p$ and $P(x, x-1) := 1 - p$ for all $x \geq 1$, and $P(x,y) := 0$ if $|x-y| > 1$. As can be checked by direct computation, $P$ is reversible with respect to the stationary distribution $\pi(x) = (1-\gamma)^x \gamma$ for $x \geq 0$. 

281
where \( \gamma := \frac{p}{1-p} \). For \( j \geq 1 \), define \( g_j(x) := \pi(j)^{-1/2}1_{\{x=j\}} \). Then \( \|g_j\|_\pi^2 = 1 \) for all \( j \) so \( \{g_j\} \) is bounded in \( \ell^2(V, \pi) \). On the other hand,

\[
P g_j(x) = p\pi(j)^{-1/2}1_{\{x=j-1\}} + (1-p)\pi(j)^{-1/2}1_{\{x=j+1\}}.
\]

So

\[
\|P g_j\|_\pi^2 = p^2\pi(j)^{-1}\pi(j-1) + (1-p)^2\pi(j)^{-1}\pi(j+1) = p^2\frac{1-p}{p} + (1-p)^2\frac{p}{1-p} = 2p(1-p).
\]

Hence \( \{P g_j\} \) is also bounded. However, for \( j > \ell \)

\[
\|P g_j - P g_\ell\|_\pi^2 = (1-p)^2\pi(j)^{-1}\pi(j+1) + p^2\pi(\ell)^{-1}\pi(\ell-1) = 2p(1-p).
\]

So \( \{P g_j\} \) does not have a converging subsequence and therefore is not relatively compact.

In this section, we focus on random walks on infinite networks that are either transient or null recurrent. In such cases, there is no stationary distribution to converge to. In fact, we prove the following.

**Theorem 5.23** (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 \( x, y \)

\[
\lim_{t \to \infty} P^t(x, y) = 0.
\]

**Proof.** In the transient case, \( \sum_x 1_{X_t = y} \) a.s. under \( \mathbb{P}_x \), we have

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

so \( P^t(x, y) \to 0 \).

Consider the null recurrent case. Fix \( x \in V \). We observe first that:

- It suffices to show that \( P^t(x, x) \to 0 \). Indeed, by irreducibility, for any \( y \) there is \( s > 0 \) such that \( P^s(x, y) > 0 \). So \( P^{t+s}(x, x) \geq P^t(x, y)P^s(y, x) \) so \( P^t(x, x) \to 0 \) implies \( P^t(x, y) \to 0 \).
Let \( \ell = \gcd \{ t : P^t(x,x) > 0 \} \). As \( P^t(x,x) = 0 \) for any \( t \) that is not a multiple of \( \ell \), it suffices to consider the transition matrix \( \tilde{P} := P^\ell \). That corresponds to “looking at the chain” at times \( k\ell, k \geq 0 \). We restrict the state space to \( \tilde{V} := \{ y \in V : \exists s \geq 0, \tilde{P}^s(x,y) > 0 \} \). Let \( (\tilde{X}_t) \) be the corresponding chain, and let \( \tilde{P}_x \) and \( \tilde{E}_x \) be the corresponding measure and expectation. Clearly we still have \( \tilde{P}_x[\tau^+_x < +\infty] = 1 \) and \( \tilde{E}_x[\tau^+_x] = +\infty \) because returns to \( x \) under \( P \) can only happen at times that are multiples of \( \ell \). The reason to consider \( \tilde{P} \) is that it is irreducible and aperiodic, as we show next. Note that the irreducibility of \( \tilde{P} \) also implies that \( \tilde{P} \) is null recurrent.

We first show that \( \tilde{P} \) is irreducible. By definition of \( \tilde{V} \), it suffices to prove that, for any \( w \in \tilde{V} \), there exists \( s \geq 0 \) such that \( \tilde{P}^s(w,x) > 0 \). Indeed that then implies that all states in \( \tilde{V} \) communicate through \( x \). Let \( r \geq 0 \) be such that \( \tilde{P}^r(x,w) > 0 \). If it were the case that \( \tilde{P}^s(w,x) = 0 \) for all \( s \geq 0 \), that would imply that \( \tilde{P}_x[\tau^+_x = +\infty] > \tilde{P}^r(x,w) > 0 \)—a contradiction.

We claim further that \( \tilde{P} \) is aperiodic. Indeed, if \( \tilde{P} \) had period \( k > 1 \), then the greatest common divisor of \( \{ t : P^t(x,x) > 0 \} \) would be \( \geq k\ell \)—a contradiction.

The chain \( (\tilde{X}_t) \) has stationary measure

\[
\mu_x(w) = \tilde{E}_x \left[ \sum_{s=0}^{\tau^+_x-1} 1\{ \tilde{X}_s = w \} \right] < +\infty,
\]

which satisfies \( \mu_x(x) = 1 \) by definition and \( \sum_w \mu_x(w) = +\infty \) by null recurrence.

**Lemma 5.24.** For any probability distribution \( \nu \) on \( \tilde{V} \),

\[
\limsup_t \nu \tilde{P}^t(x) \leq \limsup_t \tilde{P}^t(x,x).
\]

**Proof.** Since \( \tilde{P}_\nu[\tau^+_x = +\infty] = 0 \), for any \( \varepsilon > 0 \) there is \( N \) such that \( \tilde{P}_\nu[\tau^+_x > N] \leq \varepsilon \). So,

\[
\limsup_t \nu \tilde{P}^t(x) \leq \varepsilon + \limsup_t \sum_{s=1}^N \tilde{P}_\nu[\tau^+_x = s] \tilde{P}^{t-s}(x,x) \leq \varepsilon + \limsup_t \tilde{P}^t(x,x).
\]

Since \( \varepsilon \) is arbitrary, the result follows. \( \blacksquare \)
For $M \geq 0$, let $F \subseteq \tilde{V}$ be a finite set such that $\mu_x(F) \geq M$. Consider the conditional distribution

$$\nu_F(W) := \frac{\mu_x(W \cap F)}{\mu_x(F)}.$$ 

**Lemma 5.25.**

$$(\nu_F \tilde{P}^t)(x) \leq \frac{1}{M}, \quad \forall t$$

**Proof.** Indeed

$$(\nu_F \tilde{P}^t)(x) \leq \left(\frac{\mu_x \tilde{P}^t(x)}{\mu_x(F)}\right) = \frac{\mu_x(x)}{\mu_x(F)} \leq \frac{1}{M},$$

by stationarity. □

Because $F$ is finite and $Q$ is aperiodic, there is $m$ such that $\tilde{P}^m(x, z) > 0$ for all $z \in F$. Then we can choose $\delta > 0$ such that

$$\tilde{P}^m(x, \cdot) = \delta \nu_F(\cdot) + (1 - \delta) \nu_0(\cdot),$$

for some probability measure $\nu_0$. Then

$$\limsup_t \tilde{P}^t(x, x) = \delta \limsup_t (\nu_F \tilde{P}^{t-m})(x) + (1 - \delta) \limsup_t (\nu_0 \tilde{P}^{t-m})(x)$$

$$\leq \frac{\delta}{M} + (1 - \delta) \limsup_t \tilde{P}^t(x, x).$$

Rearranging gives $\limsup_t \tilde{P}^t(x, x) \leq 1/M$. Since $M$ is arbitrary, this concludes the proof. □

**Definitions** Let $(X_t)$ be an irreducible Markov chain on a countable state space $V$ with transition matrix $P$ and stationary measure $\pi > 0$. As we did in the finite case, we let $(Pf)(x) := \sum_y P(x,y)f(y)$ (which is well defined when $\sum_y P(x,y)|f(y)| < +\infty$ for all $x$). Let $\ell^2(V)$ be the set of real-valued functions on $V$ with finite support and let $\ell^2(V, \pi)$ be the Hilbert space of real-valued functions $f$ with $\|f\|^2_\pi := \sum_x \pi(x)f(x)^2 < +\infty$ equipped with the inner product

$$\langle f, g \rangle_\pi := \sum_{x \in V} \pi(x)f(x)g(x).$$

Then $P$ maps $\ell^2(V, \pi)$ to itself. In fact, we have the stronger statement:

**Lemma 5.26.** For any $f \in \ell^2(V, \pi)$, $Pf$ is well-defined and further we have

$$\|Pf\|_\pi \leq \|f\|_\pi.$$
Proof. Note that by Cauchy-Schwarz, Fubini and stationarity

\[
\sum_x \pi(x) \left[ \sum_y P(x, y) |f(y)| \right]^2 \leq \sum_x \pi(x) \sum_y P(x, y) f(y)^2 = \sum_y \sum_x \pi(x) P(x, y) f(y)^2 = \sum_y \pi(y) f(y)^2 = \|f\|_\pi^2 < +\infty.
\]

This shows that \( Pf \) is well-defined since \( \pi > 0 \). Applying the same argument to \( \|P\|_\pi \) gives the inequality.

We consider the operator norm

\[
\|P\|_\pi = \sup \left\{ \frac{\|Pf\|_\pi}{\|f\|_\pi} : f \in \ell^2(V, \pi), f \neq 0 \right\},
\]

and note that by the lemma \( \|P\|_\pi \leq 1 \). Note that, if \( V \) is finite or more generally if \( \pi \) is summable, then we have \( \|P\|_\pi = 1 \) since we can take \( f \equiv 1 \) above in that case.

**Lemma 5.27.** If in addition \( P \) is reversible with respect to \( \pi > 0 \), then \( P \) is self-adjoint on \( \ell^2(V, \pi) \), that is,

\[
\langle f, Pg \rangle_\pi = \langle Pf, g \rangle_\pi \quad \forall f, g \in \ell^2(V, \pi).
\]

**Proof.** First consider \( f, g \in \ell_0(V) \). Then by reversibility

\[
\langle f, Pg \rangle_\pi = \sum_{x, y} \pi(x) P(x, y) f(x) g(y) = \sum_{x, y} \pi(y) P(y, x) f(x) g(y) = \langle Pf, g \rangle_\pi.
\]

Because \( \ell^0(V) \) is dense in \( \ell^2(V, \pi) \) (just truncate) and the bilinear form above is continuous in \( f \) and \( g \) (because \( |\langle f, Pg \rangle_\pi| \leq \|P\|_\pi \|f\|_\pi \|g\|_\pi \) by Cauchy-Schwarz and the definition of the operator norm) the result follows for \( f, g \in \ell^2(V, \pi) \).

**Rayleigh quotient** For a reversible \( P \), we have the following characterization of the operator norm in terms of the so-called Rayleigh quotient.

**Theorem 5.28 (Rayleigh quotient).** Let \( P \) be irreducible and reversible with respect to \( \pi > 0 \). Then

\[
\|P\|_\pi = \sup \left\{ \frac{\langle f, Pf \rangle_\pi}{\langle f, f \rangle_\pi} : f \in \ell_0(V), f \neq 0 \right\}.
\]
Proof. Let $\lambda_1$ be the r.h.s. above. By Cauchy-Schwarz $|\langle f, Pf \rangle_\pi| \leq \|f\|_\pi \|Pf\|_\pi$. That gives $\lambda_1 \leq \|P\|_\pi$ by dividing both sides by $\|f\|_\pi^2$.

In the other direction, note that for a self-adjoint operator $P$ we have the following “polarization identity”

$$\langle Pf, g \rangle_\pi = \frac{1}{4} [(\langle f + g, f + g \rangle_\pi - \langle f - g, f - g \rangle_\pi)]$$

which can be checked by expanding the r.h.s. Note that if $\langle f, Pf \rangle_\pi \leq \lambda_1 \langle f, f \rangle_\pi$ for all $f \in \ell_0(V)$ then the same holds for all $f \in \ell^2(V, \pi)$ because $\ell_0(V)$ is dense in $\ell^2(V, \pi)$. So for any $f, g \in \ell^2(V, \pi)$

$$|\langle Pf, g \rangle_\pi| \leq \frac{\lambda_1}{4} [\langle f + g, f + g \rangle_\pi + \langle f - g, f - g \rangle_\pi] = \lambda_1 \frac{\langle f, f \rangle_\pi + \langle g, g \rangle_\pi}{2}.$$ 

Taking $g := Pf \|f\|_\pi / \|Pf\|_\pi$ gives

$$\|Pf\|_\pi \|f\|_\pi \leq \lambda_1 \|f\|_\pi^2,$$

or $\|P\|_\pi \leq \lambda_1$.

### Spectral radius

We define the notion of spectral radius.

**Definition 5.29** (Spectral radius). Let $P$ be irreducible. The spectral radius of $P$ is defined as

$$\rho(P) := \limsup_t P^t(x, y)^{1/t},$$

which does not depend on $x, y$.

To see that the lim sup does not depend on $x, y$, let $u, v, x, y \in V$ and $k, m \geq 0$ such that $P^m(u, x) > 0$ and $P^k(y, v)$. Then

$$P^{t+m+k}(u, v) \leq (P^m(u, x)P^t(x, y)P^k(y, v))^{1/(t+m+k)}$$

$$\geq P^m(u, x)^{1/(t+m+k)} P^t(x, y)^{1/t} P^k(y, v)^{1/(t+m+k)},$$

which shows that $\limsup_t P^t(u, v)^{1/t} \geq \limsup_t P^t(x, y)^{1/t}$ for all $u, v, x, y$.

In the positive recurrent case (for instance if the chain is finite), we have $P^t(x, y) \to \pi(y) > 0$ and so $\rho(P) = 1 = \|P\|_\pi$. The equality between $\rho(P)$ and $\|P\|_\pi$ holds in general for reversible chains.
**Theorem 5.30** (Spectral radius and norm). Let $P$ be irreducible and reversible with respect to $\pi > 0$. Then

$$\rho(P) = \|P\|_\pi.$$  

Moreover, for all $t$,

$$P^t(x, y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}} \|P\|^t_\pi.$$  

**Proof.** Because $P$ is self-adjoint and $\|\delta_x\|^2_\pi = \pi(z) \leq 1$, by Cauchy-Schwarz

$$\pi(x)P^t(x, y) = \langle \delta_x, P^t \delta_y \rangle_\pi \leq \|P\|^t_\pi \|\delta_x\|_\pi \|\delta_y\|_\pi = \|P\|^t_\pi \sqrt{\pi(x)\pi(y)}.$$  

Hence $P^t(x, y) \leq \sqrt{\frac{\pi(y)}{\pi(x)}} \|P\|^t_\pi$ and further $\rho(P) \leq \|P\|_\pi$.

For the other direction, by self-adjointness and Cauchy-Schwarz, for any $f \in \ell^2(V, \pi)$

$$\|P^{t+1}f\|^2_\pi = \langle P^{t+1}f, P^{t+1}f \rangle_\pi = \langle P^{t+2}f, P^t f \rangle_\pi \leq \|P^{t+2}f\|_\pi \|P^t f\|_\pi,$$

or

$$\frac{\|P^{t+1}f\|_\pi}{\|P^t f\|_\pi} \leq \frac{\|P^{t+2}f\|_\pi}{\|P^{t+1}f\|_\pi}.$$  

So $\|P^{t+1}f\|_\pi/\|P^t f\|_\pi$ is non-decreasing and therefore has a limit $L \leq +\infty$. Moreover $\|Pf\|_\pi/\|f\|_\pi \leq L$ so it suffices to prove $L \leq \rho(P)$. As before it suffices to prove this for $f \in \ell^2_0(V), f \neq 0$ by a density argument.

Observe that

$$\left(\frac{\|P^t f\|_\pi}{\|f\|_\pi}\right)^{1/t} = \left(\frac{\|P f\|_\pi}{\|f\|_\pi} \times \cdots \times \frac{\|P^t f\|_\pi}{\|P^{t-1} f\|_\pi}\right)^{1/t} \rightarrow L,$$

so $L = \lim_t \|P^t f\|_\pi^{1/t}$. By self-adjointness again

$$\|P^t f\|^2_\pi = \langle f, P^{2t} f \rangle_\pi = \sum_{x,y} \pi(x) f(x) f(y) P^{2t}(x, y).$$

By definition of $\rho := \rho(P)$, for any $\varepsilon > 0$, there is $t$ large enough so that $P^{2t}(x, y) \leq (\rho + \varepsilon)^{2t}$ for all $x, y$ in the support of $f$. In that case,

$$\|P^t f\|_\pi^{1/t} \leq (\rho + \varepsilon) \left(\sum_{x,y} \pi(x)|f(x) f(y)|\right)^{1/2t}.$$  

287
The sum on the l.h.s. is finite because $f$ has finite support. Since $\varepsilon$ is arbitrary, we get
\[
\limsup_t \|P^t f\|_{\pi}^{1/t} \leq \rho.
\]

**Corollary 5.31.** Let $P$ be irreducible and reversible with respect to $\pi$. If $\|P\|_{\pi} < 1$, then $P$ is transient.

**Proof.** By Theorem 5.30, $P^t(x,x) \leq \|P\|_{\pi}^t$ so $\sum_t P^t(x,x) < +\infty$. Because $\sum_t P_t(x,x) = \mathbb{E}_x[\sum_t 1_{\{X_t=x\}}]$, we have that $\sum_t 1_{\{X_t=x\}} < +\infty$, $P_x$-a.s., and $(X_t)$ is transient.

This is not an if and only if. Random walk on $\mathbb{Z}^3$ is transient, yet $P^{2t}(0,0) = \Theta(t^{-3/2})$ so $\|P\|_{\pi} = \rho(P) = 1$.

In the non-reversible case, the result in Theorem 5.30 generally does not hold.

**Example 5.32** (Counter-example). Consider asymmetric random walk on $\mathbb{Z}$ with probability $p \in (1/2, 1)$ of going to the right. Then both $\pi_0(x) := \left(\frac{p}{1-p}\right)^x$ and $\pi_1(x) := 1$ define stationary measures, but only $\pi_0$ is reversible. Under $\pi_1$, we have $\|P\|_{\pi_1} = 1$. Indeed, let $f_n(x) := 1_{\{|x|\leq n\}}$ and note that
\[
(Pf_n)(x) = 1_{\{|x|\leq n-1\}} + p1_{\{x = -n - 1 \text{ or } -n\}} + (1 - p)1_{\{x = n \text{ or } n + 1\}},
\]
so $\|f_n\|_{\pi_1}^2 = 2n + 1$ and $\|Pf_n\|_{\pi_1}^2 \geq 2(n - 1) + 1$. Hence
\[
\limsup_n \frac{\|Pf_n\|_{\pi_1}}{\|f_n\|_{\pi_1}} \geq 1.
\]

On the other hand, $\mathbb{E}_0[X_t] = (2p - 1)t$ and $X_t$, as a sum of $t$ independent increments in $\{-1, +1\}$, is a 2-Lipschitz function. So, by the Azuma-Hoeffding inequality (Theorem 3.52),
\[
P^t(0,0)^{1/t} \leq \mathbb{P}_0[|X_t| \leq 0]^{1/t} = \mathbb{P}_0[|X_t - (2p - 1)t| \leq -(2p - 1)t]^{1/t} \leq e^{-\frac{(2p-1)^2t^2}{2t^2}}.
\]

Therefore $\rho(P) \leq e^{-\frac{(2p-1)^2}{2}} < 1$. \hfill \blacktriangleleft
The goal of this section is to relate the spectral gap and the spectral radius to certain geometric properties of the underlying network. More specifically, we will encounter isoperimetric properties of the network, that is, relationships between the “volume” of sets and their “circumference.” The classical *isoperimetric inequality* states that the area enclosed by any (rectifiable) closed curve in the plane is at most the length of the curve squared divided by $4\pi$. Moreover equality is achieved if and only if the curve is a circle. We will derive similar inequalities for networks which will involve the eigenvalues of the corresponding random walk.

**Some notation**  We will need the following definitions. Let $\mathcal{N} = (G, c)$ be a finite or infinite network with $G = (V, E)$. For a subset $S \subseteq V$, we let the edge boundary of $S$ be

$$\partial_E S := \{ e = (x, y) \in E : x \in S, y \in S^c \},$$

and the vertex boundary of $S$ be

$$\partial_V S := \{ y \in S^c : \exists x \in S \text{ s.t. } x \sim y \}.$$

Let $g : E \to \mathbb{R}_+$ and $h : V \to \mathbb{R}_+$ be edge and vertex weight functions. For $F \subseteq E$ and $W \subseteq V$ we define

$$|F|_g := \sum_{e \in F} g(e),$$

and

$$|W|_h := \sum_{v \in W} h(v).$$

For $S \subseteq V$, we let

$$\Phi_E(S; g, h) := \frac{|\partial_E S|_g}{|S|_h},$$

and

$$\Phi_V(S; g, h) := \frac{|\partial_V S|_h}{|S|_h}.$$

### 5.2.1 Bottleneck ratio

Let $(X_t)_{t \geq 0}$ be a finite, irreducible Markov chain on $V$ reversible with respect to its stationary distribution $\pi > 0$. We think of $(X_t)_{t \geq 0}$ as a random walk on the
network $\mathcal{N} = (G, c)$ where $G$ is the transition graph and $c(x, y) := \pi(x)P(x, y) = \pi(y)P(y, x)$. For disjoint subsets $S_0, S_1 \subseteq V$, we let

$$c(S_0, S_1) := \sum_{x_0 \in S_0} \sum_{x_1 \in S_1} c(x_0, x_1).$$

**Definition 5.33** (Bottleneck ratio). For a subset of states $S \subseteq V$, the bottleneck ratio of $S$ is

$$\Phi_E(S; c, \pi) = \frac{\partial_E S}{|S|_\pi} = \frac{c(S, S^c)}{\pi(S)}.$$  

The bottleneck ratio (or Cheeger isoperimetric constant) of $\mathcal{N}$ is

$$\Phi_* := \min \left\{ \Phi_E(S; c, \pi) : S \subseteq V, 0 < \pi(S) \leq \frac{1}{2} \right\}.$$

Equivalently,

$$\Phi_* := \min \left\{ \frac{c(S, S^c)}{\min \{\pi(S), \pi(S^c)\}} : S \subseteq V, 0 < \pi(S) < 1 \right\}.$$

**Example 5.34** (Bottleneck ratio: complete graph). Let $G = K_n$ be the complete graph on $n$ vertices and assume $c(x, y) = 1$ for all $x \neq y$. For simplicity, take $n$ even. Then for a subset $S$ of size $|S| = k$,

$$\Phi_E(S; c, \pi) = \frac{\partial_E S}{|S|_\pi} = \frac{k(n-k)}{k/n} = \frac{n-k}{n}.$$  

Thus, the minimum is achieved for $k = n/2$ and

$$\Phi_* = \frac{n-n/2}{n} = \frac{1}{2}.$$  

The spectral gap and the bottleneck ratio are related through the following isoperimetric inequalities. The lower bound is known as Cheeger’s inequality.

**Theorem 5.35** (Spectral gap and the bottleneck ratio). Let $P$ be a finite, irreducible, reversible Markov transition matrix and let $\gamma = 1 - \lambda_2$ be the spectral gap of $P$. Then

$$\frac{\Phi_*^2}{2} \leq \gamma \leq 2\Phi_*.$$  

In terms of the relaxation time $t_{rel} = \gamma^{-1}$, these inequalities have an intuitive meaning: the presence or absence of a strong bottleneck in the state space leads to slow or fast mixing respectively. See Figure 5.1. Before giving a proof of the theorem, we start with a trivial example.
Example 5.36 (Two-state chain). Let $V := \{0, 1\}$ and, for $\alpha, \beta \in (0, 1)$,

$$P := \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

which has stationary distribution

$$\pi := \left( \frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta} \right).$$

Recall that the second right eigenvector is

$$f_2 := \left( \sqrt{\frac{\alpha}{\beta}}, \sqrt{\frac{\beta}{\alpha}} \right)' = \left( \sqrt{\frac{\pi_1}{\pi_0}}, \sqrt{\frac{\pi_0}{\pi_1}} \right)' ,$$

with eigenvalue $\lambda_2 := 1 - \alpha - \beta$, so the spectral gap is $\alpha + \beta$. Assume that $\beta \leq \alpha$.

Then the bottleneck ratio is

$$\Phi_* = \frac{c(0,1)}{\pi(0)} = P(0,1) = \alpha.$$

Then Theorem 5.35 reads

$$\frac{\alpha^2}{2} \leq \alpha + \beta \leq 2\alpha,$$
which is indeed satisfied for all $0 < \beta \leq \alpha < 1$. Note that the upper bound is tight when $\alpha = \beta$.

**Proof.** We start with the upper bound. Thanks to the variational characterization of $\lambda_2$, to get an upper bound on the spectral gap, it suffices to plug in a well-chosen function in Rayleigh’s quotient. Taking a hint from Example 5.36, for $S \subseteq V$ with $\pi(S) \in (0, 1/2]$, we let

$$f_S(x) := \begin{cases} -\sqrt{\frac{\pi(S)}{\pi(S^c)}}, & x \in S, \\ \sqrt{\frac{\pi(S)}{\pi(S^c)}}, & x \in S^c. \end{cases}$$

Then

$$\sum_x \pi(x)f_S(x) = \pi(S) \left[ -\sqrt{\frac{\pi(S^c)}{\pi(S)}} + \pi(S^c) \sqrt{\frac{\pi(S)}{\pi(S^c)}} \right] = 0,$$

and

$$\sum_x \pi(x)f_S(x)^2 = \pi(S) \left[ -\frac{\pi(S^c)}{\pi(S)} \right]^2 + \pi(S^c) \left[ \frac{\pi(S)}{\pi(S^c)} \right]^2 = 1.$$  

From the Poincaré inequality for $\mathcal{N}$ (Lemma A.9),

$$\gamma \leq \frac{\mathcal{E}(f_S)}{\text{Var}_\pi[f_S]} = \mathcal{E}(f_S) = \frac{c(S, S^c)}{\pi(S)\pi(S^c)} \leq 2 \frac{c(S, S^c)}{\pi(S)},$$

as claimed.

The other direction is somewhat trickier. Because we seek an upper bound on the bottleneck ratio $\Phi_*$, our goal is to find a cut $(S, S^c)$ such that

$$\frac{c(S, S^c)}{\pi(S) \land \pi(S^c)} \leq \sqrt{2\gamma}.$$  

We use a probabilistic argument, that is, we construct a random cut $(Z, Z^c)$. Observe that it suffices that

$$\mathbb{E}[c(Z, Z^c)] \leq \sqrt{2\gamma} \mathbb{E} \left[ \pi(Z) \land \pi(Z^c) \right],$$

since that implies that $\mathbb{E} \left[ \sqrt{2\gamma} \pi(Z) \land \pi(Z^c) - c(Z, Z^c) \right] \geq 0$ which in turn implies that $\mathbb{P} \left[ \sqrt{2\gamma} \pi(Z) \land \pi(Z^c) - c(Z, Z^c) > 0 \right] > 0$. We now describe the random cut $(Z, Z^c):$
1. (Cuts from $f_2$) Let $f_2$ be an eigenfunction corresponding to the eigenvalue $\lambda_2$ with $\|f_2\|^2_\pi = 1$. Order the vertices $V := \{v_1, \ldots, v_n\}$ in such a way that

$$f_2(v_i) \leq f_2(v_{i+1}), \quad \forall i = 1, \ldots, n - 1.$$ 

The function $f_2$ naturally produces a series of cuts $(S_i, S_i^c)$ where $S_i := \{v_1, \ldots, v_i\}$. By definition of the bottleneck ratio,

$$\Phi_* \leq \frac{c(S_i, S_i^c)}{\pi(S_i) \wedge \pi(S_i^c)}. \quad (5.2)$$

2. (Normalization) Let

$$m := \min\{i : \pi(S_i) > 1/2\},$$

and define

$$f := f_2 - f_2(v_m).$$

We further define $g := \alpha f$ where $\alpha > 0$ is chosen so that

$$g(v_1)^2 + g(v_n)^2 = 1.$$

Note that $g(v_m) = 0$ and $g(v_1) \leq \cdots g(v_m) = 0 \leq g(v_{m+1}) \leq \cdots g(v_n)$. Further, the function $g$ satisfies:

Lemma 5.37.

$$\frac{1}{2} \sum_{x,y} c(x, y)(g(x) - g(y))^2 \leq \gamma \sum_x \pi(x)g(x)^2.$$

Proof. By Lemma A.9, $\gamma = \frac{\mathcal{E}(f_2)}{\Var \pi[f_2]}$. Because neither the numerator nor the denominator are affected by adding a constant, we have $\gamma = \frac{\mathcal{E}(f)}{\Var \pi[f]}$. Also a constant multiplying $f$ cancels out in the ratio so $\gamma = \frac{\mathcal{E}(g)}{\Var \pi[g]}$. Now use the fact that $\Var \pi[g] \leq \sum_x \pi(x)g(x)^2$. \hfill $\blacksquare$

3. (Random cut) Pick $\Theta$ in $[g(v_1), g(v_n)]$ with density $2|\theta|$. Note that

$$\int_{g(v_1)}^{g(v_n)} 2|\theta| \, d\theta = g(v_1)^2 + g(v_n)^2 = 1.$$

Finally define

$$Z := \{v_i : g(v_i) < \Theta\}.$$
We compute the expectations on both sides of (5.1). Because of our centering, \( \Theta \leq 0 \) implies that \( \pi(Z) \land \pi(Z^c) = \pi(Z) \) and vice versa. Thus

\[
\mathbb{E}[\pi(Z) \land \pi(Z^c)] = \mathbb{E} \left[ \sum_{\ell < m} \pi(v_\ell) 1_{\{v_\ell \in Z\}} 1_{\{\Theta \leq 0\}} + \sum_{\ell \geq m} \pi(v_\ell) 1_{\{v_\ell \in Z^c\}} 1_{\{\Theta > 0\}} \right]
\]

\[
= \mathbb{E} \left[ \sum_{\ell < m} \pi(v_\ell) 1_{\{g(v_\ell) \leq 0\}} + \sum_{\ell \geq m} \pi(v_\ell) 1_{\{0 < \Theta \leq g(v_\ell)\}} \right]
\]

\[
= \sum_{\ell < m} \pi(v_\ell) \mathbb{P}[g(v_\ell) < \Theta \leq 0] + \sum_{\ell \geq m} \pi(v_\ell) \mathbb{P}[0 < \Theta \leq g(v_\ell)]
\]

\[
= \sum_{\ell < m} \pi(v_\ell) g(v_\ell)^2 + \sum_{\ell \geq m} \pi(v_\ell) g(v_\ell)^2
\]

\[
= \sum_x \pi(x) g(x)^2,
\] (5.3)

where we integrated over the density of \( \Theta \) to obtain the fourth line.

To compute \( \mathbb{E}[c(Z, Z^c)] \), we note that \( x_k \in Z \) and \( x_\ell \in Z^c \) if and only if \( g(v_k) < \Theta \leq g(v_\ell) \). The probability of that event depends on the signs of \( g(v_k) \) and \( g(v_\ell) \). If \( g(v_k) g(v_\ell) \geq 0 \),

\[
\mathbb{P}[g(v_k) < \Theta \leq g(v_\ell)] = |g(v_k)^2 - g(v_\ell)^2|
\]

\[
= |g(v_k) - g(v_\ell)||g(v_k) + g(v_\ell)|
\]

\[
= |g(v_k) - g(v_\ell)|( |g(v_k)| + |g(v_\ell)|).
\]

If \( g(v_k) g(v_\ell) < 0 \),

\[
\mathbb{P}[g(v_k) < \Theta \leq g(v_\ell)] = g(v_k)^2 + g(v_\ell)^2
\]

\[
\leq g(v_k)^2 + g(v_\ell)^2 - 2g(v_k)g(v_\ell)
\]

\[
= (g(v_k) - g(v_\ell))^2
\]

\[
= |g(v_k) - g(v_\ell)|( |g(v_k)| + |g(v_\ell)|).
\]

We apply Cauchy-Schwarz to get

\[
\mathbb{E}[c(Z, Z^c)] \leq \sum_{k < \ell} c(v_k, v_\ell) |g(v_k) - g(v_\ell)|( |g(v_k)| + |g(v_\ell)|)
\]

\[
\leq \left( \sum_{k < \ell} c(v_k, v_\ell)(g(v_k) - g(v_\ell))^2 \right)^{1/2}
\]

\[
\times \left( \sum_{k < \ell} c(v_k, v_\ell)( |g(v_k)| + |g(v_\ell)|)^2 \right)^{1/2}.
\]

294
To bound the l.h.s., we use the following lemma.

**Lemma 5.38.**

\[
\sum_{k<\ell} c(v_k, v_\ell) (|g(v_k)| + |g(v_\ell)|)^2 \leq 2 \sum_x \pi(x) g(x)^2.
\]

**Proof.** Note that

\[
(|g(x)| + |g(y)|)^2 = 2g(x)^2 + 2g(y)^2 - (|g(x)| - |g(y)|)^2 \leq 2g(x)^2 + 2g(y)^2.
\]

Therefore, since \(\sum_y c(x, y) = \sum_y c(y, x) = \pi(x)\),

\[
\sum_{k<\ell} c(v_k, v_\ell) (|g(v_k)| + |g(v_\ell)|)^2 \leq \frac{1}{2} \sum_{x,y} c(x, y) (|g(x)| + |g(y)|)^2
\]

\[
= \sum_x \pi(x) g(x)^2 + \sum_y \pi(y) g(y)^2.
\]

Combining (5.3) and Lemmas 5.37 and 5.38, we get finally

\[
\mathbb{E}[c(Z, Z^c)] \leq \left(\gamma \sum_x \pi(x) g(x)^2\right)^{1/2} \left(2 \sum_x \pi(x) g(x)^2\right)^{1/2}
\]

\[
= \sqrt{2\gamma \mathbb{E}[\pi(Z) \wedge \pi(Z^c)]}.
\]

That concludes the proof.

**Expander graphs** In many applications, it is useful to construct “bottleneck-free” graphs. In particular, random walks mix rapidly on such graphs. Formally:

**Definition 5.39 (Expander family).** Let \(\{G_n\}_n\) be a collection of finite \(d\)-regular graphs with \(\lim_n |V(G_n)| = +\infty\), where \(V(G_n)\) is the vertex set of \(G_n\). Let

\[
\Phi_s(G_n) := \min \left\{ \frac{|\partial_E S|}{d|S|} : S \subseteq V(G_n), 0 < |S| \leq \frac{|V(G_n)|}{2} \right\}
\]

denote the bottleneck ratio of \(G_n\) with unit conductances. Let \(\alpha > 0\). We say that \(\{G_n\}_n\) is a \((d, \alpha)\)-expander family if for all \(n\)

\[
\Phi_s(G_n) \geq \alpha.
\]
The point of the definition is that the bottleneck ratio of all graphs in an expander family is bounded away from 0 uniformly in $n$. Note that it is trivial to construct such a family if we drop the bounded degree assumption: the bottleneck ratio of the complete graph $K_n$ is $1/2$ by Example 5.34. On the other hand, it is far from obvious that one can construct a family of sparse graphs (i.e., such that $|E(G_n)| = O(|V(G_n)|)$) with a bottleneck ratio uniformly bounded away from 0. It turns out that a simple probabilistic construction does the trick.

Existence of expander graphs For simplicity, we allow multigraphs and consider the case $d = 3$. We construct a random bipartite multigraph $G_n = (L_n, R_n, E_n)$ on $2n$ vertices known as Pinsker’s model. Denote the vertices by $L_n = \{\ell_1, \ldots, \ell_n\}$ and $R_n = \{r_1, \ldots, r_n\}$. Let $\sigma_1^n$ and $\sigma_2^n$ be independent uniform random permutations of $[n]$. The edge set of $G_n$ is given by

$$E_n := \{(\ell_i, r_i) : i \in [n]\} \cup \{(\ell_i, r_{\sigma_1^n(i)}) : i \in [n]\} \cup \{(\ell_i, r_{\sigma_2^n(i)}) : i \in [n]\}.$$ 

In words, $G_n$ is a union of three independent uniform perfect matchings (and its vertices are labeled so that one of the matchings is $\{(\ell_i, r_i)\}_i$). See Figure 5.2. Observe that, as a multigraph, all vertices of $G_n$ have degree 3. We show that there exists $\alpha > 0$ such that, for all $n$ large enough, with positive probability $G_n$ has a bottleneck ratio bounded below by $\alpha$. In particular, such a $G_n$ exists for all $n$. 

Figure 5.2: An instance of Pinsker’s model.
large enough and, thus, there exists a $(3, \alpha)$-expander family. In fact, we prove something stronger: with high probability, $\Phi_*(G_n) \geq \alpha$.

**Claim 5.40** (Pinsker’s model: bottleneck ratio). There exists $\alpha > 0$ such that

$$\lim_n \mathbb{P}[\Phi_*(G_n) \geq \alpha] = 1.$$  

**Proof.** For convenience, assume $n$ is even. We need to show that, for any $S$ with $|S| \leq |V(G_n)|/2 = n$, we have $|\partial E S| \geq \alpha d|S|$ for some $\alpha > 0$. We first reduce the proof to a statement about sets of vertices lying on one side of $G_n$.

**Lemma 5.41**. There is $\beta > 0$ such that

$$\lim_n \mathbb{P}\left[|\partial V K| \geq (1 + \beta)|K|, \forall K \subseteq L, |K| \leq n/2\right] = 1.$$  

Note that the lemma concerns the vertex boundary of $K$. Observe moreover that by construction we always have $|\partial V K| \geq |K|$. Before proving the lemma, we argue that it implies Claim 5.40. Assume the lemma holds. Let $S$ with $|S| \leq n$, and let $S_L := S \cap L$ and $S_R := S \cap R$. Assume w.l.o.g. that $|S_L| \geq |S_R|$. Let $K$ be any subset of $S_L$ with $|K| = |S_L| \wedge (n/2) \geq |S_R|$. In particular, $|K| \geq |S|/2$ because $|S_L| \geq |S|/2$ and $|S| \leq n$. Assume the event in the lemma holds. Then $|\partial V K| \geq (1 + \beta)|K|$. Of course some of the vertices in $\partial V K$ may in fact be in $S_R$. But $|S_R| \leq |K|$ so

$$|\partial E S| \geq |\partial V K| - |S_R| \geq \beta|K| \geq \frac{\beta}{2}|S|.$$  

That proves the claim with $\alpha = \beta/2$.

It remains to prove the lemma.

**Proof of Lemma 5.41.** Let $K \subseteq L$ with $|K| \leq n/2$. W.l.o.g. assume $K = \{\ell_1, \ldots, \ell_k\}$. Observe that $\partial V K \supseteq K'$ where $K' = \{r_1, \ldots, r_k\}$. We bound the probability that $|\partial V K| \leq k + \lfloor \beta k \rfloor$ by taking a union bound over all subsets of $\{r_{k+1}, \ldots, r_n\}$ of size $\lfloor \beta k \rfloor$. See Figure 5.3. Since $\sigma_n^1$ and $\sigma_n^2$ are uniform and independent, we have

$$\mathbb{P}[|\partial V K| \leq k + \lfloor \beta k \rfloor] \leq \binom{n-k}{\lfloor \beta k \rfloor} \left(\frac{\lfloor \beta k \rfloor}{\binom{n}{k}}\right)^2 \leq \binom{n}{\lfloor \beta k \rfloor} \left(\frac{\lfloor \beta k \rfloor}{\binom{n}{k}}\right)^2,$$  

297
where we used that $\binom{n}{s} = \binom{n}{n-s}$. Again, by a union bound,

\[ P[\exists K \subseteq L, |K| \leq n/2, |\partial \gamma K| \leq (1 + \beta)|K|] \leq \sum_{k=1}^{n/2} \binom{n}{k} \left( \binom{n}{[\beta k]} \right)^2 \frac{(k+\beta k)}{(n\beta k)^2}. \]

We use the bound $\frac{n^s}{s^s} \leq \binom{n}{s} \leq \frac{e^{n^s}}{s^s}$ for $s \leq t < n$ (indeed $(\log(\frac{e^t}{t}))' = \log(\frac{e^t}{t}) > 0$ for $0 < t < n$) to obtain

\[ \sum_{k=1}^{n/2} \binom{n}{k} \left( \binom{n}{[\beta k]} \right)^2 \leq \sum_{k=1}^{n/2} e^{\beta k} \frac{1}{\beta k} \frac{(e^{\beta k} (1 + \beta)^{\beta k})^2}{n^{\beta k}} \leq \sum_{k=1}^{n/2} \left( \frac{k}{n} \right)^{k(1-\beta)} \left( \frac{e^{3(1+\beta)^2}}{\beta^3} \right)^{\beta k}. \quad (5.4) \]

Note that $\frac{k}{n} \leq \frac{1}{2}$ for $k \leq \frac{n}{2}$ and

\[ \gamma_{\beta} := \left( \frac{1}{2} \right)^{1-\beta} \left( \frac{e^{3(1+\beta)^2}}{\beta^3} \right)^{\beta} < 1, \]

for $\beta$ small enough since $\beta^\beta \to 1$ as $\beta \to 0$. Thus the $k$-th term in the sum (5.4) goes to 0 as $n \to +\infty$ and is, in fact, bounded by $\gamma_{\beta}^k$ which is summable. So, by
dominated convergence, the sum itself tends to 0 as \( n \to +\infty \). That concludes the proof.

Claim 5.40 implies:

**Theorem 5.42** (Existence of expander family). For \( \alpha > 0 \) small enough, there exists a \((3, \alpha)\)-expander (multigraph) family.

**Proof.** By Claim 5.40, for all \( n \) large enough, there exists \( G_n \) with \( \Phi_\ast(G_n) \geq \alpha \) for some fixed \( \alpha > 0 \).

**Fast mixing on expander graphs**  As we mentioned above, an important property of an expander graph is that random walk on such a graph mixes rapidly. We make this precise.

**Claim 5.43** (Mixing on expanders). Let \( \{G_n\} \) be a \((d, \alpha)\)-expander family. Then \( t_{\text{mix}}(\varepsilon) = \Theta(\log |V(G_n)|) \), where the constant depends on \( \varepsilon \) and \( \alpha \).

**Proof.** Because of the degree assumption, random walk on \( G_n \) is reversible with respect to the uniform distribution. So the mixing time is upper bounded by

\[
t_{\text{mix}}(\varepsilon) \leq \log \left( \frac{1}{\varepsilon \pi_{\text{min}}} \right) t_{\text{rel}} \leq \log \left( \frac{|V(G_n)|}{\varepsilon} \right) \alpha^{-1} = O(\log |V(G_n)|).
\]

By the diameter-based lower bound on mixing for reversible chains, Claim 2.74, for \( n \) large enough

\[
t_{\text{mix}}(\varepsilon) \geq \frac{\Delta^2}{12 \log |V(G_n)| + 4 |\log \pi_{\text{min}}|},
\]

where \( \Delta \) is the diameter of \( G_n \). For a \( d \)-regular graph \( G_n \), the diameter is at least \( \log |V(G_n)| \). Indeed, by induction, the number of vertices within graph distance \( k \) of any vertex is at most \( d^k \). For \( d^k \) to be greater than \( |V(G_n)| \), we need \( k \geq \log_d |V(G_n)| \). Finally,

\[
t_{\text{mix}}(\varepsilon) \geq \frac{(\log_d |V(G_n)|)^2}{16 \log |V(G_n)|} = \Omega(\log |V(G_n)|).
\]

That concludes the proof.
5.2.2 Markov chains: random walk on trees, cycles, and hypercubes revisited

We use the techniques of the previous section to bound the mixing time of random walk on some simple graphs. In particular we revisit the examples of Section 4.4.2.

b-ary tree Let \((Z_t)\) be lazy simple random walk on the \(\ell\)-level rooted \(b\)-ary tree, \(\hat{T}_{b,\ell}\). The root, 0, is on level 0 and the leaves, \(L\), are on level \(\ell\). All vertices have degree \(b + 1\), except for the root which has degree \(b\) and the leaves which have degree 1. Hence the stationary distribution is

\[
\pi(x) := \frac{\delta(x)}{2(n - 1)},
\]

(5.5)

where \(n\) is the number of vertices and \(\delta(x)\) is the degree of \(x\).

It is intuitively clear that the bottleneck of this walk is at the root. Let \(x_0\) be a leaf of \(\hat{T}_{b,\ell}\) and let \(A\) be the set of vertices “on the other side of root (inclusively),” i.e., vertices whose graph distance from \(x_0\) is at least \(\ell\). See Figure 4.7. Let \(S\) be the remaining vertices. Then by symmetry \(\pi(S) \leq 1/2\). Note that there is a single edge connecting \(S\) and \(S^c = A\), namely, the edge linking the root of \(S\) and 0. More precisely, let \(v_S\) be the root of \(S\). From (5.5),

\[
P(v_S, 0) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6},
\]

\[
\pi(v_S) = \frac{3}{2n - 2},
\]

and, by symmetry,

\[
\pi(S) = \frac{(2n - 2 - 2)/2}{2n - 2} = \frac{n - 2}{2n - 2}.
\]

Hence

\[
\Phi_* \leq \frac{\frac{1}{6} \left(\frac{3}{2n - 2}\right)}{\frac{n - 2}{2n - 2}} = \frac{1}{2(n - 2)}.
\]

By Theorem 5.35,

\[
\gamma \leq 2\Phi_* = \frac{1}{n - 2} \quad \text{and} \quad t_{\text{rel}} = \gamma^{-1} \geq n - 2.
\]

Thus

\[
t_{\text{mix}}(\varepsilon) \geq (t_{\text{rel}} - 1) \log \left(\frac{1}{2\varepsilon}\right) = \Omega(n).
\]

We showed in Section 4.4.2 that \(t_{\text{mix}}(\varepsilon) = \Theta(n)\).
Cycle  Let \((Z_t)\) be lazy simple random walk on the cycle of size \(n\), \(\mathbb{Z}_n := \{0, 1, \ldots, n - 1\}\), where \(i \sim j\) if \(|j - i| = 1 \mod n\). Assume \(n\) is even.

Consider a subset of vertices \(S\). Note that by symmetry \(\pi(S) = \frac{|S|}{n}\). Moreover, for all \(i \sim j\), \(c(i, j) = \pi(i)P(i, j) = \frac{1}{n} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4n}\). Among all sets of size \(|S|\), consecutive vertices minimize the size of the boundary. So

\[
\Phi_* \leq \frac{2\frac{1}{\ell n}}{\ell n} = \frac{1}{2\ell},
\]

for all \(\ell \leq n/2\). This expression is minimized for \(\ell = n/2\) so

\[
\Phi_* = \frac{1}{n}.
\]

By Theorem 5.35,

\[
\frac{1}{2n^2} \leq \frac{\Phi_*^2}{2} \leq \gamma \leq 2\Phi_* = \frac{2}{n}
\]

and

\[
\frac{n}{2} \leq t_{\text{rel}} = \gamma^{-1} \leq 2n^2.
\]

Thus

\[
t_{\text{mix}}(\varepsilon) \geq (t_{\text{rel}} - 1) \log \left(\frac{1}{2\varepsilon}\right) = \Omega(n),
\]

and

\[
t_{\text{mix}}(\varepsilon) \leq \log \left(\frac{1}{\varepsilon \pi_{\text{min}}}\right) t_{\text{rel}} = O(n^2 \log n).
\]

We know from exact eigenvalue computations that in fact \(\gamma = \frac{2\pi^2}{n^2} + O(n^{-4})\).

We showed in Section 4.4.2 (and Exercise 4.9) that \(t_{\text{mix}}(\varepsilon) = \Theta(n^2)\).

Hypercube  Let \((Z_t)\) be lazy simple random walk on the \(n\)-dimensional hypercube \(\mathbb{Z}_2^n := \{0, 1\}^n\) where \(i \sim j\) if \(\|i - j\|_1 = 1\).

To get a bound on the bottleneck ratio, consider the set \(S = \{x \in \mathbb{Z}_2^n : x_1 = 0\}\). By symmetry \(\pi(S) = \frac{1}{2}\). For each \(i \sim j\), \(c(i, j) = \frac{1}{2n} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{n^2}\). Hence

\[
\Phi_* \leq \frac{2n - 1}{\frac{1}{2}} = \frac{1}{2n},
\]

and, by Theorem 5.35,

\[
\gamma \leq 2\Phi_* \leq \frac{1}{n}.
\]
Thus
\[ t_{\text{mix}}(\varepsilon) \geq (t_{\text{rel}} - 1) \log \left( \frac{1}{2\varepsilon} \right) = \Omega(n). \]

We know from exact eigenvalue computations that indeed \( \gamma = \frac{1}{n} \).

We showed in Section 4.4.2 that \( t_{\text{mix}}(\varepsilon) = \Theta(n \log n) \).

5.2.3 \( \triangleright \) Ising model: Glauber dynamics on the complete graph and expander graphs

Let \( G = (V, E) \) be a finite, connected graph with maximal degree \( \bar{\delta} \). Define \( \mathcal{X} := \{-1, +1\}^V \). Recall that the (ferromagnetic) Ising model on \( V \) with inverse temperature \( \beta \) is the probability distribution over spin configurations \( \sigma \in \mathcal{X} \) given by
\[
\mu_\beta(\sigma) := \frac{1}{Z(\beta)} e^{-\beta \mathcal{H}(\sigma)},
\]
where
\[
\mathcal{H}(\sigma) := -\sum_{i \sim j} \sigma_i \sigma_j,
\]
is the Hamiltonian and
\[
Z(\beta) := \sum_{\sigma \in \mathcal{X}} e^{-\beta \mathcal{H}(\sigma)},
\]
is the partition function. In this context, recall that vertices are often referred to as sites. The single-site Glauber dynamics of the Ising model is the Markov chain on \( \mathcal{X} \) which, at each time, selects a site \( i \in V \) uniformly at random and updates the spin \( \sigma_i \) according to \( \mu_\beta(\sigma) \) conditioned on agreeing with \( \sigma \) at all sites in \( V \setminus \{i\} \).

Specifically, for \( \gamma \in \{-1, +1\} \), \( i \in V \), and \( \sigma \in \mathcal{X} \), let \( \sigma^{i,\gamma} \) be the configuration \( \sigma \) with the state at \( i \) being set to \( \gamma \). Then, letting \( n = |V| \), the transition matrix of the Glauber dynamics is
\[
Q_\beta(\sigma, \sigma^{i,\gamma}) := \frac{1}{n} \left( \frac{e^{\gamma \beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}} \right) = \frac{1}{n} \left\{ \frac{1}{2} + \frac{1}{2} \tanh(\gamma \beta S_i(\sigma)) \right\}, \quad (5.6)
\]
where
\[
S_i(\sigma) := \sum_{j \sim i} \sigma_j.
\]

All other transitions have probability 0. Recall that this chain is irreducible and reversible with respect to \( \mu_\beta \). In particular \( \mu_\beta \) is the stationary distribution of \( Q_\beta \).

We showed in Claim 4.77 that the Glauber dynamics is fast mixing at high temperature. More precisely we proved that \( t_{\text{mix}}(\varepsilon) = O(n \log n) \) when \( \beta < \bar{\delta}^{-1} \). Here we prove a converse: at low temperature, graphs with good enough expansion properties produce exponentially slow mixing of the Glauber dynamics.
**Complete graph: the Curie-Weiss model**  
Let $G = K_n$ be the complete graph on $n$ vertices. In this case, the Ising model is often referred to as the *Curie-Weiss model*. It is customary to scale $\beta$ with $n$. We define $\alpha := \beta(n-1)$. Since $\bar{\delta} = n-1$, we have that, when $\alpha < 1$, $\beta = \frac{\alpha}{n-1} < \bar{\delta}^{-1}$ so $t_{\text{mix}}(\varepsilon) = O(n \log n)$. In the other direction, we prove:

**Claim 5.44** (Curie-Weiss model: exponentially slow mixing at low temperature). 
For $\alpha > 1$, $t_{\text{mix}}(\varepsilon) = \Omega(\exp(r(\alpha)n))$ for some function $r(\alpha) > 0$ not depending on $n$.

*Proof.* We first prove exponential mixing when $\alpha$ is large enough, an argument which will be useful in the generalization to expander graphs below.

The idea of the proof is to bound the bottleneck ratio and use Theorem 5.35. To simplify the proof, assume $n$ is odd. We denote the bottleneck ratio of the chain by $\Phi^X$ to avoid confusion with the underlying graph $G$. Intuitively, because the spins tend to align strongly at low temperature, it takes a considerable amount of time to travel from a configuration with a majority of $-1$s to a configuration with a majority of $+1$s. A natural place to look for a bottleneck is the set $S := \left\{ \sigma \in \mathcal{X} : \sum_i \sigma_i < 0 \right\}$, where the quantity $m(\sigma) := \sum_i \sigma_i$ is called the magnetization. Note that the magnetization is positive if and only if a majority of spins are $+1$ and that it forms a Markov chain by itself. Observe further that $\mu_\beta(S) = 1/2$. The bottleneck ratio is hence bounded by

$$\Phi^X \leq \sum_{\sigma \in S, \sigma' \notin S} \frac{\mu_\beta(\sigma)Q_\beta(\sigma, \sigma')}{\mu_\beta(S)} = 2 \sum_{\sigma \in S, \sigma' \notin S} \mu_\beta(\sigma)Q_\beta(\sigma, \sigma').$$  \hspace{1cm} (5.7)

Because the Glauber dynamics changes a single spin at a time, in order for $\sigma \in S$ to be adjacent to a configuration $\sigma' \notin S$, it must be that $$\sigma \in S_{-1} := \left\{ \sigma \in \mathcal{X} : m(\sigma) = -1 \right\},$$
and that $\sigma' = \sigma^{i,\pm}$ for some site $i$ such that $$i \in M_\sigma := \left\{ i \in V : \sigma_i = -1 \right\}.$$ 
Because the number of such sites is $(n+1)/2$ on $S_{-1}$, that is, $|M_\sigma| = (n+1)/2$ for all $\sigma \in S_{-1}$, and the Glauber dynamics picks a site uniformly at random, it
follows that for $\sigma \in S_{-1}$

$$
\sum_{\sigma' \in S} \mu_\beta(\sigma) Q_\beta(\sigma, \sigma') \leq \mu_\beta(\sigma) \left( \frac{n+1}{2} \right) = \frac{1}{2} \left( 1 + \frac{1}{n} \right) \mu_\beta(\sigma).
$$

Thus plugging this back in (5.7) gives

$$
\Phi^\mathcal{X}_* \leq \left( 1 + \frac{1}{n} \right) \mu_\beta(S_{-1})
$$

$$
= (1 + o(1)) \sum_{\sigma \in S_{-1}} \frac{e^{-\beta H(\sigma)}}{\mathcal{Z}(\beta)}
$$

$$
= (1 + o(1)) \sum_{\sigma \in S_{-1}} \exp\left( \frac{\alpha}{n-1} \left[ \left( \frac{|M wounds|}{2} \right) + \left( \frac{|M c\sigma|}{2} \right) - |M_\sigma||M_{\sigma}'| \right] \right) \mathcal{Z}(\beta).
$$

We bound the partition function $\mathcal{Z}(\beta) = \sum_{\sigma \in \mathcal{X}} e^{-\beta H(\sigma)}$ with the term for the all-$(-1)$ configuration, leading to

$$
\Phi^\mathcal{X}_* \leq (1 + o(1)) \sum_{\sigma \in S_{-1}} \exp\left( \frac{\alpha}{n-1} \left[ \left( \frac{|M wounds|}{2} \right) + \left( \frac{|M c\sigma|}{2} \right) - |M_\sigma||M_{\sigma}'| \right] \right)
$$

$$
= (1 + o(1)) \sum_{\sigma \in S_{-1}} \exp\left( -\frac{2\alpha}{n-1} |M_\sigma||M_{\sigma}'| \right)
$$

$$
= (1 + o(1)) \left( \frac{n}{n/2} \right) \exp\left( -\frac{2\alpha}{n-1} \left[ \frac{n+1}{2} \left[ \frac{n-1}{2} \right] \right] \right)
$$

$$
= (1 + o(1)) \sqrt{\frac{2}{\pi n}} 2^{n/2} (1 + o(1)) \exp\left( -\frac{\alpha(n+1)}{2} \right)
$$

$$
= C_\alpha \sqrt{\frac{2}{\pi n}} \exp\left( -n \left[ \frac{\alpha}{2} - \ln 2 \right] \right),
$$

for some constant $C_\alpha > 0$ depending on $\alpha$, where we used Stirling’s formula. Hence, by Theorem 5.35, for $\alpha > 2 \ln 2$ there is $r(\alpha) > 0$

$$
t_{\text{mix}}(\varepsilon) \geq (t_{\text{rel}} - 1) \log \left( \frac{1}{2 \varepsilon} \right) \geq \exp(r(\alpha)n) \log \left( \frac{1}{2 \varepsilon} \right).
$$

That proves the weaker result.
We now show that $\alpha > 1$ in fact suffices. For this, we need to improve our bound on the partition function in (5.10). Writing

$$Z(\beta) = \sum_{\sigma \in X} e^{-\beta H(\sigma)}$$

$$= \sum_{k=0}^{n} \binom{n}{k} \exp \left( \frac{\alpha n}{n-1} \left[ \frac{k}{2} + \binom{n-k}{2} - k(n-k) \right] \right)$$

$$= 2 \sum_{k=0}^{(n-1)/2} \binom{n}{k} \exp \left( \frac{\alpha n}{n-1} \left[ \frac{k}{2} + \binom{n-k}{2} - k(n-k) \right] \right)$$

$$= 2 \sum_{k=0}^{(n-1)/2} Y_{\alpha,k},$$

we see that the partition function is a sum of $O(n)$ exponentially large terms and is therefore dominated by the term corresponding to the largest exponent. Using Stirling’s formula,

$$\log \binom{n}{k} = (1 + o(1)) n H(k/n),$$

where $H(p) = -p \log p - (1-p) \log(1-p)$ is the entropy, and therefore

$$\log Y_{\alpha,k} = (1 + o(1)) n \left[ H(k/n) + \frac{\alpha (k/n)^2 + (1-k/n)^2 - 2(k/n)(1-k/n)}{2} \right].$$

where, for $p \in [0, 1]$, we let

$$J_\alpha(p) := H(p) + \frac{\alpha (1-2p)^2}{2}.$$

Note that the first term in $J_\alpha(p)$ is increasing on $[0, 1/2]$ while the second term is decreasing on $[0, 1/2]$. By a straightforward computation,

$$J_\alpha'(p) = \log \left( \frac{1-p}{p} \right) - 2\alpha (1-2p),$$

and

$$J_\alpha''(p) = -\frac{1}{p(1-p)} + 4\alpha.$$

Observe first that, when $\alpha < 1$, $J_\alpha'(1/2) = 0$ and $J_\alpha''(p) < 0$ for all $p \in [0, 1]$ since $p(1-p) \leq 1/4$. Hence, in that case, $J_\alpha$ is maximized at $p = 1/2$. In a sense, the
contribution from the entropy (i.e., how many ways are there to have \(k\) spins with value \(-1\) dwarfs that from the Hamiltonian (i.e., how much such a configuration is favored). In our case of interest, on the other hand, i.e. when \(\alpha > 1\), \(J''_\alpha(p) > 0\) in an interval around \(1/2\) so there is \(p* < 1/2\) with \(J_\alpha(p*) > J_\alpha(1/2) = 1\). Going back to (5.10) and bounding 

\[
Z(\beta) \geq 2Y_{\alpha,\{p,p\}} \left\lceil p n \right\rceil, 
\]

we get

\[
\Phi^X_* = O(\exp(-n[J_\alpha(p*) - J_\alpha(1/2)])) = O(\exp(-n[J_\alpha(p*) - J_\alpha(1/2)])).
\]

That concludes the proof.

**Expander graphs** In the proof of Claim 5.44, the bottleneck slowing down the chain arises as a result of the fact that, when \(m(\sigma) = -1\), there is a large number of edges in the underlying graph \(K_n\) connecting \(M_\sigma\) and \(M_c\). That produces a low probability for such configurations under the Ising model. The same argument easily extends to expander graphs.

**Claim 5.45** (Ising model on expander graphs: slow mixing of the Glauber dynamics). Let \(\{G_n\}_{n}\) be a \((d,\gamma)\)-expander family. For large enough inverse temperature \(\beta > 0\), the Glauber dynamics of the Ising model on \(G_n\) satisfies \(t_{\text{mix}}(\varepsilon) = \Omega(\exp(r(\beta)|V(G_n)|))\) for some function \(r(\beta) > 0\) not depending on \(n\).

In words, the absence of a bottleneck in the underlying graph produces a bottleneck in the Glauber dynamics.

**Proof.** Let \(\mu_\beta\) be the probability distribution over spin configurations under the Ising model over \(G_n = (V,E)\) with inverse temperature \(\beta\). Let \(Q_\beta\) be the transition matrix of the Glauber dynamics. For a subsets of vertices \(W_0, W_1 \subseteq V\) in the underlying graph \(G_n\), let

\[
E(W_0, W_1) := \{(u,v) : u,v \in V, (u,v) \in E\},
\]

be the set of edges with one endpoint in \(W_0\) and one endpoint in \(W_1\). Let \(N = |V(G_n)|\). We use the notation in the proof of Claim 5.44. Following the argument in that proof, we observe that (5.8) and (5.9) still hold. Thus

\[
\Phi^X_* \leq (1 + o(1)) \sum_{\sigma \in S_{-1}} \frac{\exp(\beta \left[|E(M_\sigma, M_\sigma)| + |E(M_c, M_c)| - |E(M_\sigma, M_c)|\right])}{Z(\beta)}.
\]

As we did in (5.10), we bound the partition function 

\[
Z(\beta) = \sum_{\sigma \in X} e^{-\beta \mathcal{H}(\sigma)}
\]

with
the term for the all-\((-1)\) configuration, leading to

\[
\Phi^{X} \leq (1 + o(1)) \sum_{\sigma \in S_{-1}} \frac{\exp (\beta [ | E(M_{\sigma}, M_{\sigma}) | + | E(M_{\sigma}^{c}, M_{\sigma}^{c}) | - | E(M_{\sigma}, M_{\sigma}^{c}) | ])}{\exp (\beta [ | E(M_{\sigma}, M_{\sigma}) | + | E(M_{\sigma}^{c}, M_{\sigma}^{c}) | + | E(M_{\sigma}, M_{\sigma}^{c}) | ])}
\]

\[
\leq (1 + o(1)) \sum_{\sigma \in S_{-1}} \exp (-2\beta | E(M_{\sigma}, M_{\sigma}^{c}) |)
\]

\[
= (1 + o(1)) \sum_{\sigma \in S_{-1}} \exp (-2\beta | \partial_{E} M_{\sigma}^{c} |)
\]

\[
= (1 + o(1)) \left( \frac{N}{N/2} \right) \exp (-2\beta \gamma | M_{\sigma}^{c} |)
\]

\[
= (1 + o(1)) \sqrt{\frac{2}{\pi N}} 2^{N}(1 + o(1)) \exp (-\beta \gamma (N - 1))
\]

\[
= C_{\beta, \gamma} \sqrt{\frac{2}{\pi N}} \exp (-N [\beta \gamma - \ln 2]),
\]

for some constant $C_{\beta, \gamma} > 0$. Taking $\beta > 0$ large enough gives the result. \hfill \blacksquare

**Exercises**

**Exercise 5.1** (Zhang’s proof of Harris’ theorem. From [Ste].)

**Exercise 5.2** (Mixing time: necessary condition for cutoff). Consider a sequence of Markov chains indexed by $n = 1, 2, \ldots$. Assume that each chain has a finite state space and is irreducible, aperiodic, and reversible. Let $t_{\text{mix}}^{(n)}(\varepsilon)$ and $t_{\text{rel}}^{(n)}$ be respectively the mixing time and relaxation time of the $n$-th chain. The sequence is said to have pre-cutoff if

\[
\sup_{0<\varepsilon<1/2} \limsup_{n \to +\infty} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{mix}}^{(n)}(1-\varepsilon)} < +\infty.
\]

Show that if for some $\varepsilon > 0$

\[
\sup_{n \geq 1} \frac{t_{\text{mix}}^{(n)}(\varepsilon)}{t_{\text{rel}}^{(n)}} < +\infty,
\]

then there is no pre-cutoff. In particular, there is no cutoff, as defined in Remark 4.71.
Notes

Bibliographic Remarks

Section 5.2  For a fascinating introduction to expander graphs and their applications, see [HLW06].
Chapter 6

Branching processes

Branching processes arise naturally in the study of stochastic processes on trees and locally tree-like graphs. After a review of the basic extinction theory of branching processes, we give a few classical examples of applications in discrete probability.

6.1 Background

We begin with a review of the extinction theory of Galton-Watson branching processes.

6.1.1 Basic definitions

Recall the definition of a Galton-Watson process.

**Definition 6.1.** A Galton-Watson branching process is a Markov chain of the following form:

- Let $Z_0 := 1$.
- Let $X(i,t)$, $i \geq 1$, $t \geq 1$, be an array of i.i.d. $\mathbb{Z}_+^+$-valued random variables with finite mean $m = \mathbb{E}[X(1,1)] < +\infty$, and define inductively,

$$Z_t := \sum_{1 \leq i \leq Z_{t-1}} X(i,t).$$

To avoid trivialities we assume $\mathbb{P}[X(1,1) = i] < 1$ for all $i \geq 0$. 
In words, \( Z_t \) models the size of a population at time (or generation) \( t \). The random variable \( X(i,t) \) corresponds to the number of offspring of the \( i \)-th individual (if there is one) in generation \( t - 1 \). Generation \( t \) is formed of all offsprings of the individuals in generation \( t - 1 \).

We denote by \( \{p_k\}_{k \geq 0} \) the law of \( X(1,1) \). We also let \( f(s) := \mathbb{E}[s^{X(1,1)}] \) be the corresponding probability generating function.

By tracking genealogical relationships, that is, who is whose child, we obtain a tree \( T \) rooted at the single individual in generation 0 with a vertex for each individual in the progeny and an edge for each parent-child relationship. We refer to \( T \) as a *Galton-Watson tree*.

A basic observation about Galton-Watson processes is that their growth is exponential in \( t \).

**Lemma 6.2 (Exponential growth I).** Let \( M_t := m^{-t}Z_t \). Then \( (M_t) \) is a nonnegative martingale with respect to the filtration \( \mathcal{F}_t = \sigma(Z_0, \ldots, Z_t) \). In particular, \( \mathbb{E}[Z_t] = m^t \).

**Proof.** Recall the following measure-theoretic lemma (see e.g. [Dur10, Exercise 5.1.1]).

**Lemma 6.3.** Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space. If \( Y_1 = Y_2 \) a.s. on \( B \in \mathcal{F} \) then \( \mathbb{E}[Y_1 \mid \mathcal{F}] = \mathbb{E}[Y_2 \mid \mathcal{F}] \) a.s. on \( B \).

Returning to the proof, observe that on \( \{Z_{t-1} = k\} \)

\[
\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \mathbb{E}\left[ \sum_{1 \leq j \leq k} X(j,t) \mid \mathcal{F}_{t-1} \right] = mk = mZ_{t-1}.
\]

This is true for all \( k \). Rearranging shows that \( (M_t) \) is a martingale. For the second claim, note that \( \mathbb{E}[M_t] = \mathbb{E}[M_0] = 1 \).

In fact, the martingale convergence theorem gives the following.

**Lemma 6.4 (Exponential growth II).** We have \( M_t \to M_\infty < +\infty \) a.s. for some nonnegative random variable \( M_\infty \in \sigma(\cup_t \mathcal{F}_t) \) with \( \mathbb{E}[M_\infty] \leq 1 \).

**Proof.** This follows immediately from the martingale convergence theorem for nonnegative martingales (Corollary 3.36) and Fatou’s lemma.
6.1.2 Extinction

Observe that \( 0 \) is a fixed point of the process. The event

\[
\{ Z_t \to 0 \} = \{ \exists t : Z_t = 0 \},
\]

is called extinction. Establishing when extinction occurs is a central question in branching process theory. We let \( \eta \) be the probability of extinction. Throughout, we assume that \( p_0 > 0 \) and \( p_1 < 1 \). Here is a first observation about extinction.

**Lemma 6.5.** A.s. either \( Z_t \to 0 \) or \( Z_t \to +\infty \).

**Proof.** The process \( (Z_t) \) is integer-valued and 0 is the only fixed point of the process under the assumption that \( p_1 < 1 \). From any state \( k \), the probability of never coming back to \( k > 0 \) is at least \( p_k^0 > 0 \), so every state \( k > 0 \) is transient. The claim follows.

In the critical case, that immediately implies almost sure extinction.

**Theorem 6.6** (Extinction: critical case). Assume \( m = 1 \). Then \( Z_t \to 0 \) a.s., i.e., \( \eta = 1 \).

**Proof.** When \( m = 1 \), \( (Z_t) \) itself is a martingale. Hence \( (Z_t) \) must converge to 0 by Lemma 6.4.

We address the general case using generating functions. Let \( f_t(s) = \mathbb{E}[s^{Z_t}] \).

Note that, by monotonicity,

\[
\eta = \mathbb{P}[\exists t \geq 0 : Z_t = 0] = \lim_{t \to +\infty} \mathbb{P}[Z_t = 0] = \lim_{t \to +\infty} f_t(0), \tag{6.1}
\]

Moreover, by the Markov property, \( f_t \) has a natural recursive form

\[
\begin{align*}
    f_t(s) &= \mathbb{E}[s^{Z_t}] \\
    &= \mathbb{E}[\mathbb{E}[s^{Z_t} \mid \mathcal{F}_{t-1}]] \\
    &= \mathbb{E}[f(s)^{Z_{t-1}}] \\
    &= f_{t-1}(f(s)) = \cdots = f^{(t)}(s), \tag{6.2}
\end{align*}
\]

where \( f^{(t)} \) is the \( t \)-th iterate of \( f \).

**Theorem 6.7** (Extinction: subcritical and supercritical cases). The probability of extinction \( \eta \) is given by the smallest fixed point of \( f \) in \([0, 1]\). Moreover:

- (Subcritical regime) If \( m < 1 \) then \( \eta = 1 \).
• (Supercritical regime) If \( m > 1 \) then \( \eta < 1 \).

**Proof.** The case \( p_0 + p_1 = 1 \) is straightforward: the process dies almost surely after a geometrically distributed time. So we assume \( p_0 + p_1 < 1 \) for the rest of the proof.

We first summarize some properties of \( f \).

**Lemma 6.8.** On \([0, 1]\), the function \( f \) satisfies:

(a) \( f(0) = p_0, \ f(1) = 1; \)

(b) \( f \) is indefinitely differentiable on \([0, 1); \)

(c) \( f \) is strictly convex and increasing;

(d) \( \lim_{s \uparrow 1} f'(s) = m < +\infty. \)

**Proof.** See e.g. [Rud76] for the relevant power series facts. Observe that (a) is clear by definition. The function \( f \) is a power series with radius of convergence \( R \geq 1 \). This implies (b). In particular,

\[
f'(s) = \sum_{i \geq 1} i p_i s^{i-1} \geq 0, \quad \text{and} \quad f''(s) = \sum_{i \geq 2} i(i-1) p_i s^{i-2} > 0,
\]

because we must have \( p_i > 0 \) for some \( i > 1 \) by assumption. This proves (c). Since \( m < +\infty, f'(1) = m \) is well defined and \( f' \) is continuous on \([0, 1], \) which implies (d). \( \Box \)

We first characterize the fixed points of \( f \). See Figure 6.1 for an illustration.

**Lemma 6.9.** We have:

- If \( m > 1 \) then \( f \) has a unique fixed point \( \eta_0 \in [0, 1) \).

- If \( m < 1 \) then \( f(t) > t \) for \( t \in [0, 1) \). Let \( \eta_0 := 1 \) in that case.

**Proof.** Assume \( m > 1 \). Since \( f'(1) = m > 1 \), there is \( \delta > 0 \) s.t. \( f(1 - \delta) < 1 - \delta. \) On the other hand \( f(0) = p_0 > 0 \) so by continuity of \( f \) there must be a fixed point in \((0, 1 - \delta) \). Moreover, by strict convexity and the fact that \( f(1) = 1, \) if \( x \in (0, 1) \) is a fixed point then \( f(y) < y \) for \( y \in (x, 1) \), proving uniqueness.

The second part follows by strict convexity and monotonicity. \( \Box \)

It remains to prove convergence of the iterates to the appropriate fixed point. See Figure 6.2 for an illustration.

**Lemma 6.10.** We have:
Figure 6.1: Fixed points of $f$ in subcritical (left) and supercritical (right) cases.

Figure 6.2: Convergence of iterates to a fixed point.
• If $x \in [0, \eta_0)$, then $f^{(t)}(x) \uparrow \eta_0$
• If $x \in (\eta_0, 1)$ then $f^{(t)}(x) \downarrow \eta_0$

Proof. We only prove 1. The argument for 2. is similar. By monotonicity, for $x \in [0, \eta_0)$, we have $x < f(x) < f(\eta_0) = \eta_0$. Iterating
\[ x < f^{(1)}(x) < \cdots < f^{(t)}(x) < f^{(t)}(\eta_0) = \eta_0. \]
So $f^{(t)}(x) \uparrow L \leq \eta_0$. By continuity of $f$ we can take the limit inside of
\[ f^{(t)}(x) = f(f^{(t-1)}(x)), \]
to get $L = f(L)$. So by definition of $\eta_0$ we must have $L = \eta_0$.

The result then follows from the above lemmas together with Equations (6.1) and (6.2).

Example 6.11 (Poisson branching process). Consider the offspring distribution $X(1, 1) \sim \text{Poi}(\lambda)$ with $\lambda > 0$. We refer to this case as the Poisson branching process. Then
\[
    f(s) = \mathbb{E}[s^{X(1, 1)}] = \sum_{i \geq 0} e^{-\lambda} \frac{\lambda^i}{i!} s^i = e^{\lambda(s-1)}.
\]
So the process goes extinct with probability 1 when $\lambda \leq 1$. For $\lambda > 1$, the probability of extinction $\eta_\lambda$ is the smallest solution in $[0, 1]$ to the equation
\[
e^{-\lambda(1-x)} = x.\]
The survival probability $\zeta_\lambda := 1 - \eta_\lambda$ satisfies $1 - e^{-\lambda \zeta_\lambda} = \zeta_\lambda$. ♦

We can use the extinction results to obtain more information on the limit in Lemma 6.4. Of course, conditioned on extinction, $M_\infty = 0$ a.s. On the other hand:

Lemma 6.12 (Exponential growth III). Conditioned on nonextinction, either $M_\infty = 0$ a.s. or $M_\infty > 0$ a.s. In particular, $\mathbb{P}[M_\infty = 0] \in \{\eta, 1\}$.

Proof. A property of rooted trees is said to be inherited if all finite trees satisfy this property and whenever a tree satisfies the property then so do all the descendant trees of the children of the root. The property $\{M_\infty = 0\}$ is inherited. The result then follows from the following 0-1 law.
Lemma 6.13 (0-1 law for inherited properties). For a Galton-Watson tree $T$, an inherited property $A$ has, conditioned on nonextinction, probability 0 or 1.

Proof. Let $T^{(1)}, \ldots, T^{(Z_1)}$ be the descendant subtrees of the children of the root. We use the notation $T \in A$ to mean that the tree $T$ satisfies $A$. Then, by independence,

$$P[A] = E[P[T \in A \mid Z_1] \leq E[P[T^{(i)} \in A, \forall i \leq Z_1] \mid Z_1] = E[P[A]^{Z_1}] = f(P[A]),$$

so $P[A] \in [0, \eta] \cup \{1\}$. Also $P[A] \geq \eta$ because $A$ holds for finite trees.

That concludes the proof.

A further moment assumption provides a more detailed picture.

Lemma 6.14 (Exponential growth: finite second moment). Let $(Z_t)$ be a branching process with $m = E[X(1, 1)] > 1$ and $\sigma^2 = \text{Var}[X(1, 1)] < +\infty$. Then, $(M_t)$ converges in $L^2$ and, in particular, $E[M_\infty] = 1$. Further, $P[M_\infty = 0] = \eta$.

Proof. We bound $E[M_t^2]$ by computing it explicitly by induction. From the orthogonality of increments (Lemma 3.40), it holds that

$$E[M_t^2] = E[M_{t-1}^2] + E[(M_t - M_{t-1})^2].$$

On $\{Z_{t-1} = k\}$,

$$E[(M_t - M_{t-1})^2 \mid \mathcal{F}_{t-1}] = m^{-2t}E[(Z_t - mZ_{t-1})^2 \mid \mathcal{F}_{t-1}]$$

$$= m^{-2t}E\left[\left(\sum_{i=1}^{k} X(i, t) - mk\right)^2 \mid \mathcal{F}_{t-1}\right]$$

$$= m^{-2t}k\sigma^2$$

$$= m^{-2t}Z_{t-1}\sigma^2.$$

Hence

$$E[M_t^2] = E[M_{t-1}^2] + m^{-t-1}\sigma^2.$$ 

Since $E[M_0^2] = 1$,

$$E[M_t^2] = 1 + \sigma^2 \sum_{i=2}^{t+1} m^{-i},$$

which is uniformly bounded when $m > 1$. Therefore $(M_t)$ converges in $L^2$ (see e.g. [Dur10, Theorem 5.4.5]). Finally, by Fatou’s lemma,

$$E|M_\infty| \leq \sup \|M_t\|_1 \leq \sup \|M_t\|_2 < +\infty.$$
and
\[ |\mathbb{E}[M_t] - \mathbb{E}[M_\infty]| \leq \|M_t - M_\infty\|_1 \leq \|M_t - M_\infty\|_2, \]
implies the convergence of expectations.

The last statement follows from Lemma 6.12.

\[ \textbf{Remark 6.15.} \text{ The Kesten-Stigum Theorem gives a necessary and sufficient condition for } \mathbb{E}[M_\infty] = 1 \text{ to hold [KS66b]. See e.g. [LP, Chapter 12].} \]

6.1.3 \quad Bond percolation on Galton-Watson trees

Let \( T \) be a Galton-Watson tree for an offspring distribution with mean \( m > 1 \). Perform bond percolation on \( T \) with density \( p \).

**Theorem 6.16** (Bond percolation on Galton-Watson trees). *Conditioned on nonextinction,
\[
p_c(T) = \frac{1}{m} \text{ a.s.}
\]

**Proof.** Let \( C_0 \) be the cluster of the root in \( T \) with density \( p \). We can think of \( C_0 \) as being generated by a Galton-Watson branching process where the offspring distribution is the law of \( \sum_{i=1}^{X(1,1)} I_i \) where the \( I_i \)s are i.i.d. Ber(\( p \)) and \( X(1,1) \) is distributed according to the offspring distribution of \( T \). In particular, by conditioning on \( X(1,1) \), the offspring mean under \( C_0 \) is \( mp \). If \( mp \leq 1 \) then
\[
1 = \mathbb{P}_p[|C_0| < +\infty] = \mathbb{E}[\mathbb{P}_p[|C_0| < +\infty | T]],
\]
and we must have \( \mathbb{P}_p[|C_0| < +\infty | T] = 1 \) a.s. In other words, \( p_c(T) \geq \frac{1}{m} \) a.s.

On the other hand, the property of trees \( \{\mathbb{P}_p[|C_0| < +\infty | T] = 1\} \) is inherited. So by Lemma 6.13, conditioned on nonextinction, it has probability 0 or 1. That probability is of course 1 on extinction. So by
\[
\mathbb{P}_p[|C_0| < +\infty] = \mathbb{E}[\mathbb{P}_p[|C_0| < +\infty | T]],
\]
if the probability is 1 conditioned on nonextinction then it must be that \( mp \leq 1 \).

In other words, for any fixed \( p \) such that \( mp > 1 \), conditioned on nonextinction \( \mathbb{P}_p[|C_0| < +\infty | T] = 0 \) a.s. By monotonicity of \( \mathbb{P}_p[|C_0| < +\infty | T] \) in \( p \), taking a limit \( p_n \to 1/m \) proves the result.

\[ \textbf{6.1.4 \quad Random walk on Galton-Watson trees} \]

To be written. See [LP, Theorem 3.5 and Corollary 5.10].

*Requires: Section 2.3.3 and 3.1.1.*

316
6.2 Random-walk representation

In this section, we develop a useful random-walk representation of the Galton-Watson process.

6.2.1 Exploration process

Consider the following exploration process of a Galton-Watson tree $T$. The exploration process, started at the root 0, has 3 types of vertices:

- $A_t$: active vertices,
- $E_t$: explored vertices,
- $N_t$: neutral vertices.

We start with $A_0 := \{0\}$, $E_0 := \emptyset$, and $N_0$ contains all other vertices in $T$. At time $t$, if $A_{t-1} = \emptyset$ we let $(A_t, E_t, N_t) := (A_{t-1}, E_{t-1}, N_{t-1})$. Otherwise, we pick an element, $a_t$, from $A_{t-1}$ and set:

- $A_t := A_{t-1} \cup \{x \in N_{t-1} : \{x, a_t\} \in T\}\{a_t\}$,
- $E_t := E_{t-1} \cup \{a_t\}$,
- $N_t := N_{t-1}\{x \in N_{t-1} : \{x, a_t\} \in T\}$.

To be concrete, we choose $a_t$ in breadth-first search (or first-come-first-serve) manner: we exhaust all vertices in generation $t$ before considering vertices in generation $t + 1$.

We imagine revealing the edges of $T$ as they are encountered in the exploration process and we let $(\mathcal{F}_t)$ be the corresponding filtration. In words, starting with 0, the Galton-Watson tree $T$ is progressively grown by adding to it at each time a child of one of the previously explored vertices and uncovering its children in $T$. In this process, $E_t$ is the set of previously explored vertices and $A_t$ is the set of vertices who are known to belong to $T$ but whose full neighborhood is waiting to be uncovered. The rest of the vertices form the set $N_t$.

Let $A_t := |A_t|$, $E_t := |E_t|$, and $N_t := |N_t|$. Note that $(E_t)$ is non-decreasing while $(N_t)$ is non-increasing. Let

$$\tau_0 := \inf\{t \geq 0 : A_t = 0\},$$

(which by convention is $+\infty$ if there is no such $t$). The process is fixed for all $t > \tau_0$. Notice that $E_t = t$ for all $t \leq \tau_0$, as exactly one vertex is explored at each
time until the set of active vertices is empty. Moreover, for all \( t \), \((A_t, \mathcal{E}_t, N_t)\) forms a partition of \([n]\) so
\[
A_t + t + N_t = n, \quad \forall t \leq \tau_0.
\]

**Lemma 6.17** (Total progeny). Let \( W \) be the total progeny. Then
\[
W = \tau_0.
\]

The random-walk representation is the following. Observe that the process \((A_t)\) admits a simple recursive form. Recall that \( A_0 := 1 \). Conditioning on \( \mathcal{F}_{t-1} \):

- If \( A_{t-1} = 0 \), the exploration process has finished its course and \( A_t = 0 \).
- Otherwise, (a) one active vertex becomes an explored vertex and (b) its neutral neighbors become active vertices. That is,
\[
A_t = \begin{cases} 
A_{t-1} + \left[ \frac{1}{(a)} + X_t \right], & t - 1 < \tau_0, \\
0, & \text{o.w.}
\end{cases}
\]

where \( X_t \) is distributed according to the offspring distribution.

We let \( Y_t = X_t - 1 \geq -1 \) and
\[
S_t := 1 + \sum_{i=1}^{t} Y_i,
\]
with \( S_0 := 1 \). Then
\[
\tau_0 = \inf\{t \geq 0 : S_t = 0\}
= \inf\{t \geq 0 : 1 + [X_1 - 1] + \cdots + [X_t - 1] = 0\}
= \inf\{t \geq 0 : X_1 + \cdots + X_t = t - 1\},
\]
and \((A_t)\) is a random walk started at 1 with steps \((Y_t)\) stopped when it hits 0 for the first time:
\[
A_t = (S_t \wedge \tau_0).
\]

We give two applications.
6.2.2 Duality principle

The random-walk representation above is useful to prove the following duality principle.

**Theorem 6.18 (Duality principle).** Let \((Z_t)\) be a branching process with offspring distribution \(\{p_k\}_{k \geq 0}\) and extinction probability \(\eta < 1\). Let \((Z'_t)\) be a branching process with offspring distribution \(\{p'_k\}_{k \geq 0}\) where

\[
p'_k = \eta^{k-1} p_k.
\]

Then \((Z_t)\) conditioned on extinction has the same distribution as \((Z'_t)\), which is referred to as the duality branching process.

**Proof.** We use the random walk representation. Let \(H = (X_1, \ldots, X_{\tau_0})\) and \(H' = (X'_1, \ldots, X'_{\tau'_0})\) be the histories of the processes \((Z_t)\) and \((Z'_t)\) respectively. (Under breadth-first search, the process \((Z_t)\) can be reconstructed from \(H\).) In the case of extinction, the history of \((Z_t)\) has finite length. We call \((x_1, \ldots, x_t)\) a valid history if \(x_1 + \cdots + x_i - (i - 1) > 0\) for all \(i < t\) and \(x_1 + \cdots + x_t - (t - 1) = 0\). By definition of the conditional probability, for a valid history \((x_1, \ldots, x_t)\) with a finite \(t\),

\[
\mathbb{P}[H = (x_1, \ldots, x_t) \mid \tau_0 < +\infty] = \frac{\mathbb{P}[H = (x_1, \ldots, x_t)]}{\mathbb{P}[\tau_0 < +\infty]} = \eta^{-1} \prod_{i=1}^{t} p_{x_i}.
\]

Because \(x_1 + \cdots + x_t = t - 1\),

\[
\eta^{-1} \prod_{i=1}^{t} p_{x_i} = \eta^{-1} \prod_{i=1}^{t} \eta^{1-x_i} p'_{x_i} = \prod_{i=1}^{t} p'_{x_i} = \mathbb{P}[H' = (x_1, \ldots, x_t)].
\]

Note that

\[
\sum_{k \geq 0} p'_k = \sum_{k \geq 0} \eta^{k-1} p_k = \eta^{-1} f(\eta) = 1,
\]

because \(\eta\) is a fixed point of \(f\). So \(\{p'_k\}_{k \geq 0}\) is indeed a probability distribution. Note further that

\[
\sum_{k \geq 0} kp'_k = \sum_{k \geq 0} k\eta^{k-1} p_k = f'(\eta) < 1,
\]

since \(f'\) is strictly increasing, \(f(\eta) = \eta < 1\) and \(f(1) = 1\). So the dual branching process is subcritical.
Example 6.19 (Poisson branching process). Let \((Z_t)\) be a Galton-Watson branching process with offspring distribution \(\text{Poi}(\lambda)\) where \(\lambda > 1\). Then the dual probability distribution is given by

\[
p'_k = \eta^{k-1} p_k = \eta^{k-1} e^{-\lambda \frac{\lambda^k}{k!}} = \eta^{-1} e^{-\lambda \frac{(\lambda \eta)^k}{k!}},
\]

where recall that \(e^{-\lambda(1-\eta)} = \eta\), so

\[
p'_k = e^{\lambda(1-\eta)} \frac{\lambda^k}{k!} = e^{-\lambda \frac{(\lambda \eta)^k}{k!}}.
\]

That is, the dual branching process has offspring distribution \(\text{Poi}(\lambda \eta)\).

6.2.3 Hitting-time theorem

The random-walk representation also gives a formula for the distribution of the size of the progeny.

We start with a combinatorial lemma of independent interest. Let \(u_1, \ldots, u_t \in \mathbb{R}\) and define \(r_0 := 0\) and \(r_i := u_1 + \cdots + u_i\) for \(1 \leq i \leq t\). We say that \(j\) is a ladder index if \(r_j > r_0 \lor \cdots \lor r_{j-1}\). Consider the cyclic permutations of \(u = (u_1, \ldots, u_t)\): \(u^{(0)} = u, \ u^{(1)} = (u_2, \ldots, u_t, u_1), \ldots, u^{(t-1)} = (u_t, u_1, \ldots, u_{t-1})\). Define the corresponding partial sums \(r^{(\beta)}_j := u_1^{(\beta)} + \cdots + u_j^{(\beta)}\) for \(j = 1, \ldots, t\) and \(\beta = 0, \ldots, t - 1\). Observe that

\[
(r_1^{(\beta)}, \ldots, r_t^{(\beta)}) = (r_{\beta+1}^j - r_{\beta}, r_{\beta+2} - r_{\beta}, \ldots, r_t - r_{\beta},
[r_{t} - r_{\beta}] + r_1, [r_t - r_{\beta}] + r_2, \ldots, [r_t - r_{\beta}] + r_{\beta}) = (r_{\beta+1}^j - r_{\beta}, r_{\beta+2} - r_{\beta}, \ldots, r_t - r_{\beta},
[r_{t} - r_{\beta}] + r_1, [r_{t} - r_{\beta}] + r_2, \ldots, [r_{t} - r_{\beta}] + r_{t} - [r_{\beta} - r_{\beta-1}], r_t).
\]

(6.3)

Lemma 6.20 (Spitzer’s combinatorial lemma). Assume \(r_t > 0\). Let \(\ell\) be the number of cyclic permutations such that \(t\) is a ladder index. Then \(\ell \geq 1\). Moreover, each such cyclic permutation has exactly \(\ell\) ladder indices.

Proof. We first show that \(\ell \geq 1\), i.e., there is at least one cyclic permutation where \(t\) is a ladder index. Let \(\beta\) be the smallest index achieving the maximum of \(r_1, \ldots, r_t\), i.e.,

\[
\ r_{\beta} \geq r_1 \lor \cdots \lor r_{\beta-1} \quad \text{and} \quad r_{\beta} \geq r_{\beta+1} \lor \cdots \lor r_t.
\]

From (6.3),

\[
r_{\beta+i} - r_{\beta} \leq 0 < r_t, \quad \forall i = 1, \ldots, t - \beta,
\]

320
and
\[ r_t - [r_\beta - r_j] < r_t, \quad \forall j = 1, \ldots, \beta - 1. \]
Moreover, \( r_t > 0 = r_0 \) by assumption. So, in \( u^{(\beta)} \), \( t \) is a ladder index.

Since \( \ell \geq 1 \), we can assume w.l.o.g. that \( u \) is such that \( t \) is a ladder index. Then \( \beta \) is a ladder index in \( u \) if and only if
\[ r_\beta > r_0 \lor \cdots \lor r_{\beta-1}, \]
if and only if
\[ r_t > r_t - r_\beta \quad \text{and} \quad r_t - [r_\beta - r_j] < r_t, \quad \forall j = 1, \ldots, \beta - 1. \]
Moreover, because \( r_t > r_j \) for all \( j \), we have \( r_t - [r_{\beta+i} - r_\beta] = (r_t - r_{\beta+i}) + r_\beta \) and the last equation is equivalent to
\[ r_t > r_t - [r_{\beta+i} - r_\beta], \quad \forall i = 1, \ldots, t - \beta \]
and
\[ r_t - [r_\beta - r_j] < r_t, \quad \forall j = 1, \ldots, \beta - 1. \]
That is, \( t \) is a ladder index in the \( \beta \)-th cyclic permutation. \( \blacksquare \)

We are now ready to prove the hitting-time theorem.

**Theorem 6.21** (Hitting-time theorem). Let \( (Z_t) \) be a Galton-Watson branching process with total progeny \( W \). In the random walk representation of \( (Z_t) \),
\[ \mathbb{P}[W = t] = \frac{1}{t} \mathbb{P}[X_1 + \cdots + X_t = t - 1], \]
for all \( t \geq 1 \).

**Proof.** Let \( R_i := 1 - S_i \) and \( U_i := 1 - X_i \) for all \( i = 1, \ldots, t \) and let \( R_0 := 0 \).
Then
\[ \{ X_1 + \cdots + X_t = t - 1 \} = \{ R_t = 1 \}, \]
and
\[ \{ W = t \} = \{ t \text{ is the first ladder index in } R_1, \ldots, R_t \}. \]
By symmetry, for all \( \beta \)
\[ \mathbb{P}[t \text{ is the first ladder index in } R_1, \ldots, R_t] \]
\[ = \mathbb{P}[t \text{ is the first ladder index in } R_1^{(\beta)}, \ldots, R_t^{(\beta)}]. \]
Let $\mathcal{E}_\beta$ be the event on the last line. Hence
\[
\mathbb{P}[W = t] = \mathbb{E}[1_{\mathcal{E}_1}] = \frac{1}{t} \mathbb{E} \left[ \sum_{\beta=1}^{t} 1_{\mathcal{E}_\beta} \right].
\]

By Spitzer’s combinatorial lemma, there is at most one cyclic permutation where $t$ is the first ladder index. In particular, $\sum_{\beta=1}^{t} 1_{\mathcal{E}_\beta} \in \{0, 1\}$. So
\[
\mathbb{P}[W = t] = \frac{1}{t} \mathbb{P}[\cup_{\beta=1}^{t} \mathcal{E}_\beta].
\]

Finally observe that, because $R_0 = 0$ and $U_i \leq 1$ for all $i$, the partial sum at the $j$-th ladder index must take value $j$. So the event $\{\cup_{\beta=1}^{t} \mathcal{E}_\beta\}$ implies that $\{R_t = 1\}$ because the last partial sum of all cyclic permutations is $R_t$. Similarly, because there is at least one cyclic permutation such that $t$ is a ladder index, the event $\{R_t = 1\}$ implies $\{\cup_{\beta=1}^{t} \mathcal{E}_\beta\}$. Therefore,
\[
\mathbb{P}[W = t] = \frac{1}{t} \mathbb{P}[R_t = 1],
\]
which concludes the proof.

Note that the formula in the hitting-time theorem is somewhat remarkable as the probability on the l.h.s. is $\mathbb{P}[S_i > 0, \forall i < t$ and $S_t = 0]$ while the probability on the r.h.s. is $\frac{1}{t} \mathbb{P}[S_t = 0]$.

**Example 6.22 (Poisson branching process).** Let $(Z_t)$ be a Galton-Watson branching process with offspring distribution $\text{Poi}(\lambda)$ where $\lambda > 0$. Let $W$ be its total progeny. By the hitting-time theorem, for $t \geq 1$,
\[
\mathbb{P}[W = t] = \frac{1}{t} \mathbb{P}[X_1 + \cdots + X_t = t - 1]
= \frac{1}{t} e^{-\lambda t} \frac{(\lambda t)^{t-1}}{(t-1)!}
= e^{-\lambda t} \frac{(\lambda t)^{t-1}}{t!},
\]
where we used that a sum of independent Poisson is Poisson.
6.3 Comparison to branching processes

We begin with an example whose connection to branching processes is clear: percolation on trees. Translating standard branching process results into their percolation counterpart immediately gives a more detailed picture of the behavior of the process than was derived in Section 2.3.3. We then tackle the phase transition of Erdős-Rényi graphs using a comparison to branching processes.

6.3.1 Percolation on trees: critical exponents

In this section, we use branching processes to study bond percolation on the infinite $b$-ary tree $\hat{T}_b$. The same techniques can be adapted to $\hat{T}_d$ with $d = b + 1$ in a straightforward manner.

We denote the root by $0$. We think of the open cluster of the root, $C_0$, as the progeny of a branching process as follows. Denote by $\partial_n$ the $n$-th level of $\hat{T}_b$, that is, the vertices of $\hat{T}_b$ at graph distance $n$ from the root. In the branching process interpretation, we think of the immediate descendants in $C_0$ of a vertex $v$ as the “children” of $v$. By construction, $v$ has at most $b$ children, independently of all other vertices in the same generation. In this branching process, the offspring distribution is binomial with parameters $b$ and $p$; $Z_n := |C_0 \cap \partial_n|$ represents the size of the progeny at generation $n$; and $W := |C_0|$ is the total progeny of the process. In particular $|C_0| < +\infty$ if and only if the process goes extinct. Because the mean number of offspring is $bp$, by Theorem 6.7, this leads immediately to a (second) proof of:

Claim 6.23.

$$p_c(\hat{T}_b) = \frac{1}{b}.$$  

The generating function of the offspring distribution is $\phi(s) := ((1 - p) + ps)^b$. So, by Theorems 6.6 and 6.7, the percolation function

$$\theta(p) = \mathbb{P}_p[|C_0| = +\infty],$$

is 0 on $[0, 1/b]$, while on $(1/b, 1]$ the quantity $\eta(p) := (1 - \theta(p))$ is the unique solution in $[0, 1)$ of the fixed point equation

$$s = ((1 - p) + ps)^b. \quad (6.4)$$

For $b = 2$, for instance, we can compute the fixed point explicitly by noting that

$$0 = ((1 - p) + ps)^2 - s = p^2 s^2 + [2p(1 - p) - 1]s + (1 - p)^2,$$
whose solution for \( p \in (1/2, 1] \) is

\[
s^* = \frac{-2p(1 - p) - 1 \pm \sqrt{4p^2(1 - p)^2 - 4p^2(1 - p)^2}}{2p^2}
\]

\[
= \frac{-2p(1 - p) - 1 \pm \sqrt{1 - 4p(1 - p)}}{2p^2}
\]

\[
= \frac{-2p(1 - p) - 1 \pm (2p - 1)}{2p^2}
\]

\[
= \frac{2p^2 + [(1 - 2p) \pm (2p - 1)]}{2p^2}.
\]

So, rejecting the fixed point 1,

\[
\theta(p) = 1 - \frac{2p^2 + 2(1 - 2p)}{2p^2} = \frac{2p - 1}{p^2}.
\]

We have proved:

**Claim 6.24.** For \( b = 2 \),

\[
\theta(p) = \begin{cases} 
0, & 0 \leq p \leq \frac{1}{2}, \\
\frac{2(p - \frac{1}{2})}{p^2}, & \frac{1}{2} < p \leq 1.
\end{cases}
\]

The expected size of the population at generation \( n \) is \((bp)^n\) so for \( p \in [0, \frac{1}{b}]\)

\[
\mathbb{E}|C_0| = \sum_{n \geq 0} (bp)^n = \frac{1}{1 - bp}.
\]

For \( p \in (\frac{1}{b}, 1) \), the total progeny is almost surely infinite, but it is of interest to compute the expected cluster size **conditioned on** \(|C_0| < +\infty\). We use the duality principle, Theorem 6.18. For \( 0 \leq k \leq b \), let

\[
\hat{p}_k := [\eta(p)]^{k-1}p_k
\]

\[
= [\eta(p)]^{k-1}\binom{b}{k} p^k (1 - p)^{b-k}
\]

\[
= \frac{[\eta(p)]^k}{((1 - p) + p \eta(p))^b} \binom{b}{k} p^k (1 - p)^{b-k}
\]

\[
= \binom{b}{k} \left( \frac{p \eta(p)}{(1 - p) + p \eta(p)} \right)^k \left( \frac{1 - p}{(1 - p) + p \eta(p)} \right)^{b-k}
\]

\[
= \binom{b}{k} \hat{p}_k^{b-k} (1 - \hat{p})^{b-k}
\]

324
where we used (6.4) and implicitly defined the dual density

\[
\hat{p} := \frac{p \eta(p)}{(1 - p) + p \eta(p)}.
\] (6.5)

In particular \(\{\hat{p}_k\}\) is indeed a probability distribution. In fact it is binomial with parameters \(b\) and \(\hat{p}\). The corresponding generating function is

\[
\hat{\phi}(s) := ((1 - \hat{p}) + \hat{p} s)^b = \eta(p)^{-1} \phi(s \eta(p)),
\]

where the second expression can be seen directly from the definition of \(\{\hat{p}_k\}\). Moreover,

\[
\hat{\phi}'(s) = \eta(p)^{-1} \phi'(s \eta(p)) \eta(p) = \phi'(s \eta(p)),
\]

so \(\hat{\phi}'(1^-) = \phi'(\eta(p)) < 1\) by the proof of Theorem 6.7, confirming that percolation with density \(\hat{p}\) is subcritical. Summarizing:

**Claim 6.25.** Conditioned on \(|C_0| < +\infty\), (supercritical) percolation on \(\hat{T}_b\) with density \(p \in (\frac{1}{b}, 1)\) has the same distribution as (subcritical) percolation on \(\hat{T}_b\) with density defined by (6.5).

Therefore:

**Claim 6.26.**

\[
\chi^f(p) := \mathbb{E}_p [\mathbb{I}_{|C_0| < +\infty}] = \begin{cases} 
\frac{1}{1 - bp}, & p \in [0, \frac{1}{b}), \\
\frac{\eta(p)}{1 - bp}, & p \in (\frac{1}{b}, 1).
\end{cases}
\]

For \(b = 2\), \(\eta(p) = 1 - \theta(p) = \left(\frac{1-p}{p}\right)^2\) so

\[
\hat{p} = \frac{p \left(\frac{1-p}{p}\right)^2}{(1 - p) + p \left(\frac{1-p}{p}\right)^2} = \frac{(1 - p)^2}{p(1 - p) + (1 - p)^2} = 1 - p,
\]

and

**Claim 6.27.** For \(b = 2\),

\[
\chi^f(p) = \begin{cases} 
\frac{1}{2 - \hat{p}}, & p \in [0, \frac{1}{2}), \\
\frac{1}{2} \left(\frac{1-p}{p}\right)^2, & p \in (\frac{1}{2}, 1).
\end{cases}
\]
In fact, the hitting-time theorem, Theorem 6.21, gives an explicit formula for
the distribution of $|C_0|$. Namely, because $|C_0| \overset{d}{=} \tau_0$ for $S_t = \sum_{\ell \leq t} X_\ell - (t - 1)$
where $S_0 = 1$ and the $X_\ell$s are i.i.d. binomial with parameters $b$ and $p$ and further

$$P[\tau_0 = \ell] = \frac{1}{\ell} P[S_\ell = 0],$$

we have

$$P_p[|C_0| = \ell] = \frac{1}{\ell} P\left[ \sum_{\ell \leq \ell} X_\ell = \ell - 1 \right] = \frac{1}{\ell} \binom{b\ell}{\ell - 1} p^{\ell - 1} (1 - p)^{b\ell - (\ell - 1)}, \quad (6.6)$$

where we used that a sum of independent binomials with the same $p$ is still binomial. In particular, at criticality, using Stirling’s formula it can be checked that

$$P_{p_c}[|C_0| = \ell] \sim \frac{1}{\ell} \frac{1}{\sqrt{2\pi p_c (1 - p_c) b\ell}} \frac{1}{\sqrt{2\pi (1 - p_c) \ell^3}}.$$

as $\ell \to +\infty$.

Close to criticality, physicists predict that many quantities behave according to
critical exponent
power laws of the form $|p - p_c|^\beta$, where the exponent is referred to as a critical exponent. The critical exponents are believed to satisfy certain “universality” properties. But even proving the existence of such exponents in general remains a major open problem. On trees, though, we can simply read off the critical exponents from the above formulas. For $b = 2$, Claims 6.24 and 6.27 imply for instance that, as $p \to p_c$,

$$\theta(p) \sim 8(p - p_c) 1_{\{p > 1/2\}},$$

and

$$\chi^f(p) \sim \frac{1}{2} |p - p_c|^{-1}.$$

In fact, as can be seen from Claim 6.26, the critical exponent of $\chi^f(p)$ does not depend on $b$. The same holds for $\theta(p)$. See Exercise 6.5. Using (6.6), the higher
moments of $|C_0|$ can also be studied around criticality. See Exercise 6.6.

6.3.2 Random binary search trees: height

To be written. See [Dev98, Section 2.1].
6.3.3 Erdős-Rényi graph: the phase transition

A compelling way to view Erdős-Rényi graphs as the density varies is the following coupling or "evolution."† For each pair \( \{i,j\} \), let \( U_{\{i,j\}} \) be independent uniform random variables in \([0,1]\) and set \( \mathcal{G}(p) := ([n], \mathcal{E}(p)) \) where \( \{i,j\} \in \mathcal{E}(p) \) if and only if \( U_{\{i,j\}} \leq p \). Then \( \mathcal{G}(p) \) is distributed according to \( \mathbb{G}_{n,p} \). As \( p \) varies from 0 to 1, we start with an empty graph and progressively add edges until the complete graph is obtained. We showed in Section 2.3.2 that \( \frac{\log n}{n} \) is a threshold function for connectivity. Before connectivity occurs in the evolution of the random graph, a quantity of interest is the size of the largest connected component. As we show in this section, this quantity itself undergoes a remarkable phase transition: when \( p = \frac{\lambda}{n} \) with \( \lambda < 1 \), the largest component has size \( \Theta(\log n) \); as \( \lambda \) crosses 1, many components quickly merge to form a so-called "giant component" of size \( \Theta(n) \).

This celebrated result of Erdős and Rényi, which is often referred to as “the” phase transition of the Erdős-Rényi graph, is related to the phase transition in percolation. That should be clear from the similarities between the proofs, specifically the branching process approach to percolation on trees (Section 6.3.1). Although the proof is quite long, it is well worth studying in details. It employs most tools we have seen up to this point: first and second moment methods, Chernoff-Cramér bound, martingale techniques, coupling and stochastic domination, and branching processes. It is quintessential discrete probability.

Statements and proof sketch Before stating the main theorems, we recall a basic result from Chapter 2.

- (Poisson tail) Let \( S_n \) be a sum of \( n \) i.i.d. \( \text{Po}i(\lambda) \) variables. Recall from (2.33) and (2.34) that for \( a > \lambda \)

\[
-\frac{1}{n} \log \mathbb{P}[S_n \geq an] \geq a \log \left( \frac{a}{\lambda} \right) - a + \lambda =: I_{\lambda}^{\text{Poi}}(a),
\]

and similarly for \( a < \lambda \)

\[
-\frac{1}{n} \log \mathbb{P}[S_n \leq an] \geq I_{\lambda}^{\text{Poi}}(a).
\]

To simplify the notation, we let

\[
I_{\lambda} := I_{\lambda}^{\text{Poi}}(1) = \lambda - 1 - \log \lambda \geq 0,
\]

where the inequality follows from the convexity of \( I_{\lambda} \) and the fact that it attains its minimum at \( \lambda = 1 \) where it is 0.

† Requires: Sections 2.3.1, 2.4.1, 4.1 and 4.3.1
We let $p = \frac{\lambda}{n}$ and denote by $C_{\text{max}}$ a largest connected component. In the subcritical case, that is, when $\lambda < 1$, we show that the largest connected component has logarithmic size in $n$.

**Theorem 6.28** (Subcritical case: upper bound on the largest cluster). Let $G_n \sim \mathbb{G}_{n,p_n}$ where $p_n = \frac{\lambda}{n}$ with $\lambda \in (0, 1)$. For all $\kappa > 0$,

$$
\Pr_{n,p_n}[|C_{\text{max}}| > (1 + \kappa)I_\lambda^{-1}\log n] = o(1),
$$

where $I_\lambda$ is defined in (6.9).

(We also give a matching logarithmic lower bound on the size of $C_{\text{max}}$ in Theorem 6.36.) In the supercritical case, that is, when $\lambda > 1$, we prove the existence of a unique connected component of size linear in $n$, which is referred to as the giant component.

**Theorem 6.29** (Supercritical regime: giant component). Let $G_n \sim \mathbb{G}_{n,p_n}$ where $p_n = \frac{\lambda}{n}$ with $\lambda > 1$. For any $\gamma \in (1/2, 1)$ and $\delta < 2\gamma - 1$,

$$
\Pr_{n,p_n}[|C_{\text{max}} - \zeta \lambda n| \geq n^\gamma] \leq O(n^{-\delta}).
$$

In fact, with probability $1 - o(1)$, there is a unique largest component and the second largest cluster has size $\Omega(\log n)$.

See Figure 6.3 for an illustration.

At a high level, the proof goes as follows:

- **(Subcritical regime)** In the subcritical case, we use an exploration process and a domination argument to approximate the size of the connected components with the progeny of a branching process. The result then follows from the hitting-time theorem and the Poisson tail above.

- **(Supercritical regime)** In the supercritical case, a similar argument gives a bound on the expected size of the giant component, which is related to the survival of the branching process. Chebyshev’s inequality gives concentration. The hard part there is to bound the variance.

**Exploration process** For a vertex $v \in [n]$, let $C_v$ be the connected component containing $v$, also referred to as the *cluster* of $v$. To analyze the size of $C_v$, we introduce a natural procedure to explore $C_v$ and show that it is dominated above and below by branching processes. This procedure is similar to the exploration process defined in Section 6.2.1.

The exploration process started at $v$ has 3 types of vertices:
- $A_t$: active vertices,
- $E_t$: explored vertices,
- $N_t$: neutral vertices.

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

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

We imagine revealing the edges of $G_n$ as they are encountered in the exploration process and we let $(F_t)$ be the corresponding filtration. In words, starting with $v$, the cluster of $v$ is progressively grown by adding to it at each time a vertex adjacent to one of the previously explored vertices and uncovering its neighbors in $G_n$. In this process, $E_t$ is the set of previously explored vertices and $A_t$—the frontier of the process—is the set of vertices who are known to belong to $C_v$ but whose full neighborhood is waiting to be uncovered. The rest of the vertices form the set $N_t$. See Figure 6.4.
Let $A_t := |A_t|$, $E_t := |\mathcal{E}_t|$, and $N_t := |\mathcal{N}_t|$. Note that $(E_t)$ is non-decreasing while $(N_t)$ is non-increasing. Let

$$\tau_0 := \inf\{t \geq 0 : A_t = 0\}.$$ 

The process is fixed for all $t > \tau_0$. Notice that $E_t = t$ for all $t \leq \tau_0$, as exactly one vertex is explored at each time until the set of active vertices is empty. Moreover, for all $t$, $(A_t, \mathcal{E}_t, \mathcal{N}_t)$ forms a partition of $[n]$ so

$$A_t + t + N_t = n, \quad \forall t \leq \tau_0.$$ (6.10)

Hence, in tracking the size of the exploration process, we can work alternatively with $A_t$ or $N_t$. Specifically, the size of the cluster of $v$ can be characterized as follows.

**Lemma 6.30.**

$$\tau_0 = |\mathcal{C}_v|.$$ 

**Proof.** Indeed a single vertex of $\mathcal{C}_v$ is explored at each time until all of $\mathcal{C}_v$ has been visited. At that point, $A_t$ is empty. 

The processes $(A_t)$ and $(N_t)$ admit a simple recursive form. Conditioning on $\mathcal{F}_{t-1}$:
- (Active vertices) If $A_{t-1} = 0$, the exploration process has finished its course and $A_t = 0$. Otherwise, (a) one active vertex becomes an explored vertex and (b) its neutral neighbors become active vertices. That is,

$$A_t = A_{t-1} + 1_{\{A_{t-1} > 0\}} \left[ \frac{1}{(a)} + Z_t \right], \quad (6.11)$$

where $Z_t$ is binomial with parameters $N_{t-1} = n - (t - 1) - A_{t-1}$ and $p_n$. For the coupling arguments below, it will be useful to think of $Z_t$ as a sum of independent Bernoulli variables. That is, let $(I_{t,j} : t \geq 1, j \geq 1)$ be an array of independent, identically distributed $\{0, 1\}$-variables with $P[I_{1,1} = 1] = p_n$. We write

$$Z_t = \sum_{i=1}^{N_{t-1}} I_{t,i}. \quad (6.12)$$

- (Neutral vertices) Similarly, if $A_{t-1} > 0$, i.e. $N_{t-1} < n - (t - 1)$, $Z_t$ neutral vertices become active vertices. That is,

$$N_t = N_{t-1} - 1_{\{N_{t-1} < n - (t - 1)\}} Z_t. \quad (6.13)$$

**Branching process arguments** With these observations, we now relate the cluster size of $v$ to the total progeny of a certain branching process. This is the key lemma.

**Lemma 6.31** (Cluster size: branching process approximation). Let $G_n \sim G_{n,p_n}$ where $p_n = \frac{\lambda}{n}$ with $\lambda > 0$ and let $C_v$ be the connected component of $v \in [n]$. Let $W_\lambda$ be the total progeny of a branching process with offspring distribution $\text{Poi}(\lambda)$. Then, for $k_n = o(\sqrt{n})$,

$$P[W_\lambda \geq k_n] - O\left( \frac{k_n^2}{n} \right) \leq P_{n,p_n}[|C_v| \geq k_n] \leq P[W_\lambda \geq k_n].$$

Before proving the lemma, recall the following simple domination results from Chapter 4:

- (Binomial domination) We have

$$n \geq m \implies \text{Bin}(n, p) \succeq \text{Bin}(m, p). \quad (6.14)$$

The binomial distribution is also dominated by the Poisson distribution in the following way:

$$\lambda \in (0, 1) \implies \text{Poi}(\lambda) \succeq \text{Bin} \left( n - 1, \frac{\lambda}{n} \right). \quad (6.15)$$

For the proofs, see Examples 4.24 and 4.28.
We use these domination results to relate the size of the connected components to the progeny of the branching process.

**Proof of Lemma 6.31.** We start with the upper bound.

**Upper bound:** Because $N_{t-1} = n - (t - 1) - A_{t-1} \leq n - 1$, conditioned on $\mathcal{F}_{t-1}$, the following stochastic domination relations hold

$$\text{Bin} \left( \frac{\lambda}{n} \right) \preceq \text{Bin} \left( n - 1, \frac{\lambda}{n} \right) \preceq \text{Poi}(\lambda),$$

by (6.14) and (6.15). Observe that the r.h.s. does not depend on $N_{t-1}$. Let $(Z_{t} \succ^{\omega} A_{t})$ be a sequence of independent $\text{Poi}(\lambda)$. Using the coupling in Example 4.28, we can couple the processes $(I_{t,j})_{j}$ and $(Z_{t} \succ^{\omega} A_{t})$ in such way that $Z_{t} \succ^{\omega} A_{t} \geq \sum_{j=1}^{n-1} I_{t,j}$ a.s. for all $t$. Then by induction on $t$, for all $t$, $0 \leq A_{t} \leq A_{t} \succ^{\omega} A_{t}$ a.s. where we define

$$A_{t} \succ^{\omega} := A_{t-1} + 1 \{ A_{t-1} > 0 \} \left[ -1 + Z_{t} \succ^{\omega} \right],$$

with $A_{0} \succ^{\omega} := 1$. (In fact, this is a domination of Markov transition matrices, as defined in Definition 4.36.) In words, $(A_{t})$ is stochastically dominated by the exploration process of a branching process with offspring distribution $\text{Bin}(n - k_{n}, p_{n})$. As a result, letting

$$\tau_{0}^{\omega} := \inf \{ t \geq 0 : A_{t} \succ^{\omega} = 0 \},$$

be the total progeny of the branching process, we immediately get

$$P_{n,p_{n}}[|C_{0}| \geq k_{n}] = P_{n,p_{n}}[\tau_{0} \geq k_{n}] \leq P[\tau_{0}^{\omega} \geq k_{n}] = P[W_{\lambda} \geq k_{n}].$$

**Lower bound:** In the other direction, we proceed in two steps. We first show that, up to a certain time, the process is bounded from below by a branching process with binomial offspring distribution. In a second step, we show that this binomial branching process can be approximated by a Poisson branching process.

1. **(Domination from below)** Let $A_{t} \prec^{\omega}$ be defined as

$$A_{t} \prec^{\omega} := A_{t-1} + 1 \{ A_{t-1} > 0 \} \left[ -1 + Z_{t} \prec^{\omega} \right],$$

with $A_{0} \prec^{\omega} := 1$, where

$$Z_{t} \prec^{\omega} := \sum_{i=1}^{n-k_{n}} I_{t,j}.$$

Note that $(A_{t} \prec^{\omega})$ is the size of the active set in the exploration process of a branching process with offspring distribution $\text{Bin}(n - k_{n}, p_{n})$. Let

$$\tau_{0}^{\omega} := \inf \{ t \geq 0 : A_{t} \prec^{\omega} = 0 \},$$
be the total progeny of this branching process. We claim that \( A_t \) is bounded from below by \( A_t^\prec \) up to time

\[
\sigma_{n-k_n} := \inf\{ t \geq 0 : N_t \leq n - k_n \}.
\]

Indeed, for all \( t \leq \sigma_{n-k_n} \), \( N_{t-1} > n - k_n \). Hence, by (6.12) and (6.18), \( Z_t \geq Z_t^\prec \) for all \( t \leq \sigma_{n-k_n} \) and as a result, by induction on \( t \),

\[
A_t \geq A_t^\prec, \quad \forall t \leq \sigma_{n-k_n}.
\]

Because the inequality between \( A_t \) and \( A_t^\prec \) holds only up to time \( \sigma_{n-k_n} \), we cannot compare directly \( \tau_0 \) and \( \tau_0^\prec \). However, observe that the size of the cluster of \( v \) is at least the total number of active and explored vertices at any time \( t \); in particular, when \( \sigma_{n-k_n} = +\infty \),

\[
|C_v| \geq A_{\sigma_{n-k_n}} + E_{\sigma_{n-k_n}} = n - N_{\sigma_{n-k_n}} \geq k_n.
\]

On the other hand, when \( \sigma_{n-k_n} = +\infty \), \( N_t > n - k_n \) for all \( t \)—in particular for all \( t \geq \tau_0 \)—and therefore \( |C_v| < k_n \). Moreover in that case, because \( A_t \geq A_t^\prec \) for all \( t \), it holds in addition that \( \tau_0^\prec \leq \tau_0 < k_n \). To sum up, we have proved the implications

\[
\tau_0^\prec \geq k_n \implies \sigma_{n-k_n} < +\infty \implies \tau_0 \geq k_n.
\]

In particular,

\[
\mathbb{P}[\tau_0^\prec \geq k_n] \leq \mathbb{P}_{n,p_n}[\tau_0 \geq k_n]. \quad (6.19)
\]

2. (Poisson approximation) By Theorem 6.21,

\[
\mathbb{P}[\tau_0^\prec = t] = \frac{1}{t!}\mathbb{P}\left[ \sum_{i=1}^{t} Z_i^\prec = t-1 \right], \quad (6.20)
\]

where the \( Z_i^\prec \)'s are independent Bin\((n - k_n, p_n)\). Note that \( \sum_{i=1}^{t} Z_i^\prec \sim \text{Bin}(t(n-k_n), p_n) \). Recall the definition of \((Z_t^\prec)\) from (6.16). By Example 4.10, Theorem 4.16, and the triangle inequality for total variation distance,

\[
\left| \mathbb{P}\left[ \sum_{i=1}^{t} Z_i^\prec = t-1 \right] - \mathbb{P}\left[ \sum_{i=1}^{t} Z_i^\succ = t-1 \right] \right| \\
\leq \frac{1}{2} t(n-k_n)(-\log(1-p_n))^2 + [t\lambda - t(n-k_n)(-\log(1-p_n))] \\
\leq \frac{1}{2} tn \left( \frac{\lambda}{n} + O(n^{-2}) \right)^2 + \left[ t\lambda - t(n-k_n) \left( \frac{\lambda}{n} + O(n^{-2}) \right) \right] \\
= O \left( \frac{tk_n}{n} \right).
\]
So by (6.20)

\[ P[\tau_0^\prec \geq k_n] = 1 - P[\tau_0^\prec < k_n] \]

\[ = 1 - P[\tau_0^\prec < k_n] + O\left(\frac{k_n^2}{n}\right) \]

\[ = P[\tau_0^\prec \geq k_n] + O\left(\frac{k_n^2}{n}\right). \]

Plugging this approximation back into (6.19) gives

\[ P_{n,p_n}[|C_v| \geq k_n] = P_{n,p_n}[\tau_0 \geq k_n] \]

\[ \geq P[\tau_0^\prec \geq k_n] - O\left(\frac{k_n^2}{n}\right) \]

\[ = P[W_\lambda \geq k_n] - O\left(\frac{k_n^2}{n}\right). \]

\[ \square \]

**Remark 6.32.** In fact one can get a slightly better lower bound. See Exercise 6.7.

When \( k_n \) is large, the branching process approximation above is not as accurate because of the saturation effect: an Erdős-Rényi graph has a finite pool of vertices from which to draw edges; as the number of neutral vertices decreases, so does the expected number of uncovered edges at each time. Instead we use the following lemma.

**Lemma 6.33** (Cluster size: saturation). Let \( G_n \sim \mathcal{G}_{n,p_n} \) where \( p_n = \frac{\lambda}{n} \) with \( \lambda > 0 \) and let \( C_v \) be the connected component of \( v \in [n] \). Let \( Y_t \sim \text{Bin}(n - 1, 1 - (1 - p_n)^t) \). Then, for any \( t \),

\[ P_{n,p_n}[|C_v| = t] \leq P[Y_t = t - 1]. \]

**Proof.** We work with neutral vertices. By Lemma 6.30 and Equation (6.10), for any \( t \),

\[ P_{n,p_n}[|C_v| = t] = P_{n,p_n}[\tau_0 = t] \leq P_{n,p_n}[N_t = n - t]. \]  

(6.21)

Recall that \( N_0 = n - 1 \) and

\[ N_t = N_{t-1} - 1_{\{N_{t-1} < n-(t-1)\}} \sum_{i=1}^{N_{t-1}} I_{t,i}. \]  

(6.22)
It is easier to consider the process without the indicator as it has a simple distribution. Define \( N_0^0 := n - 1 \) and

\[
N_t^0 := N_{t-1}^0 - \sum_{i=1}^{N_{t-1}^0} I_{t,i}, \tag{6.23}
\]

and observe that \( N_t \geq N_t^0 \) for all \( t \), as the two processes agree up to time \( \tau_0 \) at which point \( N_t \) stays fixed. The interpretation of \( N_t^0 \) is straightforward: starting with \( n - 1 \) vertices, at each time each remaining vertex is discarded with probability \( p_n \). Hence, the number of surviving vertices at time \( t \) has distribution

\[
N_t^0 \sim \text{Bin}(n - 1, (1 - p_n)^t),
\]

by the independence of the steps. Arguing as in (6.21),

\[
\mathbb{P}_{n,p_n}[|C_v| = t] \leq \mathbb{P}_{n,p_n}[N_t^0 = n - t]
= \mathbb{P}_{n,p_n}[(n - 1) - N_t^0 = t - 1]
= \mathbb{P}[Y_t = t - 1].
\]

which concludes the proof.

Combining the previous lemmas we get:

**Lemma 6.34** (Bound on the cluster size). Let \( G_n \sim G_{n,p_n} \) where \( p_n = \frac{\lambda}{n} \) with \( \lambda > 0 \) and let \( C_v \) be the connected component of \( v \in [n] \).

- (Subcritical case) Assume \( \lambda \in (0, 1) \). For all \( \kappa > 0 \),

\[
\mathbb{P}_{n,p_n}[|C_v| > (1 + \kappa)I_{\lambda}^{-1}\log n] = O(n^{-(1+\kappa)}).
\]

- (Supercritical case) Assume \( \lambda > 1 \). Let \( \zeta_\lambda \) be the unique solution in \((0, 1)\) to the fixed point equation

\[
1 - e^{-\lambda \zeta} = \zeta.
\]

Note that, Example 6.11, \( \zeta_\lambda \) is the survival probability of a branching process with offspring distribution Pois(\( \lambda \)). For any \( \kappa > 0 \),

\[
\mathbb{P}_{n,p_n}[|C_v| > (1 + \kappa)I_{\lambda}^{-1}\log n] = \zeta_\lambda + O\left(\frac{\log^2 n}{n}\right).
\]

Moreover, for any \( \alpha < \zeta_\lambda \) and any \( \delta > 0 \), there exists \( \kappa_{\delta,\alpha} > 0 \) large enough so that

\[
\mathbb{P}_{n,p_n}[I_{\lambda}^{-1}\log n \leq |C_v| \leq \alpha n] = O(n^{-(1+\delta)}). \tag{6.24}
\]
Proof. In both cases we use Lemma 6.31. To apply the lemma we need to bound the tail of the progeny $W_\lambda$ of a Poisson branching process. Using the notation of Lemma 6.31, by Theorem 6.21,

$$\mathbb{P}(W_\lambda > k_n) = \mathbb{P}(W_\lambda = +\infty) + \sum_{t > k_n} \frac{1}{t^\lambda} \mathbb{P}\left(\sum_{i=1}^{t} Z_i^\succ = t - 1\right), \quad (6.25)$$

where the $Z_i^\succ$ are i.i.d. Poi($\lambda$). Both terms on the r.h.s. depend on whether or not the mean $\lambda$ is smaller or larger than 1. We start with the first term. When $\lambda < 1$, the Poisson branching process goes extinct with probability 1. Hence $\mathbb{P}(W_\lambda = +\infty) = 0$. When $\lambda > 1$ on the other hand, $\mathbb{P}(W_\lambda = +\infty) = \zeta_\lambda$, where $\zeta_\lambda > 0$ is the survival probability of the branching process. As to the second term, the sum of the $Z_i^\succ$ is $\lambda t$. When $\lambda < 1$, using (6.7),

$$\sum_{t > k_n} \frac{1}{t^\lambda} \mathbb{P}\left(\sum_{i=1}^{t} Z_i^\succ = t - 1\right) \leq \sum_{t > k_n} \mathbb{P}\left(\sum_{i=1}^{t} Z_i^\succ \geq t - 1\right) \leq \sum_{t > k_n} \exp\left(-\left(tI_\lambda^{\text{Poi}}\left(\frac{t-1}{t}\right)\right)\right) \leq \sum_{t > k_n} \exp\left(-t(I_\lambda - O(t^{-1}))\right) \leq \sum_{t > k_n} C' \exp\left(-tI_\lambda\right) \leq C \exp\left(-I_\lambda k_n\right), \quad (6.26)$$

for some constants $C, C' > 0$, where we assume that $k_n = \omega(1)$. When $\lambda > 1$,

$$\sum_{t > k_n} \frac{1}{t^\lambda} \mathbb{P}\left(\sum_{i=1}^{t} Z_i^\succ = t - 1\right) \leq \sum_{t > k_n} \mathbb{P}\left(\sum_{i=1}^{t} Z_i^\succ \leq t\right) \leq \sum_{t > k_n} \exp\left(-tI_\lambda\right) \leq C \exp\left(-I_\lambda k_n\right), \quad (6.27)$$

for a possibly different $C > 0$.

**Subcritical case:** Assume $0 < \lambda < 1$ and let $c = (1 + \kappa)I_\lambda^{-1}$ for $\kappa > 0$. By Lemma 6.31,

$$\mathbb{P}_{n,p_n} \left[|C_1| > c \log n\right] \leq \mathbb{P}[W_\lambda > c \log n].$$

336
By (6.25) and (6.26),
\[
P[W_\lambda > c \log n] = O (\exp (-I_\lambda c \log n)),
\] (6.28)
which proves the claim.

**Supercritical case:** Now assume \( \lambda > 1 \) and again let \( c = (1 + \kappa)I_\lambda^{-1} \) for \( \kappa > 0 \).

By Lemma 6.31,
\[
P_{n,p_n} [ |C_v| > c \log n] = P[W_\lambda > c \log n] + O \left( \frac{\log^2 n}{n} \right),
\] (6.29)

By (6.25) and (6.27),
\[
P[W_\lambda > c \log n] = \zeta_\lambda + O (\exp (-cI_\lambda \log n)) = \zeta_\lambda + O (n^{-(1+\kappa)}).
\] (6.30)

Combining (6.29) and (6.30), for any \( \kappa > 0 \),
\[
P_{n,p_n} [ |C_v| > c \log n] = \zeta_\lambda + O \left( \frac{\log^2 n}{n} \right).
\] (6.31)

Next, we show that in the supercritical case when \( |C_v| > c \log n \) the cluster size is in fact linear in \( n \) with high probability. By Lemma 6.33
\[
P_{n,p_n} [ |C_v| = t] \leq P[Y_t = t - 1] \leq P[Y_t \leq t],
\]
where \( Y_t \sim \text{Bin}(n-1, 1 - (1 - p_n)^t) \). Roughly, the r.h.s. is negligible until the mean \( \mu_t := (n-1)(1 - (1 - \lambda/n)^t) \) is of the order of \( t \). Let \( \zeta_\lambda \) be the unique solution in \( (0,1) \) to the fixed point equation
\[
f(\zeta) := 1 - e^{-\lambda \zeta} = \zeta.
\]
The solution is unique because \( f(0) = 0, f(1) < 1 \), and the \( f \) is increasing, strictly concave and has derivative \( \lambda > 1 \) at 0. Note in particular that, when \( t = \zeta_\lambda n \), \( \mu_t \approx t \). Let \( \alpha < \zeta_\lambda \). For any \( t \in [c \log n, \alpha n] \), by a Chernoff bound for Poisson trials (Theorem 2.37),
\[
P[Y_t \leq t] \leq \exp \left( -\frac{\mu_t}{2} \left( 1 - \frac{t}{\mu_t} \right)^2 \right).
\] (6.32)
For \( t/n \leq \alpha < \zeta \), using \( 1 - x \leq e^{-x} \) for \( x \in (0, 1) \), there is \( \gamma_{\alpha} > 1 \) such that

\[
\mu_t \geq (n-1)(1-e^{-\lambda(t/n)}) = t \left( \frac{n-1}{n} \right) \frac{1-e^{-\lambda(t/n)}}{t/n} = t \left( \frac{n-1}{n} \right) f(t/n) \frac{t/n}{t/n} \geq t \left( \frac{n-1}{n} \right) \frac{1-e^{-\lambda\alpha}}{\alpha} \geq \gamma_{\alpha} t,
\]

for \( n \) large enough, by the properties of \( f \) mentioned above. Plugging this back into (6.32), we get

\[
P[Y_t \leq t] \leq \exp \left( -t \left( \frac{\gamma_{\alpha}}{2} \left( 1 - \frac{1}{\gamma_{\alpha}} \right)^2 \right) \right).
\]

Therefore

\[
\sum_{t=c \log n}^{\alpha n} \mathbb{P}_{n,p_n}[|C_v| = t] \leq \sum_{t=c \log n}^{\alpha n} \mathbb{P}[Y_t \leq t] \leq \sum_{t=c \log n}^{+\infty} \exp \left( -t \left( \frac{\gamma_{\alpha}}{2} \left( 1 - \frac{1}{\gamma_{\alpha}} \right)^2 \right) \right) = O \left( \exp \left( -c \log n \left( \frac{\gamma_{\alpha}}{2} \left( 1 - \frac{1}{\gamma_{\alpha}} \right)^2 \right) \right) \right).
\]

Taking \( \kappa > 0 \) large enough proves (6.24).

Let \( C_{\text{max}} \) be the largest connected component of \( G_n \) (choosing the component containing the lowest label if there is more than one such component). Our goal is to characterize the size of \( C_{\text{max}} \). Let

\[
X_k := \sum_{v \in [n]} 1_{\{|C_v| > k\}},
\]

be the number of vertices in clusters of size at least \( k \). There is a natural connection between \( X_k \) and \( C_{\text{max}} \), namely,

\[
|C_{\text{max}}| > k \iff X_k > 0 \iff X_k > k.
\]

338
A first moment argument on $X_k$ and the previous lemma immediately imply an upper bound on the size of $C_{\text{max}}$ in the subcritical case.

**Proof of Theorem 6.28.** Let $c = (1 + \kappa)I^{-1}_\lambda$ for $\kappa > 0$. We use the first moment method on $X_k$. By symmetry and the first moment method (Corollary ??),

$$
\mathbb{P}_{n,p_n} [C_{\text{max}} > c \log n] = \mathbb{P}_{n,p_n} [X_{c \log n} > 0] \\
\leq \mathbb{E}_{n,p_n} [X_{c \log n}] \\
= n \mathbb{P}_{n,p_n} [|C_1| > c \log n].
$$

(6.33)

By Lemma 6.34,

$$
\mathbb{P}_{n,p_n} [C_{\text{max}} > c \log n] = O(n \cdot n^{-(1+\kappa)}) = O(n^{-\kappa}) \to 0,
$$
as $n \to +\infty$.

In fact we prove below that the largest component is of size roughly $I^{-1}_\lambda \log n$. But first we turn to the supercritical regime.

**Second moment arguments** To characterize the size of the largest cluster in the supercritical case, we need a second moment argument. Assume $\lambda > 1$. For $\delta > 0$ and $\alpha < \zeta_\lambda$, let $\kappa_{\delta,\alpha}$ be as defined in Lemma 6.34. Set

$$
\kappa_n := (1 + \kappa_{\delta,\alpha})I^{-1}_\lambda \log n \quad \text{and} \quad \kappa_n := \alpha n.
$$

We call a vertex $v$ such that $|C_v| \leq \kappa_n$ a **small vertex**. Let

$$
Y_k := \sum_{v \in [n]} 1_{\{|C_v| \leq \kappa_n\}}.
$$

Then $Y_k$ is the number of small vertices. Observe that by definition $Y_k = n - X_k$. Hence by Lemma 6.34, the expectation of $Y_k$ is

$$
\mathbb{E}_{n,p_n}[Y_k] = n(1 - \mathbb{P}_{n,p_n} [\{|C_v| > \kappa_n\}]) = (1 - \zeta_\lambda)n + O(\log^2 n).
$$

(6.34)

Using Chebyshev’s inequality (Theorem 2.2), we prove that $Y_k$ is close to its expectation:

**Lemma 6.35 (Concentration of $Y_k$).** For any $\gamma \in (1/2, 1)$ and $\delta < 2\gamma - 1$,

$$
\mathbb{P}_{n,p_n} [Y_k - (1 - \zeta_\lambda)n \geq n^\gamma] = O(n^{-\delta}).
$$

Lemma 6.35, which is proved below, leads to our main result in the supercritical case: the existence of the **giant component**, a unique cluster of size linear in $n$. 

339
Proof of Theorem 6.29. Take $\alpha \in (\zeta_\lambda/2, \zeta_\lambda)$ and let $k_n$ and $\bar{k}_n$ be as above. Let $B_{1,n} := \{|X_{k_n} - \zeta_\lambda n| \geq n^\gamma\}$. Because $\gamma < 1$, for $n$ large enough, the event $B_{1,n}$ implies that $X_{k_n} \geq 1$ and, in particular, that

$$|C_{\text{max}}| \leq X_{k_n}.$$ 

Let $B_{2,n} := \{\exists v, |C_v| \in [k_n, \bar{k}_n]\}$. If, in addition to $B_{1,n}$, $B_{2,n}$ also holds then

$$|C_{\text{max}}| \leq X_{k_n} = X_{\bar{k}_n}.$$ 

There is equality in the last display if there is a unique cluster of size greater than $\bar{k}_n$. This is indeed the case under $B_{1,n} \cap B_{2,n}$: if there were two distinct clusters of size $\bar{k}_n$, then since $2\alpha > \zeta_\lambda$ we would have for $n$ large enough

$$X_{k_n} = X_{\bar{k}_n} > 2\bar{k}_n = 2\alpha n > \zeta_\lambda n + n^\gamma,$$

a contradiction. Hence we have proved that, under $B_{1,n} \cap B_{2,n}$, we have

$$|C_{\text{max}}| = X_{k_n} = X_{\bar{k}_n}.$$ 

Applying Lemmas 6.34 and 6.35 concludes the proof.

It remains to prove Lemma 6.35.

Proof of Lemma 6.35. The main task is to bound the variance of $Y_{k_n}$. Note that

$$\mathbb{E}_{n,p_n}[Y_{k_n}^2] = \sum_{u,v \in [n]} \mathbb{P}_{n,p_n}[|C_u| \leq k, |C_v| \leq k]$$

$$= \sum_{u,v \in [n]} \left\{ \mathbb{P}_{n,p_n}[|C_u| \leq k, |C_v| \leq k, u \leftrightarrow v] + \mathbb{P}_{n,p_n}[|C_u| \leq k, |C_v| \leq k, u \not\leftrightarrow v] \right\}, \quad (6.35)$$

where $u \leftrightarrow v$ indicates that $u$ and $v$ are in the same connected component.

To bound the first term in (6.35), we note that $u \leftrightarrow v$ implies that $C_u = C_v$.

Hence,

$$\sum_{u,v \in [n]} \mathbb{P}_{n,p_n}[|C_u| \leq k, |C_v| \leq k, u \leftrightarrow v] = \sum_{u,v \in [n]} \mathbb{P}_{n,p_n}[|C_u| \leq k, v \in C_u]$$

$$= \sum_{u,v \in [n]} \mathbb{E}_{n,p_n}[1_{|C_u| \leq k} \mathbf{1}_{v \in C_u}]$$

$$= \sum_{u \in [n]} \mathbb{E}_{n,p_n}[|C_u| \mathbf{1}_{|C_u| \leq k}]$$

$$= n \mathbb{E}_{n,p_n}[|C_1| \mathbf{1}_{|C_1| \leq k}]$$

$$\leq nk. \quad (6.36)$$
To bound the second term in (6.35), we sum over the size of $C_u$ and note that, conditioned on $\{|C_u| = \ell, u \leftrightarrow v\}$, the size of $C_v$ has the same distribution as the unconditional size of $C_1$ in a $G_{n-\ell,p_n}$ random graph, that is,

$$P_{n,p_n}[|C_v| \leq k | |C_u| = \ell, u \leftrightarrow v] = P_{n-\ell,p_n}[|C_1| \leq k].$$

Observe that the probability on the l.h.s. is increasing in $\ell$. Hence

$$\sum_{u,v \in [n]} \sum_{\ell \leq k} P_{n,p_n}[|C_u| = \ell, |C_v| \leq k, u \leftrightarrow v]$$

$$= \sum_{u,v \in [n]} \sum_{\ell \leq k} P_{n,p_n}[|C_u| = \ell, u \leftrightarrow v] P_{n,p_n}[|C_v| \leq k | |C_u| = \ell, u \leftrightarrow v]$$

$$\leq \sum_{u,v \in [n]} \sum_{\ell \leq k} P_{n,p_n}[|C_u| = \ell] P_{n-k,p_n}[|C_v| \leq k]$$

$$= \sum_{u,v \in [n]} P_{n,p_n}[|C_u| \leq k] P_{n-k,p_n}[|C_v| \leq k].$$

To get a bound on the variance of $Y_k$, we need to relate this last expression to $(E_{n,p_n}[Y_k])^2$. For this purpose we define

$$\Delta_k := P_{n-k,p_n}[|C_1| \leq k] - P_{n,p_n}[|C_1| \leq k].$$

Then, plugging this back above, we get

$$\sum_{u,v \in [n]} \sum_{\ell \leq k} P_{n,p_n}[|C_u| = \ell, |C_v| \leq k, u \leftrightarrow v]$$

$$\leq \sum_{u,v \in [n]} P_{n,p_n}[|C_u| \leq k](P_{n,p_n}[|C_v| \leq k] + \Delta_k)$$

$$\leq (E_{n,p_n}[Y_k])^2 + n^2|\Delta_k|,$n

and it remains to bound $\Delta_k$. We use a coupling argument. Let $H \sim G_{n-\ell,p_n}$ and construct $H' \sim G_{n,p_n}$ in the following manner: let $H'$ coincide with $H$ on the first $n-\ell$ vertices then pick the rest the edges independently. Then clearly $\Delta_k \geq 0$ since the cluster of 1 in $H'$ includes the cluster of 1 in $H$. In fact, $\Delta_k$ is the probability that under this coupling the cluster of 1 has at most $k$ vertices in $H$ but not in $H'$. That implies in particular that at least one of the vertices in the cluster of 1 in $H$ is connected to a vertex in $\{n-k+1, \ldots, n\}$. Hence by a union bound over those edges

$$\Delta_k \leq k^2 p_n,$$
and

\[ \sum_{u,v \in [n]} \mathbb{P}_{n,p_n}[|C_u| \leq k, |C_v| \leq k, u \leftrightarrow v] \leq (\mathbb{E}_{n,p_n}[Y_k])^2 + \lambda k^2 n. \] (6.37)

Combining (6.36) and (6.37), we get

\[ \text{Var}[Y_k] \leq 2\lambda k^2 n. \]

The result follows from Chebyshev’s inequality (Theorem 2.2) and Equation (6.34).

A similar second moment argument also gives a lower bound on the size of the largest component in the subcritical case. We proved in Theorem 6.28 that, when \( \lambda < 1 \), the probability of observing a connected component of size significantly larger than \( I^{-1}_\lambda \log n \) is vanishingly small. In the other direction, we get:

**Theorem 6.36** (Subcritical regime: lower bound on the largest cluster). Let \( G_n \sim G_{n,p_n} \) where \( p_n = \frac{\lambda}{n} \) with \( \lambda \in (0,1) \). For all \( \kappa \in (0,1) \),

\[ \mathbb{P}_{n,p_n}[|C_{\text{max}}| \leq (1 - \kappa)I^{-1}_\lambda \log n] = o(1). \]

**Proof.** Recall that

\[ X_k := \sum_{v \in [n]} 1_{\{|C_v| > k\}}. \]

It suffices to prove that with probability \( 1 - o(1) \) we have \( X_k > 0 \) when \( k = (1 - \kappa)I^{-1}_\lambda \log n \). To apply the second moment method, we need an upper bound on the second moment of \( X_k \) and a lower bound on its first moment. The following lemma is closely related to Lemma 6.35. Exercise 6.8 asks for a proof.

**Lemma 6.37** (Second moment of \( X_k \)). Assume \( \lambda < 1 \). There is \( C > 0 \) such that

\[ \mathbb{E}_{n,p_n}[X_k^2] \leq (\mathbb{E}_{n,p_n}[X_k])^2 + C nke^{-kI_\lambda}, \quad \forall k \geq 0. \]

**Lemma 6.38** (First moment of \( X_k \)). Let \( k_n = (1 - \kappa)I^{-1}_\lambda \log n \). Then, for any \( \beta \in (0, \kappa) \) we have that

\[ \mathbb{E}_{n,p_n}[X_{k_n}] = \Omega(n^\beta), \]

for \( n \) large enough.
Proof. By Lemma 6.31,
\[
\mathbb{E}_{n,p_n}[X_{k_n}] = n \mathbb{P}_{n,p_n}|C_1| > k_n
= n \mathbb{P}[W > k_n] - O \left( k_n^2 \right). \tag{6.38}
\]

Once again, we use the random-walk representation of the total progeny of a branching process (Theorem 6.21). Using the notation of Lemma 6.31,
\[
\mathbb{P}[W > k_n] = \sum_{t > k_n} \frac{1}{t} \mathbb{P} \left[ \sum_{i=1}^{t} Z_i^+ = t - 1 \right]
= \sum_{t > k_n} \frac{1}{t} e^{-\lambda t} \frac{(\lambda t)^{t-1}}{(t-1)!}.
\]

Using Stirling’s formula, we note that
\[
e^{-\lambda t} \frac{(\lambda t)^{t-1}}{t!} = e^{-\lambda t} \frac{(\lambda t)^t}{\lambda t (t/e)^t \sqrt{2\pi t} (1 + o(1))}
= \frac{1 + o(1)}{\lambda t \sqrt{2\pi t}} \exp (-t\lambda + t \log \lambda + t)
= \frac{1 + o(1)}{\lambda \sqrt{2\pi t^3}} \exp (-tI_\lambda).
\]

For any \( \epsilon > 0 \), for \( n \) large enough,
\[
\mathbb{P}[W > k_n] \geq \lambda^{-1} \sum_{t > k_n} \exp (-t(I_\lambda + \epsilon))
= \Omega (\exp (-k_n(I_\lambda + \epsilon))).
\]

For any \( \beta \in (0, \kappa) \), plugging the last line back into (6.38) and taking \( \epsilon \) small enough gives
\[
\mathbb{E}_{n,p_n}[X_{k_n}] = \Omega (n \exp (-k_n(I_\lambda + \epsilon)))
= \Omega \left( \exp \left( \{1 - (1 - \kappa)I_\lambda^{-1} (I_\lambda + \epsilon)\} \log n \right) \right)
= \Omega(n^\beta).
\]

We return to the proof of Theorem 6.28. Let \( k_n = (1 - \kappa)I_\lambda^{-1} \log n \). By the
second moment method (Theorem 2.20) and Lemmas 6.37 and 6.38,

\[ P_{n,p} [X_{k_n} > 0] \geq \frac{(\mathbb{E}X_{k_n})^2}{\mathbb{E}[X_{k_n}^2]} \]

\[ \geq \left( 1 + O(nk_ne^{-\kappa_\lambda}) \frac{1}{\Omega(n^{2\beta})} \right)^{-1} \]

\[ = \left( 1 + O(k_ne^{\kappa \log n}) \frac{1}{\Omega(n^{2\beta})} \right)^{-1} \]

\[ \rightarrow 1, \]

for \( \beta \) close enough to \( \kappa \). That proves the claim. \( \blacksquare \)

**Critical regime via martingales** To be written. See [NP10].

**Exercises**

**Exercise 6.1** (Galton-Watson process: geometric offspring). Let \( (Z_t) \) be a Galton-Watson branching process with geometric offspring distribution (started at 0), i.e., \( p_k = p(1-p)^k \) for all \( k \geq 0 \), for some \( p \in (0,1) \). Let \( q := 1-p \), let \( m \) be the mean of the offspring distribution, and let \( M_t = m^{-t}Z_t \).

a) Compute the probability generating function \( f \) of \( \{p_k\}_{k \geq 0} \) and the extinction probability \( \eta := \eta_p \) as a function of \( p \).

b) If \( G \) is a \( 2 \times 2 \) matrix, define

\[ G(s) := \frac{G_{11}s + G_{12}}{G_{21}s + G_{22}}. \]

Show that \( G(H(s)) = (GH)(s) \).

c) Assume \( m \neq 1 \). Use b) to derive

\[ f_t(s) = \frac{pm^t(1-s)+qs-p}{qm^t(1-s)+qs-p}. \]

Deduce that when \( m > 1 \)

\[ \mathbb{E}[\exp(-\lambda M_\infty)] = \eta + (1-\eta) \frac{(1-\eta)}{\lambda + (1-\eta)}. \]
d) Assume \( m = 1 \). Show that

\[
    f_t(s) = \frac{t - (t - 1)s}{t + 1 - ts},
\]

and deduce that

\[
    \mathbb{E}[e^{-\lambda Z_t/t} | Z_t > 0] \to \frac{1}{1 + \lambda}.
\]

**Exercise 6.2** (Supercritical branching process: infinite line of descent). Let \((Z_t)\) be a supercritical Galton-Watson branching process with offspring distribution \(\{p_k\}_{k \geq 0}\). Let \(\eta\) be the extinction probability and define \(\zeta := 1 - \eta\). Let \(Z_t^\infty\) be the number of individuals in the \(t\)-th generation with an infinite line of descent, i.e., whose descendant subtree is infinite. Denote by \(S\) the event of nonextinction of \((Z_t)\). Define \(p_0^\infty := 0\) and

\[
    p_k^\infty := \zeta^{-1} \sum_{j \geq k} \binom{j}{k} \eta^{j-k} \zeta^k p_j.
\]

a) Show that \(\{p_k^\infty\}_{k \geq 0}\) is a probability distribution and compute its expectation.

b) Show that for any \(k \geq 0\)

\[
    \mathbb{P}[Z_t^\infty = k | S] = p_k^\infty.
\]

[Hint: Condition on \(Z_1\).]

c) Show by induction on \(t\) that, conditioned on nonextinction, the process \((Z_t^\infty)\) has the same distribution as a Galton-Watson branching process with offspring distribution \(\{p_k^\infty\}_{k \geq 0}\).

**Exercise 6.3** (Hitting-time theorem: nearest-neighbor walk). Let \(X_1, X_2, \ldots\) be i.i.d. random variables taking value \(+1\) with probability \(p\) and \(-1\) with probability \(q := 1 - p\). Let \(S_t := \sum_{i=1}^t X_i\) with \(S_0 := 0\) and \(M_t := \max\{S_i : 0 \leq i \leq t\}\).

a) For \(r \geq 1\), use the reflection principle to show that

\[
    \mathbb{P}[M_t \geq r, S_t = b] = \begin{cases} 
        \mathbb{P}[S_t = b], & b \geq r, \\
        (q/p)^{r-b} \mathbb{P}[S_t = 2r - b], & b < r. 
    \end{cases}
\]

b) Use a) to give a proof of the hitting-time theorem in the following special case: letting \(\tau_b\) be the first time \(S_t\) hits \(b > 0\), show that for all \(t \geq 1\)

\[
    \mathbb{P}[\tau_b = t] = \frac{b}{t} \mathbb{P}[S_t = b].
\]

[Hint: Consider the probability \(\mathbb{P}[M_{t-1} = S_{t-1} = b - 1, S_t = b]\).]
Exercise 6.4 (Percolation on bounded-degree graphs). Let $G = (V, E)$ be a countable graph such that all vertices have degree bounded by $b + 1$ for $b \geq 2$. Let 0 be a distinguished vertex in $G$. For bond percolation on $G$, prove that

$$p_c(G) \geq p_c(\hat{T}_b),$$

by bounding the expected size of the cluster of 0. [Hint: Consider self-avoiding paths started at 0.]

Exercise 6.5 (Percolation on $\hat{T}_b$: critical exponent of $\theta(p)$). Consider bond percolation on the rooted infinite $b$-ary tree $\hat{T}_b$ with $b > 2$. For $\varepsilon \in [0, 1 - \frac{1}{b}]$ and $u \in [0, 1]$, define

$$h(\varepsilon, u) := u - ((1 - \frac{1}{b} - \varepsilon)(1 - u) + \frac{1}{b} + \varepsilon)^b.$$

a) Show that there is a constant $C > 0$ not depending on $\varepsilon, u$ such that

$$\left| h(\varepsilon, u) - b\varepsilon u + \frac{b - 1}{2b} u^2 \right| \leq C(u^3 \vee \varepsilon u^2).$$

b) Use a) to prove that

$$\lim_{p \uparrow p_c(\hat{T}_b)} \frac{\theta(p)}{(p - p_c(\hat{T}_b))} = \frac{2b^2}{b - 1}.$$ 

Exercise 6.6 (Percolation on $\hat{T}_2$: higher moments of $|C_0|$). Consider bond percolation on the rooted infinite binary tree $\hat{T}_2$. For density $p < \frac{1}{2}$, let $Z_p$ be an integer-valued random variable with distribution

$$\mathbb{P}_p[Z_p = \ell] = \frac{\ell \mathbb{P}_p[|C_0| = \ell]}{\mathbb{P}_p[|C_0|]}, \quad \forall \ell \geq 1.$$

a) Using the explicit formula for $\mathbb{P}_p[|C_0| = \ell]$ derived in Section 6.3.1, show that for all $0 < a < b < +\infty$

$$\mathbb{P}_p\left[ \frac{Z_p}{(1/4)(1/2 - p)^{-2}} \in [a, b] \right] \to C \int_a^b x^{-1/2} e^{-x} dx,$$

as $p \uparrow \frac{1}{2}$, for some constant $C > 0$.

b) Show that for all $k \geq 2$ there is $C_k > 0$ such that

$$\lim_{p \uparrow p_c(\hat{T}_2)} \frac{\mathbb{E}_p[|C_0|^k]}{(p_c(\hat{T}_2) - p)^{-1 - 2(k-1)}} = C_k.$$
c) What happens when $p \downarrow p_c(\widehat{T}_2)$?

**Exercise 6.7** (Branching process approximation: improved bound). Let $p_n = \frac{\lambda}{n}$ with $\lambda > 0$. Let $W_{n,p_n}$, respectively $W_{\lambda}$, be the total progeny of a branching process with offspring distribution $\text{Bin}(n, p_n)$, respectively $\text{Poi}(\lambda)$.

a) Show that

$$\left| \mathbb{P}[W_{n,p_n} \geq k] - \mathbb{P}[W_{\lambda} \geq k] \right| \leq \max\{ \mathbb{P}[W_{n,p_n} \geq k, W_{\lambda} < k], \mathbb{P}[W_{n,p_n} < k, W_{\lambda} \geq k] \}.$$  

b) Couple the two processes step-by-step and use a) to show that

$$\left| \mathbb{P}[W_{n,p_n} \geq k] - \mathbb{P}[W_{\lambda} \geq k] \right| \leq \frac{\lambda^2}{n} \sum_{i=1}^{k-1} \mathbb{P}[W_{\lambda} \geq i].$$


**Bibliographic remarks**

**Section 6.1** See [Dur10, Section 5.3.4] for a quick introduction to branching processes. A more detailed overview relating to its use in discrete probability can be found in [vdH14, Chapter 3]. The classical reference on branching processes is [AN04]. The Kesten-Stigum theorem is due to Kesten and Stigum [KS66a, KS66b, KS67]. Our proof of a weaker version with an $L^2$ condition follows [Dur10, Example 5.4.3]. Spitzer’s combinatorial lemma is due to Spitzer [Spi56]. The proof presented here follows [Fel71, Section XII.6]. The hitting-time theorem was first proved by Otter [Ott49]. Several proofs of a generalization can be found in [Wen75]. The critical percolation threshold for percolation on Galton-Watson trees is due to R. Lyons [Lyo90].

**Section 6.3** The presentation in Section 6.3.1 follows [vdH10]. See also [Dur85]. For much more on the phase transition of Erdős-Rényi graphs, see e.g. [vdH14, Chapter 4], [JLR11, Chapter 5] and [Bol01, Chapter 6]. In particular a central limit theorem for the giant component, proved by several authors including Martin-Löf [ML98], Pittel [Pit90], and Barraez, Boucheron, and de la Vega [BBFdlV00], is established in [vdH14, Section 4.5].
Appendix A

Appendix

A.1 Some definitions and useful facts

These lemmas will eventually be inserted in the main text in a suitable place.

A.1.1 Binomial coefficients

Recall the following bounds on factorials and binomial coefficients:

\[ e \left( \frac{n}{e} \right)^n \leq n! \leq e \left( \frac{n+1}{e} \right)^{n+1} \]

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

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

and

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

where \( H(p) := -p \log p - (1 - p) \log(1 - p) \).
A.1.2 Conditional expectation: definition and properties

Recall the definition of the conditional expectation (see e.g. [Wil91, Section 9.2]).

**Theorem A.1** (Conditional expectation). Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$ a sub $\sigma$-field. Then there exists a (a.s.) unique $Y \in L^1(\Omega, \mathcal{G}, \mathbb{P})$ (note the $\mathcal{G}$-measurability) s.t.

$$E[Y; \mathcal{G}] = E[X; \mathcal{G}], \forall G \in \mathcal{G}.$$ 

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

In $L^2$ conditional expectation reduces to an orthogonal projection (see e.g. [Wil91, Section 9.4]).

**Theorem A.2** (Conditional expectation: $L^2$ case). Let $\langle X, Y \rangle := E[XY]$. Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$ a sub $\sigma$-field. Then there exists a (a.s.) unique $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ s.t.

$$\|X - Y\|_2 = \inf \{\|X - W\|_2 : W \in L^2(\Omega, \mathcal{G}, \mathbb{P})\},$$

and, moreover, $\langle Z, X - Y \rangle = 0, \forall Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$. Such $Y$ is called an orthogonal projection of $X$ on $L^2(\Omega, \mathcal{G}, \mathbb{P})$.

In addition to linearity and the usual inequalities (e.g. Jensen’s inequality, etc.) and convergence theorems (e.g. dominated convergence, etc.), we highlight the following three properties of the conditional expectation (see e.g. [Wil91, Section 9.7]).

**Lemma A.3** (Taking out what is known). If $Z \in \mathcal{G}$ is bounded then $E[ZX \mid \mathcal{G}] = Z E[X \mid \mathcal{G}]$. This is also true if $X, Z \geq 0$ and $E[ZX] < +\infty$ or $X \in L^p(\mathcal{F})$ and $Z \in L^q(\mathcal{G})$ with $p^{-1} + q^{-1} = 1$ and $p > 1$.

**Lemma A.4** (Role of independence). If $X$ is independent of $\mathcal{H}$ then $E[X \mid \mathcal{H}] = E[X]$. In fact, if $\mathcal{H}$ is independent of $\sigma(X, \mathcal{G})$, then $E[X \mid \sigma(\mathcal{G}, \mathcal{H})] = E[X \mid \mathcal{G}]$.

**Lemma A.5** (Tower property (or law of total probability)). We have $E[E[X \mid \mathcal{G}]] = E[X]$. In fact, if $\mathcal{H} \subseteq \mathcal{G}$ is a $\sigma$-field

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

That is, the smallest $\sigma$-field wins.

The following fact will also prove useful (see e.g. [Dur10, Example 5.1.5] for a proof).
Lemma A.6 (Conditioning on an independent RV). Suppose \( X \) and \( Y \) are independent. Let \( \phi \) be a function with \( \mathbb{E}|\phi(X,Y)| < +\infty \) and let \( g(x) = \mathbb{E}(\phi(x,Y)) \). Then,

\[
\mathbb{E}(\phi(X,Y)|X) = g(X).
\]

A.1.3 A Taylor expansion

To be written. See [LL10, Lemmas 12.1.1, 12.1.4].

A.1.4 Spectral representation of reversible matrices

Let \( P \) be the transition matrix of a finite, irreducible Markov chain on \( V \) reversible with respect to \( \pi \). Define \( n := |V| \). We let \( \ell^2(\pi) \) be the vector space of real-valued functions with inner product

\[
\langle f, g \rangle_\pi := \sum_{x \in V} \pi(x)f(x)g(x).
\]

Lemma A.7 (Spectral representation: reversible matrices). The space \( \ell^2(\pi) \) has an orthonormal basis of eigenfunctions \( \{f_j\}_{j=1}^n \) with real eigenvalues \( \{\lambda_j\}_{j=1}^n \) such that \( |\lambda_j| \leq 1 \), for all \( j \). The eigenfunction \( f_1 \) corresponding to the eigenvalue 1 can be taken to be the all-1 function. Furthermore, we have the following decomposition

\[
P^t(x,y) = 1 + \sum_{j=2}^n f_j(x)f_j(y)\lambda_j^t.
\]

Proof. To be written. See [LPW06, Lemma 12.2]

A.1.5 A fact about trees

Lemma A.8. A cycle-free undirected graph with \( n \) vertices and \( n - 1 \) edges is a spanning tree.
A.1.6 A Poincaré inequality

The Dirichlet form is defined as $E(f,g) := \langle f, (I - P)g \rangle_\pi$. Note that

$$2\langle f, (I - P)f \rangle_\pi = 2\langle f, f \rangle_\pi - 2\langle f, Pf \rangle_\pi = \sum_{x} \pi(x) f(x)^2 + \sum_{y} \pi(y) f(y)^2 - \sum_{x} \pi(x) f(x) f(y) P(x,y)$$

$$= \sum_{x,y} f(x)^2 \pi(x) P(x,y) + \sum_{x,y} f(y)^2 \pi(y) P(y,x) - \sum_{x} \pi(x) f(x) f(y) P(x,y)$$

$$= \sum_{x,y} \pi(x) P(x,y) [f(x) - f(y)]^2 = 2\mathcal{E}(f)$$

where

$$\mathcal{E}(f) := \frac{1}{2} \sum_{x,y} c(x,y) [f(x) - f(y)]^2,$$

is the Dirichlet energy encountered previously. We note further that if $\sum_{x} \pi(x) f(x) = 0$ then

$$\langle f, f \rangle_\pi = \langle f - \langle 1, f \rangle_\pi, f - \langle 1, f \rangle_\pi \rangle_\pi = \text{Var}_\pi[f],$$

where the last expression denotes the variance under $\pi$. So the variational characterization of $\lambda_2$ translates into

$$\text{Var}_\pi[f] \leq \gamma \mathcal{E}(f),$$

for all $f$ such that $\sum_{x} \pi(x) f(x) = 0$ (in fact for any $f$ by considering $f - \langle 1, f \rangle_\pi$ and noticing that both sides are unaffected by adding a constant), which is known as a Poincaré inequality.

**Lemma A.9 (Poincaré inequality).**

$$\text{Var}_\pi[f] \leq \gamma \mathcal{E}(f), \quad \forall f,$$

with equality for $f_2$, the eigenfunction of $P$ corresponding to the second largest eigenvalue $\lambda_2$. 

351
Bibliography


Index

ε-packing, 64, 93
, 286
adapted process, 99
approximate counting, 267
Azuma-Hoeffding inequality, 117, 119,
120, 129, 130, 136, 140, 145,
182, 288
balancing vectors, 22
ballot theorem, 103
balls and bins, 126
Bernstein’s inequality, 59–61
Berry-Esséen theorem, 128
binary classification, 83
binomial variable, 51
birth-and-death chain, 183
bond percolation, 11
trees, 114
Bonferroni inequalities, 95
Boole’s inequality, see union bound
bottleneck ratio, 290
bounded differences inequality, 124
branching number, 44, 45
branching processes
  dual branching process, 319
duality principle, 319
  exploration process, 317, 328
  extinction, 309, 311
  Galton-Watson branching process,
  344
  Galton-Watson process, 309, 317
  Galton-Watson tree, 310, 316
  infinite line of descent, 345
  Poisson offspring, 314, 320, 322
  random-walk representation, 317–
  319
Chebyshev polynomials, 79
Chebyshev’s inequality, 19, 46, 47, 49,
197, 339, 342
Cheeger’s inequality, 290
Chen-Stein
  Stein coupling, 249
  Chen-Stein method, 246, 261
dissociated case, 249
Stein coupling, 246
Chernoff bound, 52, 127
Chernoff-Cramér bound, 47, 58, 62
Chernoff-Cramér method, 17, 47, 50, 52,
70, 96, 117
chi-square variable, 71
chromatic number, 130
clique number, 35, 261
commute time, 14, 173
commute time identity, 173, 184
compressed sensing, 73
sensing matrix, 73
concentration inequalities, 98
concentration phenomenon, 117
conditional expectation
  definition, 349
connectivity, 39
Gibbs random field, 12
Glauber dynamics, 15, 242, 302
  fast mixing, 243, 302
gradient, 164
graph
  definitions, 1
Green function, 148
Hamming distance, 121
harmonic function, 147, 148
Harper’s vertex isoperimetric theorem, 129
Harris’ inequality, 213, 266
Harris’ theorem, 221
hitting time, 14, 100, 173
hitting-time theorem, 346
Hoeffding’s inequality, 53, 54, 93, 117
Holley’s lemma, 118
Holley’s theorem, 215, 266
hypothesis class, 84
independent set, 23
indicator trick, 24
infinite trees, 42
inherited property, 314
Ising model, 302
  boundary conditions, 209
  complete graph, 303
  Hamiltonian, 210, 242
  magnetization, 303
  partition function, 210, 242
  spins, 209, 242
isoperimetric inequality, 289
Janson’s inequality, 266
Johnson-Lindenstrauss lemma, 70
Kesten’s theorem, 221
Kirchhoff’s resistance formula, 175
Kolmogorov’s inequality, 110
Kullback-Leibler divergence, 51
Laplacian
  matrix, 162
  operator, 149
large deviations, 52, 98
Lipschitz condition, 127
Lipschitz process, 65
loop erasure, 177
Markov chain
  construction, 6
  cover time, 101
  decomposition theorem, 103
  definitions, 8
  first return, 101
  first visit, 101
  Markov property, 7
  mixing time, 11
  positive recurrence, 103
  recurrence, 103
  reversible, 147
  strong Markov property, 101
Markov chain tree theorem, 185
Markov chains, 187, 229
  asymmetric random walk on \( \mathbb{Z} \), 288
  hitting times, 257
  relaxation time, 276
Markov’s inequality, 18, 19, 24, 47, 86, 117
martingale, 108, 117, 147, 148, 184
Doob martingale, 109, 119, 129
exposure martingale, 129
martingale difference, 118
martingales
  convergence theorem, 109, 310
  Doob’s submartingale inequality, 110, 117
  stopped process, 111
max-flow min-cut theorem, 5
maximum degree, 61
maximum principle, 149, 183
McDiarmid’s inequality, 125, 130
method of bounded differences, 125, 127, 134
method of moments, 97
method of random paths, 168, 184
mixing time, 14, 78, 229
cutoff, 236, 307
diameter bound, 82
lower bound, 81
pre-cutoff, 307
mixing times
lower bounds, 243
upper bounds, 243
moment-generating function, 18, 47, 52
moments, 17
central moments, 17
Nash-Williams inequality, 164, 184
negative association, 176
network, 13
No Free Lunch Theorem, 84
orthogonality of increments, 111, 120
Pólya’s theorem, 163, 168
Pólya’s urn, 109
packing number, 64
pairwise uncorrelated, 95
Paley-Zygmund inequality, 33
parity functions, 280
path coupling, 244
pattern matching, 126
Peierls’ argument, see contour argument
percolation, 28, 42, 188
critical exponents, 326
critical value, 28, 42, 221
dual lattice, 29
Galton-Watson tree, 316
Harris’ theorem, 266
Kesten’s theorem, 266
percolation function, 28, 42, 204, 221
percolation on $\mathbb{L}^2$, 28, 221
percolation on $\mathbb{L}^d$, 204
percolation on trees, 42
phase transition, 327
RSW theorem, 266
percolation function, 42
Poincaré inequality, 124, 274
Poisson approximation, 196
Poisson trials, 52
Poisson variable, 50
poset, 203
positive association, 199, 213
strong, 263
positive correlation, 213
predictable process, 99
preferential attachment graph, 133
probabilistic method, 21, 85, 93
probability generating function, 311
pseudo-regret, 141
random
$b$-ary tree, 236
random $k$-SAT, 25
random permutation
longest increasing subsequence, 26
random projection method, 70
random target lemma, 150
random walk
biased random walk on $\mathbb{Z}$, 189
cycle, 232, 277
hypercube, 232, 280
lazy, 231
simple random walk on $\mathbb{Z}$, 78
random walk on network, 14
Rayleigh quotient, 285
Rayleigh’s principle, 166, 176
recurrence, 14, 147, 158, 166
reflection principle, 102, 345
369
relative entropy, 51
relaxation time
  random walk on cycle, 278
  random walk on hypercube, 280
restricted isometry property, 73, 74
rough embedding, 169, 171
rough equivalence, 170, 184
rough isometry, 184
RSW theorem, 221
Russo’s formula, 221
satifiability threshold, 25
Sauer’s lemma, 89, 90, 93
scale-free trees, 133
second moment method, 17, 32, 39, 42, 339
  weighted second moment method, 45
separation distance, 269
set balancing, 49
shattering, 89
simple random walk on a graph, 13
slicing method, 145
stochastic bandit
  arm, 141
spanning arborescence, 179
sparse signal recovery, 73
sparsity, 73
spectral gap, 276
spectral radius, 286
Spitzer’s combinatorial lemma, 320
stochastic bandit
  Upper Confidence Bound, 141
stochastic domination, 187, 199, 204, 213, 332
  Markov chain, 208
stochastic matrix, 6
stochastic monotonicity, 208
stopping time, 100
Strassen’s theorem, 266
sub-exponential variable, 57
sub-Gaussian increments, 68
sub-Gaussian variable, 53, 118
submodularity, 266
symmetrization, 54, 87, 92
tail probabilities, 18
Thomson’s principle, 161
threshold phenomenon, 25, 28, 35, 42
  threshold function, 35
tilting, 56
total variation distance, 10
transience, see recurrence
trees
  branching ratio, 115
Turán graphs, 24
type, see recurrence
uniform spanning tree
  weighted uniform spanning tree, 177
  Wilson’s method, 175
uniform spanning trees, 175
  cycle popping algorithm, 178
uniform uncertainty principle, 74
union bound, 24, 95
Varopoulos-Carne bound, 78, 98
VC dimension, 89, 92
vertex boundary, 289
Wasserstein distance, 265