2

Moments and Tails

In this chapter, we look at the moments of a random variable. Specifically, we demonstrate that moments capture useful information about the tail of a random variable while often being simpler to compute or at least bound. Several well-known inequalities quantify this intuition. Although they are straightforward to derive, such inequalities are surprisingly powerful. Through a range of applications, we illustrate the utility of controlling the tail of a random variable, typically by allowing one to dismiss certain "bad events" as rare. We begin in Section 2.1 by recalling the classical Markov and Chebyshev inequalities. Then we discuss three of the most fundamental tools in discrete probability and probabilistic combinatorics. In Sections 2.2 and 2.3, we derive the complementary *first* and *second moment methods*, and give several standard applications, especially to threshold phenomena in random graphs and percolation. In Section 2.4, we develop the *Chernoff–Cramér method*, which relies on the moment-generating function and is the building block for a large class of tail bounds. Two key applications in data science are briefly introduced: sparse recovery and empirical risk minimization.

2.1 Background

We start with a few basic definitions and standard inequalities. See Appendix B for a refresher on random variables and their expectation.

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 *kth moment* and *kth central moment* of X. The first moment and the second central moment are known as 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 kth moment can be expressed as

MOMENTS

$$\mathbb{E}[X^k] = \int_0^{+\infty} k\,x^{k-1}\,\mathbb{P}[X > x]\,\mathrm{d}x. \tag{2.1.1}$$

The *moment-generating function* (or *exponential moment*) of X is the function

MOMENT-GENERATING FUNCTION

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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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, for any $k \in \mathbb{Z}$,

$$\frac{\mathrm{d}^k}{\mathrm{d}s} M_X(s) = \mathbb{E}\left[X^k e^{sX}\right], \tag{2.1.2}$$

and the following expansion holds

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

The moment-generating function plays nicely with sums of independent random variables. Specifically, if X_1 and X_2 are independent random variables with M_{X_1} and M_{X_2} defined over a joint interval $(-s_0, s_0)$, then for s in that interval,

$$\begin{aligned} M_{X_1+X_2}(s) &= \mathbb{E}\left[e^{s(X_1+X_2)}\right] \\ &= \mathbb{E}\left[e^{sX_1} e^{sX_2}\right] \\ &= \mathbb{E}\left[e^{sX_1}\right] \mathbb{E}\left[e^{sX_2}\right] \\ &= M_{X_1}(s) M_{X_2}(s), \end{aligned} \tag{2.1.3}$$

where we used independence in the third equality.

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

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

TAIL

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 (much) 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 bounds 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 for two main reasons: (i) moments contain information about the tails of a random variable, as (2.1.1) makes explicit for instance; and (ii) they are typically easier to compute – or, at least, to approximate.

As we will see, tail bounds are also useful to study the maximum of a collection of random variables.

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.

MARKOV'S INEQUALITY

Theorem 2.1.1 (Markov's inequality). *Let X be a non-negative random variable. Then, for all $b > 0$,*

$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E} X}{b}. \tag{2.1.4}$$

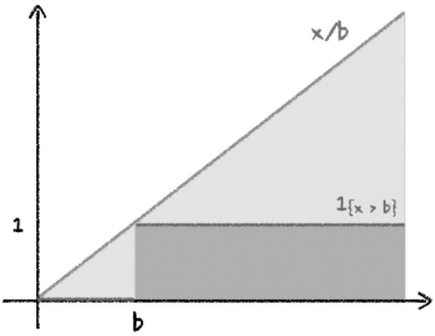

Figure 2.1 Proof of Markov's inequality: taking expectations of the two functions depicted in the figure yields the inequality.

Proof

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

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 (Theorem 2.1.1) to $|X - \mathbb{E}X|^2$ gives a classical tail bound featuring the second moment of a random variable.

Theorem 2.1.2 (Chebyshev's inequality). *Let X be a random variable with $\mathbb{E}X^2 < +\infty$. Then, for all $\beta > 0$,*

CHEBYSHEV'S INEQUALITY

$$\mathbb{P}[|X - \mathbb{E}X| > \beta] \leq \frac{\mathrm{Var}[X]}{\beta^2}. \tag{2.1.5}$$

Proof This follows immediately by applying (2.1.4) to $|X - \mathbb{E}X|^2$ with $b = \beta^2$. ■

Of course, this bound is non-trivial only when β is larger than the standard deviation. Results of this type that quantify the probability of deviating from the mean are referred to as *concentration inequalities*. Chebyshev's inequality is perhaps the simplest instance – we will derive many more. To bound the variance, the following standard formula is sometimes useful:

CONCENTRATION INEQUALITIES

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i<j} \mathrm{Cov}[X_i, X_j], \tag{2.1.6}$$

where we recall that the *covariance* of X_i and X_j is

COVARIANCE

$$\mathrm{Cov}[X_i, X_j] := \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j].$$

When X_i and X_j are independent, then $\mathrm{Cov}[X_i, X_j] = 0$.

Example 2.1.3 Let X be a Gaussian random variable with mean μ and variance σ^2, that is, whose density is

GAUSSIAN

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

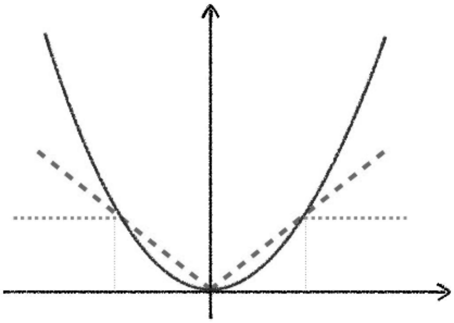

Figure 2.2 Comparison of the Markov and Chebyshev inequalities: the squared deviation from the mean (solid) gives a better approximation of the indicator function (dotted) close to the mean than the absolute deviation (dashed).

We write $X \sim \mathrm{N}(\mu, \sigma^2)$. A direct computation shows that $\mathbb{E}|X-\mu| = \sigma\sqrt{\frac{2}{\pi}}$. Hence Markov's inequality gives

$$\mathbb{P}[|X-\mu| \geq b] \leq \frac{\mathbb{E}|X-\mu|}{b} = \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{b},$$

while Chebyshev's inequality (Theorem 2.1.2) gives

$$\mathbb{P}[|X-\mu| \geq b] \leq \left(\frac{\sigma}{b}\right)^2.$$

Hence, for b large enough, Chebyshev's inequality produces a stronger bound. See Figure 2.2 for some insight. ◀

UNIFORM

Example 2.1.4 (Coupon collector's problem). Let $(X_t)_{t\in\mathbb{N}}$ be i.i.d. uniform random variables over $[n]$, that is, that are equally likely to take any value in $[n]$. Let $T_{n,i}$ be the first time that i elements of $[n]$ have been picked, that is,

$$T_{n,i} = \inf\{t \geq 1 : \ |\{X_1, \dots, X_t\}| = i\},$$

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 $n \to +\infty$:

COUPON COLLECTOR

Claim 2.1.5

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

To prove this claim we note that the time elapsed between $T_{n,i-1}$ and $T_{n,i}$, which we denote by $\tau_{n,i} := T_{n,i} - T_{n,i-1}$, is geometric with success probability $1 - \frac{i-1}{n}$. And all $\tau_{n,i}$s are independent. Recall that a geometric random variable Z with success probability p has probability mass function $\mathbb{P}[Z = z] = (1-p)^{z-1}p$ for $z \in \mathbb{N}$ and has mean $1/p$ and variance $(1-p)/p^2$. So, the expectation and variance of $T_{n,n} = \sum_{i=1}^{n} \tau_{n,i}$ are

GEOMETRIC

$$\mathbb{E}[T_{n,n}] = \sum_{i=1}^{n}\left(1 - \frac{i-1}{n}\right)^{-1} = n\sum_{j=1}^{n} j^{-1} = \Theta(n \log n) \qquad (2.1.7)$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

and

$$\mathrm{Var}[T_{n,n}] \leq \sum_{i=1}^{n} \left(1 - \frac{i-1}{n}\right)^{-2} = n^2 \sum_{j=1}^{n} j^{-2} \leq n^2 \sum_{j=1}^{+\infty} j^{-2} = \Theta(n^2). \tag{2.1.8}$$

So by Chebyshev's inequality

$$\begin{aligned}\mathbb{P}\left[\left|T_{n,n} - n\sum_{j=1}^{n} j^{-1}\right| \geq \varepsilon\, n \log n\right] &\leq \frac{\mathrm{Var}[T_{n,n}]}{(\varepsilon\, n \log n)^2} \\ &\leq \frac{n^2 \sum_{j=1}^{+\infty} j^{-2}}{(\varepsilon\, n \log n)^2} \\ &\to 0,\end{aligned}$$

by (2.1.7) and (2.1.8). ◀

A classical implication of Chebyshev's inequality is (a version of) the law of large numbers. Recall that a sequence of random variables $(X_n)_{n\geq 1}$ converges in probability to a random variable X, denoted by $X_n \to_{\mathrm{p}} X$, if for all $\varepsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}[|X_n - X| \geq \varepsilon] \to 0.$$

Theorem 2.1.6 (L^2 weak law of large numbers). *Let $X_1, X_2, \ldots$ be* uncorrelated *random variables, that is, $\mathbb{E}[X_iX_j] = \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ for $i \neq j$, with $\mathbb{E}[X_i] = \mu < +\infty$ and* $\sup_i \mathrm{Var}[X_i] < +\infty$. *Then* UNCORRELATED

$$\frac{1}{n}\sum_{k\leq n} X_k \to_{\mathrm{p}} \mu.$$

See Exercise 2.5 for a proof. When the X_ks are i.i.d. and integrable (but not necessarily square integrable), convergence is almost sure. That result, the strong law of large numbers, also follows from Chebyshev's inequality (and other ideas), but we will not prove it here.

2.2 First Moment Method

In this section, we develop some techniques based on the first moment. Recall that the expectation of a random variable has an elementary, yet handy, property: linearity. That is, if random variables $X_1, \ldots, X_k$ defined on a joint probability space have finite first moments, then

$$\mathbb{E}[X_1 + \cdots + X_k] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_k] \tag{2.2.1}$$

without any further assumption. In particular, linearity holds whether or not the X_is 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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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.

FIRST MOMENT PRINCIPLE

Theorem 2.2.1 (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 $\mathbb{P}[X \leq \mu] = 0$. We can write $\{X \leq \mu\} = \bigcap_{n \geq 1} \{X < \mu + 1/n\}$. That implies by monotonicity (see Lemma B.2.6) that, for any $\varepsilon \in (0,1)$, it holds that $\mathbb{P}[X < \mu + 1/n] < \varepsilon$ for n large enough. Hence, because we assume that $\mathbb{P}[X \leq \mu] = 0$,

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

a contradiction. ■

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

Example 2.2.2 (Balancing vectors). Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be arbitrary unit vectors in $\mathbb{R}^n$. How small can we make the 2-norm of the linear combination

$$x_1 \boldsymbol{v}_1 + \cdots + x_n \boldsymbol{v}_n$$

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 $\boldsymbol{v}_i$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_is *at random*. Let $X_1, \ldots, X_n$ be independent random variables uniformly distributed in $\{-1, +1\}$. Then, since $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\,\mathbb{E}[X_j] = 0$ for all $i \neq j$ but $\mathbb{E}[X_i^2] = 1$ for all i,

$$\begin{aligned}
\mathbb{E}\|X_1 \boldsymbol{v}_1 + \cdots + X_n \boldsymbol{v}_n\|_2^2 &= \mathbb{E}\left[\sum_{i,j} X_i X_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle\right] \\
&= \sum_{i,j} \mathbb{E}[X_i X_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle] \\
&= \sum_{i,j} \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle\, \mathbb{E}[X_i X_j] \\
&= \sum_i \|\boldsymbol{v}_i\|_2^2 \\
&= n,
\end{aligned}$$

where we used the linearity of expectation on the second line. Hence, random variable $Z = \|X_1 \boldsymbol{v}_1 + \cdots + X_n \boldsymbol{v}_n\|^2$ has expectation $\mathbb{E}Z = n$ and must take a value $\leq n$ with positive

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

probability by the first moment principle (Theorem 2.2.1). In other words, there must be a choice of X_is 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 2.2.3 (Independent sets). For $d \in \mathbb{N}$, let $G = (V, E)$ be a d-regular graph with n vertices. Such a graph necessarily has $m = nd/2$ edges. 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:

Claim 2.2.4

$$\alpha(G) \geq \frac{n}{2d}.$$

Proof 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.

Step 1. 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

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

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

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

where we also used that $\mathbb{E}[\mathbf{1}_{i \in S}\mathbf{1}_{j \in S}] = p^2$ 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 applied to $X - Y$, there must exist a set S of vertices in G such that

$$|S| - |\{\{i, j\} \in E \colon i, j \in S\}| \geq \frac{n}{2d}. \tag{2.2.2}$$

Step 2. For each edge e connecting two vertices in S, remove one of the endvertices of e. By construction, the remaining set of vertices (i) forms an independent set, and (ii) has a size larger than or equal to the left-hand side of (2.2.2). That inequality implies the claim. ■

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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 2.2.5 *The previous result can be strengthened to*

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

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

INDICATOR TRICK

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

2.2.2 Boole's Inequality

One implication of the first moment principle (Theorem 2.2.1) is that if a *non-negative, integer-valued* random variable X has expectation strictly smaller than 1, then its value is 0 with positive probability. The following application of Markov's inequality (Theorem 2.1.1) adds a quantitative twist: if that same X has a "small" expectation, then its value is 0 with "large" probability.

Theorem 2.2.6 (First moment method). *If X is a non-negative, integer-valued random variable, then*

$$\mathbb{P}[X > 0] \leq \mathbb{E}X. \tag{2.2.3}$$

Proof Take $b = 1$ in Markov's inequality. ■

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 2.2.6 reduces simply to the standard *union bound*, also known as *Boole's inequality*. We record one useful version of this setting in the next corollary.

UNION BOUND

Corollary 2.2.7 *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*

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

we have

$$\mathbb{P}[B_n] \leq \mu_n.$$

In particular, if $\mu_n \to 0$ as $n \to +\infty$, then $\mathbb{P}[B_n] \to 0$.

Proof Take $X := X_n = \sum_{i=1}^{m_n} \mathbf{1}_{A_{n,i}}$ in Theorem 2.2.6. ■

A useful generalization of the union bound is given in Exercise 2.2.

FIRST MOMENT METHOD

We will refer to applications of Theorem 2.2.6 as the *first moment method*. We give a few examples.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Example 2.2.8 (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 $\lceil rn \rceil$ clauses. That is, $\Phi_{n,r}$ is an AND of $\lceil rn \rceil$ 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 $\bar{z}_i$. The formula $\Phi_{n,r}$ is said to be satisfiable if there exists an assignment $z = (z_1, \ldots, z_n)$ 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 natural to conjecture that a sharp transition takes place, that is, that there exists an $r_k^* \in \mathbb{R}_+$ (depending on k but not on n) such that

$$\lim_{n\to\infty} \mathbb{P}[\Phi_{n,r} \text{ is satisfiable}] = \begin{cases} 0 \text{ if } r > r_k^*, \\ 1 \text{ if } r < r_k^*. \end{cases} \tag{2.2.4}$$

Studying such *threshold phenomena* is a major theme of modern discrete probability. Using the first moment method (Theorem 2.2.6), we give an upper bound on the threshold. Formally:

THRESHOLD PHENOMENON

Claim 2.2.9

$$r > 2^k \log 2 \implies \limsup_{n\to\infty} \mathbb{P}[\Phi_{n,r} \textit{ is satisfiable}] = 0.$$

Proof 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

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

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

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

There are 2^n possible assignments. Each fixed assignment satisfies the random choice of clauses $\Phi_{n,r}$ with probability $(1 - 2^{-k})^{\lceil rn \rceil}$. Indeed, note that the rn clauses are picked independently and each clause literal picked is satisfied with probability $1/2$. Therefore, by the assumption on r, for $\varepsilon > 0$ small enough and n large enough,

$$\begin{aligned} \mathbb{E} X_n &= 2^n (1 - 2^{-k})^{\lceil rn \rceil} \\ &\leq 2^n (1 - 2^{-k})^{(2^k \log 2)(1+\varepsilon)n} \\ &\leq 2^n e^{-(\log 2)(1+\varepsilon)n} \\ &= 2^{-\varepsilon n} \\ &\to 0, \end{aligned}$$

where we used $(1 - 1/\ell)^\ell \leq e^{-1}$ for all $\ell \in \mathbb{N}$ (see Exercise 1.16). Theorem 2.2.6 implies the claim. ∎

Remark 2.2.10 *Bounds in the other direction are also known. For instance, for $k \geq 3$, it has been shown that if $r < 2^k \log 2 - k$, then*

$$\liminf_{n\to\infty} \mathbb{P}[\Phi_{n,r} \textit{ is satisfiable}] = 1.$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

See [ANP05]. For the $k = 2$ case, it is known that (2.2.4) *in fact holds with $r_2^* = 1$ [CR92]. A breakthrough of [DSS22] also establishes* (2.2.4) *for large k; the threshold r_k^* is characterized as the root of a certain equation coming from statistical physics.* ◀

2.2.3 ▷ *Random Permutations: Longest Increasing Subsequence*

RANDOM PERMUTATION

In this section, we bound the expected length of a longest increasing subsequence in a random permutation. Let $\sigma_n = (\sigma_n(1), \ldots, \sigma_n(n))$ be a uniformly random permutation of $[n] := \{1, \ldots, n\}$ (i.e., a bijection of $[n]$ to itself chosen uniformly at random among all such mappings) and let L_n be the length of a longest increasing subsequence of σ_n (i.e., a sequence of indices $i_1 < \cdots < i_k$ such that $\sigma_n(i_1) < \cdots < \sigma_n(i_k)$).

Claim 2.2.11

$$\mathbb{E}L_n = \Theta(\sqrt{n}).$$

Proof We first prove that

$$\limsup_{n\to\infty} \frac{\mathbb{E}L_n}{\sqrt{n}} \leq e, \tag{2.2.5}$$

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 ℓ for which $\mathbb{P}[L_n \geq \ell]$ is "small." More formally, we bound the expectation as

$$\mathbb{E}L_n \leq \ell\,\mathbb{P}[L_n < \ell] + n\,\mathbb{P}[L_n \geq \ell] \leq \ell + n\,\mathbb{P}[L_n \geq \ell] \tag{2.2.6}$$

for an ℓ chosen below. To bound the probability on the right-hand side, we appeal to the first moment method (Theorem 2.2.6) by letting X_n be the number of increasing subsequences of length ℓ. We also use the indicator trick, that is, we think of X_n as a sum of indicators over subsequences (not necessarily increasing) of length ℓ.

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,

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

where we used a standard bound on factorials recalled in Appendix A. Note that, in order for this bound to go to 0, we need $\ell > e\sqrt{n}$. Then (2.2.5) follows by taking $\ell = (1+\delta)e\sqrt{n}$ in (2.2.6), for an arbitrarily small $\delta > 0$.

For the other half of the claim, we show that

$$\frac{\mathbb{E}L_n}{\sqrt{n}} \geq 1.$$

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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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. 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 $L_n^{(j)}$. If the opposite holds, then we have instead $D_n^{(k)} > D_n^{(j)}$. Either way, $(L_n^{(j)}, D_n^{(j)})$ and $(L_n^{(k)}, D_n^{(k)})$ must be distinct. This clever combinatorial argument is known as the *Erdős–Szekeres Theorem*. That concludes the proof of the second claim. ■

Remark 2.2.12 *It has been shown that in fact*

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

as $n \to +\infty$, where $c = -1.77\ldots$ [BDJ99].

2.2.4 ▹ *Percolation: Existence of a Non-Trivial Threshold on $\mathbb{Z}^2$*

In this section, we use the first moment method (Theorem 2.2.6) to prove the existence of a non-trivial threshold in bond percolation on the two-dimensional lattice. We begin with some background.

Threshold in bond percolation Consider bond percolation (Definition 1.2.1) on the two-dimensional lattice $\mathbb{L}^2$ (see Section 1.1.1) with density p. Let $\mathbb{P}_p$ denote the corresponding measure. Recall that paths are "self-avoiding" by definition (see Section 1.1.1). We say that a path is open if all edges in the induced subgraph are open. 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

OPEN PATH

$$\mathcal{C}_x := \{y \in \mathbb{Z}^2 \,:\, x \Leftrightarrow y\}.$$

The *percolation function* is defined as

PERCOLATION FUNCTION

$$\theta(p) := \mathbb{P}_p[|\mathcal{C}_0| = +\infty],$$

that is, $\theta(p)$ 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_es 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

CRITICAL VALUE

$$p_c(\mathbb{L}^2) := \sup\{p \geq 0 \colon \theta(p) = 0\},$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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 $\mathbb{P}_p[\exists x,\ |\mathcal{C}_x| = +\infty] = 0$. Conversely, because $\{\exists x,\ |\mathcal{C}_x| = +\infty\}$ is a tail event (see Definition B.3.9) for any enumeration of the edges, by Kolmogorov's 0-1 law (Theorem B.3.11) it holds that $\mathbb{P}_p[\exists x,\ |\mathcal{C}_x| = +\infty] = 1$ when $\theta(p) > 0$.

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

Claim 2.2.13

$$p_c(\mathbb{L}^2) \in (0, 1).$$

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 substructure. The key observation is the following:

An infinite $\mathcal{C}_0$ contains an open path starting at 0 of *infinite* length and, as a result, of *all* lengths.

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

$$\mathbb{P}_p[|\mathcal{C}_0| = +\infty] \leq \mathbb{P}_p[\cap_n \{X_n > 0\}] = \lim_n \mathbb{P}_p[X_n > 0] \leq \limsup_n \mathbb{E}_p[X_n], \tag{2.2.7}$$

where the last inequality follows from Theorem 2.2.6. We bound the number of paths by noting that they *cannot backtrack*. That gives four choices at the first step, and at most three choices at each subsequent step. Hence, we get the following bound

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

The right-hand side goes to 0 for all $p < 1/3$. When combined with (2.2.7), that proves half of the claim:

$$p_c(\mathbb{L}^2) > 0.$$

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* $\widetilde{\mathbb{L}}^2$ 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.

DUAL LATTICE

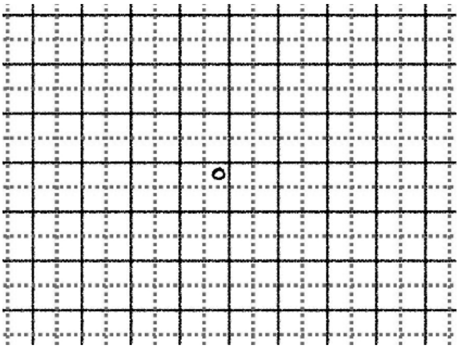

Figure 2.3 Primal (solid) and dual (dotted) lattices.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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. Recall that cycles are "self-avoiding" by definition (see Section 1.1.1). We say that a cycle is closed if all edges in the induced subgraph are closed, that is, are not open.

Lemma 2.2.14 (Contour lemma). *If $|\mathcal{C}_0| < +\infty$, then there is a closed cycle around the origin in the dual lattice $\widetilde{\mathbb{L}}^2$.*

CONTOUR LEMMA

To prove that $\theta(p) > 0$ for p close enough to 1, the idea is to apply the first moment method to Z_n equal to the number of closed 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 y-axis and for each $n-1$ subsequent non-backtracking choices. Namely,

$$\begin{aligned}
\mathbb{P}[|\mathcal{C}_0| < +\infty] &\le \mathbb{P}[\exists n \ge 4, Z_n > 0] \\
&\le \sum_{n\ge4} \mathbb{P}[Z_n > 0] \\
&\le \sum_{n\ge4} \mathbb{E}Z_n \\
&\le \sum_{n\ge4} \frac{n}{2} 3^{n-1}(1-p)^n \\
&= \frac{3^3(1-p)^4}{2} \sum_{m\ge1} (m+3)(3(1-p))^{m-1} \\
&= \frac{3^3(1-p)^4}{2} \left(\frac{1}{(1-3(1-p))^2} + 3\frac{1}{1-3(1-p)} \right)
\end{aligned}$$

when $p > 2/3$, where the first term in parentheses on the last line comes from differentiating with respect to q the geometric series $\sum_{m\ge0} q^m$ and setting $q := 1-p$. This expression can be taken smaller than 1 if we let p approach 1. We have shown that $\theta(p) > 0$ for p close enough to 1, and that concludes the proof. (Exercise 2.3 sketches a proof that $\theta(p) > 0$ for all $p > 2/3$.) ■

It is straightforward to extend the claim to $\mathbb{L}^d$. (Exercise 2.4 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 2.2.14 Assume $|\mathcal{C}_0| < +\infty$. 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 $\mathcal{C}_0$. Paint red the squares of vertices in $\mathcal{C}_0^c$ *which share a side with a green square*. Leave the other squares white. Let u_0 be a highest vertex in $\mathcal{C}_0$ along the y-axis and let v_0 and v_1 be the dual vertices corresponding to the upper left and right corners, respectively, 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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

v_0 to v_1. Notice that this edge satisfies what we call the *red-green property*: as you traverse it from v_0 to v_1, 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. 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 (i.e., it cannot be white). 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 (of length at least 4).
- *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 $\mathcal{C}_0$, as they could not be connected to the origin with open paths.

  So it remains to consider the latter case, where through a similar reasoning the origin must lie 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 v_j. Let v_k be the dual vertex preceding v_j on the second visit. After traversing the edge from v_{j-1} to v_j, v_k cannot be to the left or to the right because in those cases the red-green properties of the two corresponding edges (i.e., $\{v_{j-1}, v_j\}$ and $\{v_k, v_j\}$) are not compatible. So v_k is straight ahead and, by the priority rules, v_{j+1} must be to the left upon entering v_j the first time. But in that case, for the origin to lie outside the cycle as we are assuming and for the cycle to avoid the path $v_0, \ldots, v_{j-1}$, we must traverse the cycle with a negative orientation, that is, 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* $u_0 + (0, 1)$, again a contradiction.

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

That concludes the proof. ∎

Remark 2.2.15 *It turns out that* $p_c(\mathbb{L}^2) = 1/2$. *We will prove* $p_c(\mathbb{L}^2) \geq 1/2$, *known as Harris' Theorem, in Section 4.2.5. The other direction is due to Kesten [Kes80].*

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

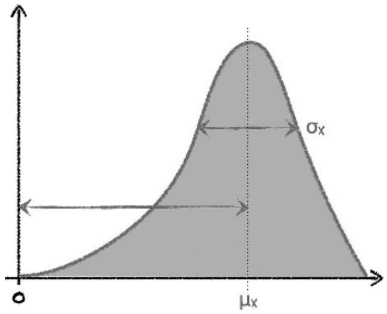

Figure 2.4 Second moment method: if the standard deviation σ_X of X is less than its expectation μ_X, then the probability that X is 0 is bounded away from 1.

2.3 Second Moment Method

The first moment method (Theorem 2.2.6) gives an *upper bound* on the probability that a non-negative, integer-valued random variable is positive – which is non-trivial 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 2.1.2), 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 random variable (not identically zero). Then

$$\mathbb{P}[X > 0] \geq 1 - \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2}. \tag{2.3.1}$$

Indeed, by (2.1.5),

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

The following tail bound, a simple application of Cauchy–Schwarz (Theorem B.4.8), leads to an improved version of this inequality.

Paley–Zygmund inequality

Theorem 2.3.1 (Paley–Zygmund inequality). *Let X be a non-negative random variable. For all $0 < \theta < 1$,*

$$\mathbb{P}[X \geq \theta\,\mathbb{E}X] \geq (1-\theta)^2 \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}. \tag{2.3.2}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Proof We have

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}[X\mathbf{1}_{\{X<\theta\mathbb{E}X\}}] + \mathbb{E}[X\mathbf{1}_{\{X\geq\theta\mathbb{E}X\}}]\\ &\leq \theta\mathbb{E}X + \sqrt{\mathbb{E}[X^2]\mathbb{P}[X \geq \theta\mathbb{E}X]},\end{aligned}$$

where we used Cauchy–Schwarz. Rearranging gives the result. ■

As an immediate application:

SECOND MOMENT METHOD

Theorem 2.3.2 (Second moment method). *Let X be a non-negative random variable (not identically zero). Then*

$$\mathbb{P}[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}. \tag{2.3.3}$$

Proof Take $\theta \downarrow 0$ in (2.3.2). ■

Since

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

we see that (2.3.3) is stronger than (2.3.1).

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). We record in the next corollary a convenient version of the method.

Corollary 2.3.3 *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 $i \overset{n}{\sim} j$ if $i \neq j$ and $A_{n,i}$ and $A_{n,j}$ are* not *independent. Then, letting*

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

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 We apply the second moment method to $X_n := \sum_{i=1}^{m_n} \mathbf{1}_{A_{n,i}}$ so that $B_n = \{X_n > 0\}$. Note that

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

where

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

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

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

whereas, if $i \overset{n}{\sim} j$,

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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Hence,

$$\frac{\mathrm{Var}[X_n]}{(\mathbb{E}\,X_n)^2} \le \frac{\mu_n + \gamma_n}{\mu_n^2} = \frac{1}{\mu_n} + \frac{\gamma_n}{\mu_n^2}.$$

Noting

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

and applying Theorem 2.3.2 gives the result. ■

2.3.2 ▹ *Random Graphs: Subgraph Containment and Connectivity in the Erdős–Rényi Model*

Threshold phenomena are also common in random graphs. We consider here the Erdős–Rényi random graph model (Definition 1.2.2). In this context, a *threshold function for a graph property P* is a function $r(n)$ such that

THRESHOLD FUNCTION

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

where $G_n \sim \mathbb{G}_{n,p_n}$ is a random 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.

Subgraph containment

We first consider the clique number, then we turn to more general subgraphs.

Cliques Let $\omega(G)$ be the *clique number* of a graph G, that is, the size of its largest clique.

CLIQUE NUMBER

Claim 2.3.4 *The property $\omega(G_n) \ge 4$ has threshold function $n^{-2/3}$.*

Proof Let X_n be the number of 4-cliques in the random graph $G_n \sim \mathbb{G}_{n,p_n}$. Then, noting that there are $\binom{4}{2} = 6$ edges in a 4-clique,

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

which goes to 0 when $p_n \ll n^{-2/3}$. Hence, the first moment method (Theorem 2.2.6) gives one direction: $\mathbb{P}_{n,p_n}[\omega(G_n) \ge 4] \to 0$ in that case.

For the other direction, we apply the second moment method for sums of indicators, that is, Corollary 2.3.3. We use the notation from that corollary. For an enumeration $S_1, \ldots, S_{m_n}$ of the 4-tuples of vertices in G_n, let $A_{n,1}, \ldots, A_{n,m_n}$ be the events that the corresponding 4-clique is present. By the calculation above we have $\mu_n = \Theta(n^4 p_n^6)$, which goes to $+\infty$ when $p_n \gg n^{-2/3}$. Also $\mu_n^2 = \Theta(n^8 p_n^{12})$, so it suffices to show that $\gamma_n = o(n^8 p_n^{12})$. Note that two 4-cliques with disjoint edge sets (but possibly sharing one vertex) are independent (i.e., their presence or absence is independent). Suppose S_i and S_j share three vertices. Then, $i \overset{n}{\sim} j$ and

$$\mathbb{P}_{n,p_n}[A_{n,i} \,|\, A_{n,j}] = p_n^3,$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

as the event $A_{n,j}$ implies that all edges between three of the vertices in S_i are already present, and there are three edges between the remaining vertex and the rest of S_i. Similarly, if $|S_i \cap S_j| = 2$, we have again $i \overset{n}{\sim} j$ and this time $\mathbb{P}_{n,p_n}[A_{n,i} \,|\, A_{n,j}] = p_n^5$. Putting these together, we get by the definition of the conditional probability (see Appendix B) and the fact that $\mathbb{P}_{n,p_n}[A_{n,j}] = p_n^6$

$$\begin{aligned}
\gamma_n &= \sum_{i \overset{n}{\sim} j} \mathbb{P}[A_{n,i} \cap A_{n,j}] \\
&= \sum_{i \overset{n}{\sim} j} \mathbb{P}_{n,p_n}[A_{n,j}]\, \mathbb{P}_{n,p_n}[A_{n,i} \,|\, A_{n,j}] \\
&= \sum_{j} \mathbb{P}_{n,p_n}[A_{n,j}] \sum_{i : i \overset{n}{\sim} j} \mathbb{P}_{n,p_n}[A_{n,i} \,|\, A_{n,j}] \\
&= \binom{n}{4} p_n^6 \left[\binom{4}{3} (n-4) p_n^3 + \binom{4}{2} \binom{n-4}{2} p_n^5 \right] \\
&= O(n^5 p_n^9) + O(n^6 p_n^{11}) \\
&= O\left(\frac{n^8 p_n^{12}}{n^3 p_n^3} \right) + O\left(\frac{n^8 p_n^{12}}{n^2 p_n} \right) \\
&= o(n^8 p_n^{12}) \\
&= o(\mu_n^2),
\end{aligned}$$

where we used that $p_n \gg n^{-2/3}$ (so that for example $n^3 p_n^3 \gg 1$). Corollary 2.3.3 gives the result: $\mathbb{P}_{n,p_n}[\cup_i A_{n,i}] \to 1$ when $p_n \gg n^{-2/3}$. ■

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 2.3.4 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 (not necessarily induced) copies of H_0 in $G_n \sim \mathbb{G}_{n,p_n}$. By the first moment method,

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

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. Indeed, note that a fixed set of v_{H_0} vertices can contain several distinct copies of H_0, depending on its structure.

From the proof of Claim 2.3.4, 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 H_0 in Figure 2.5 whose *edge density* is $\frac{e_{H_0}}{v_{H_0}} = \frac{6}{5}$. When $p_n \gg n^{-5/6}$, the expected number of copies of H_0 in G_n 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

EDGE DENSITY

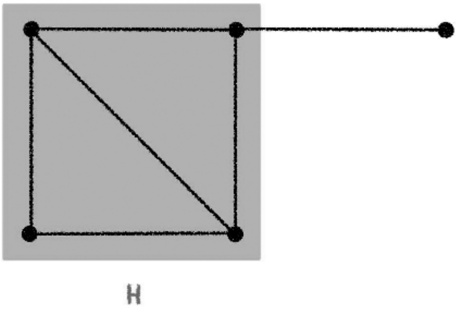

Figure 2.5 Graph H_0 and subgraph H.

that regime is asymptotically negligible despite its diverging expectation. This leads to the following definition

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

Assume H_0 has at least one edge.

Claim 2.3.5 *"Having a copy of H_0" has threshold $n^{-1/r_{H_0}}$.*

Proof We proceed as in Claim 2.3.4. 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 $S_1, \ldots, S_{m_n}$ be an enumeration of the *copies (as subgraphs) of H_0 in a complete graph* on the vertices of G_n. Let $A_{n,i}$ be the event that $S_i \subseteq G_n$. Using again the notation of Corollary 2.3.3,

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

where

$$\Phi_{H_0}(n) := \min\left\{n^{v_H} p_n^{e_H} : \text{subgraphs } H \subseteq H_0, e_H > 0\right\}.$$

Note that $\Phi_{H_0}(n) \to +\infty$ when $p_n \gg n^{-1/r_{H_0}}$ by definition of r_{H_0}. The events $A_{n,i}$ and $A_{n,j}$ are independent if S_i and S_j share no edge. Otherwise we write $i \overset{n}{\sim} j$. Note that there are $\Theta(n^{v_H} n^{2(v_{H_0} - v_H)})$ pairs S_i, S_j whose intersection is isomorphic to H. The probability that both S_i and S_j of such a pair are present in G_n is $\Theta(p_n^{e_H} p_n^{2(e_{H_0} - e_H)})$. Hence,

$$\begin{aligned}
\gamma_n &= \sum_{i \overset{n}{\sim} j} \mathbb{P}[A_{n,i} \cap A_{n,j}] \\
&= \sum_{H \subseteq H_0, e_H > 0} \Theta\left(n^{2v_{H_0} - v_H} p_n^{2e_{H_0} - e_H}\right) \\
&\le \frac{\Theta(\mu_n^2)}{\Theta(\Phi_{H_0}(n))} \\
&= o(\mu_n^2),
\end{aligned}$$

where we used that $\Phi_{H_0}(n) \to +\infty$. The result follows from Corollary 2.3.3. ■

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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 $\Phi_{H_0}(n) \to 0$. 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 small (vanishing in expectation) number of copies of H – producing a large variance that helps to explain the failure of the second moment method.

Connectivity threshold

Next we use the second moment method to show that the threshold function for connectivity in the Erdős–Rényi random graph model 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 2.3.6 *"Not having an isolated vertex" has threshold function* $\frac{\log n}{n}$.

Proof Let X_n be the number of isolated vertices in the random graph $G_n \sim \mathbb{G}_{n,p_n}$. Using $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$ (see Exercise 1.16),

$$\mathbb{E}_{n,p_n}[X_n] = n(1-p_n)^{n-1} \le e^{\log n - (n-1)p_n} \to 0, \tag{2.3.4}$$

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 $A_{n,j}$ be the event that vertex j is isolated. By the computation above, using $1 - x \ge e^{-x-x^2}$ for $x \in [0, 1/2]$ (see Exercise 1.16),

$$\mu_n = \sum_i \mathbb{P}_{n,p_n}[A_{n,i}] = n(1-p_n)^{n-1} \ge e^{\log n - np_n - np_n^2}, \tag{2.3.5}$$

which goes to $+\infty$ when $p_n \ll \frac{\log n}{n}$. Note that $A_{n,i}$ and $A_{n,j}$ are not independent for all $i \neq j$ (because the absence of an edge between i and j is part of both events) and

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

so that

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

Because γ_n is *not* $o(\mu_n^2)$, we cannot apply Corollary 2.3.3. Instead we use Theorem 2.3.2 directly. We have

$$\begin{aligned} \frac{\mathbb{E}_{n,p_n}[X_n^2]}{\mathbb{E}_{n,p_n}[X_n]^2} &= \frac{\mu_n + \gamma_n}{\mu_n^2} \\ &\le \frac{n(1-p_n)^{n-1} + n^2(1-p_n)^{2n-3}}{n^2(1-p_n)^{2n-2}} \\ &\le \frac{1}{n(1-p_n)^{n-1}} + \frac{1}{1-p_n}, \end{aligned} \tag{2.3.6}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

which is $1 + o(1)$ when $p_n \ll \frac{\log n}{n}$ by (2.3.5). The second moment method implies that $\mathbb{P}_{n,p_n}[X_n > 0] \to 1$ in that case. ■

Connectivity We use Claim 2.3.6 to study the threshold for connectivity.

Claim 2.3.7 *Connectivity has threshold function* $\frac{\log n}{n}$.

Proof We start with the easy direction. If $p_n \ll \frac{\log n}{n}$, Claim 2.3.6 implies that the graph has at least one isolated vertex – and therefore is necessarily disconnected – with probability going to 1 as $n \to +\infty$.

Assume now that $p_n \gg \frac{\log n}{n}$. Let $\mathcal{D}_n$ be the event that G_n is disconnected. To bound $\mathbb{P}_{n,p_n}[\mathcal{D}_n]$, we let Y_k be the number of subsets of k vertices that are disconnected from all other vertices in the graph for $k \in \{1, \dots, n/2\}$. Then, by the first moment method,

$$\mathbb{P}_{n,p_n}[\mathcal{D}_n] \leq \mathbb{P}_{n,p_n}\left[\sum_{k=1}^{n/2} Y_k > 0\right] \leq \sum_{k=1}^{n/2} \mathbb{E}_{n,p_n}[Y_k].$$

The expectation of Y_k is straightforward to bound. Using $k \leq n/2$ and $\binom{n}{k} \leq n^k$,

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

The expression in parentheses is $o(1)$ when $p_n \gg \frac{\log n}{n}$ by a calculation similar to (2.3.4). Summing over k,

$$\mathbb{P}_{n,p_n}[\mathcal{D}_n] \leq \sum_{k=1}^{+\infty} \left(n(1-p_n)^{n/2}\right)^k = O(n(1-p_n)^{n/2}) = o(1),$$

where we used that the geometric series (started at $k = 1$) is dominated asymptotically by its first term. So the probability of being disconnected goes to 0 when $p_n \gg \frac{\log n}{n}$ and we have proved the claim. ■

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 $k \in \{1, \dots, n/2\}$, 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, asymptotically the graph is typically composed of a large connected component together with some isolated vertices. Formally, we prove Claim 2.3.8, which says that with probability close to 1, either the graph is connected or there are some isolated vertices together with a (necessarily unique) connected component of size greater than $n/2$.

Claim 2.3.8

$$\mathbb{P}_{n,p_n}[Z_1 > 0] \geq \frac{1}{1+e^s} + o(1) \quad \textit{and} \quad \mathbb{P}_{n,p_n}\left[\sum_{k=2}^{n/2} Z_k > 0\right] = o(1).$$

The limit of $\mathbb{P}_{n,p_n}[Z_1 > 0]$ can be computed explicitly using the method of moments. See Exercise 2.19.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Proof of Claim 2.3.8 We first consider isolated vertices. From (2.3.5), (2.3.6), and the second moment method,

$$\mathbb{P}_{n,p_n}[Z_1 > 0] \geq \left(e^{-\log n + np_n + np_n^2} + \frac{1}{1-p_n}\right)^{-1} = \frac{1}{1+e^s} + o(1),$$

as $n \to +\infty$ by our choice of p_n.

To bound the number of components of size $k > 1$, we note first that the random variable Y_k used in the previous claim (which imposes no condition on the edges *between* the vertices in the subsets of size k) is too loose to provide a suitable bound. Instead, to bound the probability that a subset 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 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

$$\mathbb{P}_{n,p_n}\left[\sum_{k=2}^{n/2} Z_k > 0\right] \leq \mathbb{P}_{n,p_n}\left[\sum_{k=2}^{n/2} Z_k' > 0\right] \leq \sum_{k=2}^{n/2} \mathbb{E}_{n,p_n}[Z_k']. \tag{2.3.7}$$

To bound the expectation of Z_k', we use Cayley's formula, which states that there are k^{k-2} trees on a set of k labeled vertices. Recall further that a tree on k vertices has $k-1$ edges (see Exercise 1.7). Hence,

$$\mathbb{E}_{n,p_n}[Z_k'] = \underbrace{\binom{n}{k} k^{k-2}}_{(a)} \underbrace{p_n^{k-1}}_{(b)} \underbrace{(1-p_n)^{k(n-k)}}_{(c)},$$

where (a) is the number of trees of size k (as subgraphs) in a complete graph of size n, (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 Appendix A) and $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$ (see Exercise 1.16),

$$\begin{aligned}\mathbb{E}_{n,p_n}[Z_k'] &\leq \frac{n^k}{k!} k^{k-2} p_n^{k-1} (1-p_n)^{k(n-k)} \\ &\leq \frac{n^k e^k}{k^k} k^k n p_n^k e^{-p_n k(n-k)} \\ &\leq n\left(ec_n e^{-\left(1-\frac{k}{n}\right)c_n}\right)^k \\ &= n\left(e(\log n + s)e^{-\left(1-\frac{k}{n}\right)(\log n + s)}\right)^k.\end{aligned}$$

For $k \leq n/2$, the expression in parentheses is $o(1)$. In fact, for $2 \leq k \leq n/2$, $\mathbb{E}_{n,p_n}[Z_k'] = o(1)$. Furthermore, summing over $k > 2$,

$$\sum_{k=3}^{n/2} \mathbb{E}_{n,p_n}[Z_k'] \leq \sum_{k=3}^{+\infty} n\left(e(\log n + s)e^{-\frac{1}{2}(\log n + s)}\right)^k = O(n^{-1/2}\log^3 n) = o(1).$$

Plugging this back into (2.3.7) gives the second claim in the statement. ■

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

2.3.3 ▷ Percolation: Critical Value on Trees and Branching Number

Consider bond percolation (see Definition 1.2.1) on the infinite d-regular tree $\mathbb{T}_d$. Root the tree arbitrarily at a vertex 0 and let $\mathcal{C}_0$ be the open cluster of the root. In this section, we illustrate the use of the first and second moment methods on the identification of the critical value

$$p_c(\mathbb{T}_d) = \sup\{p \in [0,1] \colon \theta(p) = 0\},$$

where recall that the percolation function is $\theta(p) = \mathbb{P}_p[|\mathcal{C}_0| = +\infty]$. 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 $\mathbb{T}_d$ is the following.

Claim 2.3.9

$$p_c(\mathbb{T}_d) = \frac{1}{d-1}.$$

Proof Let ∂_n be the nth level of $\mathbb{T}_d$, that is, the set of vertices at graph distance n from 0. Let X_n be the number of vertices in $\partial_n \cap \mathcal{C}_0$. In order for the open cluster of the root to be infinite, there must be at least one vertex on the nth level connected to the root by an open path. By the first moment method (Theorem 2.2.6),

$$\theta(p) = \mathbb{P}_p[|\mathcal{C}_0| = +\infty] \le \mathbb{P}_p[X_n > 0] \le \mathbb{E}_p X_n = d(d-1)^{n-1}p^n \to 0, \tag{2.3.8}$$

as $n \to +\infty$, for any $p < \frac{1}{d-1}$. Here we used that there is a unique path between 0 and any vertex in the tree to deduce that $\mathbb{P}_p[x \in \mathcal{C}_0] = p^n$ for $x \in \partial_n$. Equation (2.3.8) implies half of the claim: $p_c(\mathbb{T}_d) \ge \frac{1}{d-1}$.

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*, that is, subtrees of $\mathbb{T}_d$ rooted at z made of all descendants of z. Let $x \wedge y$ be the *most recent common ancestor* of x and y, that is, 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

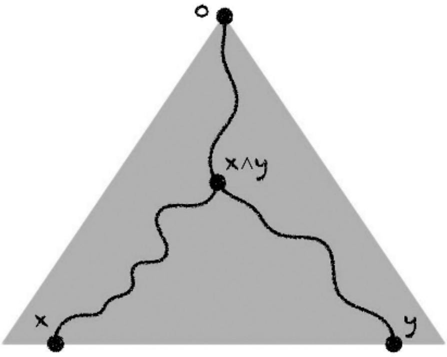

Figure 2.6 Most recent common ancestor of x and y.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

$$\mu_n := \mathbb{E}_p[X_n] = \mathbb{E}_p\left[\sum_{x\in\partial_n} \mathbf{1}_{\{x\in\mathcal{C}_0\}}\right] = (b+1)b^{n-1}p^n,$$

we have

$$\begin{aligned}
\mathbb{E}_p[X_n^2] &= \mathbb{E}_p\left[\left(\sum_{x\in\partial_n} \mathbf{1}_{\{x\in\mathcal{C}_0\}}\right)^2\right] \\
&= \sum_{x,y\in\partial_n} \mathbb{P}_p[x,y\in\mathcal{C}_0] \\
&= \sum_{x\in\partial_n} \mathbb{P}_p[x\in\mathcal{C}_0] + \sum_{m=0}^{n-1}\sum_{x,y\in\partial_n} \mathbf{1}_{\{x\wedge y\in\partial_m\}} p^m p^{2(n-m)} \\
&= \mu_n + (b+1)b^{n-1}\sum_{m=0}^{n-1}(b-1)b^{(n-m)-1}p^{2n-m} \\
&\leq \mu_n + (b+1)(b-1)b^{2n-2}p^{2n}\sum_{m=0}^{+\infty}(bp)^{-m} \\
&= \mu_n + \mu_n^2\cdot\frac{b-1}{b+1}\cdot\frac{1}{1-(bp)^{-1}},
\end{aligned}$$

where, on the fourth line, we used that all vertices on the nth level are equivalent and that, for a fixed x, the set $\{y\colon x\wedge y\in\partial_m\}$ is composed of those vertices in ∂_n that are descendants of $x\wedge y$ but not in the descendant subtree of $x\wedge y$ containing x. When $p > \frac{1}{d-1} = \frac{1}{b}$, dividing by $(\mathbb{E}_pX_n)^2 = \mu_n^2 \to +\infty$, we get

$$\begin{aligned}
\frac{\mathbb{E}_p[X_n^2]}{(\mathbb{E}_pX_n)^2} &\leq \frac{1}{\mu_n} + \frac{b-1}{b+1}\cdot\frac{1}{1-(bp)^{-1}} \\
&\leq 1 + \frac{b-1}{b+1}\cdot\frac{1}{1-(bp)^{-1}} \\
&=: C_{b,p}.
\end{aligned} \tag{2.3.9}$$

By the second moment method (Theorem 2.3.2) and monotonicity,

$$\theta(p) = \mathbb{P}_p[|\mathcal{C}_0| = +\infty] = \mathbb{P}_p[\forall n,\ X_n > 0] = \lim_n \mathbb{P}_p[X_n > 0] \geq C_{b,p}^{-1} > 0,$$

which concludes the proof. (Note that the version of the second moment method in (2.3.1) does not work here. Subtract 1 in (2.3.9) and take p close to $1/b$.) ■

The argument in the proof of Claim 2.3.9 relies crucially on the fact that, in a tree, any two vertices are connected by a unique path. For instance, approximating $\mathbb{P}_p[x \in \mathcal{C}_0]$ 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 $\mathbb{T}_d$ is a "branching process" (defined formally and studied at length in Chapter 6), 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 children bp exceeds 1, the process grows

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

exponentially on average and "explodes" with positive probability (see Theorem 6.1.6). We will come back to this point of view in Section 6.2.4 where branching processes are used to give a more refined analysis of bond percolation on $\mathbb{T}_d$.

General trees Let $\mathscr{T}$ 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. Generalizing a previous definition from Section 1.1.1 to infinite, locally finite graphs, a cutset separating 0 and $+\infty$ is a finite set of edges Π such that all infinite paths (which, recall, are self-avoiding by definition) starting at 0 go through Π. (For our purposes, it will suffice to assume that cutsets are finite by default.) For a cutset Π, we let $\Pi_{\mathrm{v}} := \{v_e \colon e \in \Pi\}$. Repeating the argument in (2.3.8), for any cutset Π, by the first moment method (i.e., union bound),

$$\begin{aligned}\theta(p) &= \mathbb{P}_p[|\mathcal{C}_0| = +\infty]\\ &\leq \mathbb{P}_p[\mathcal{C}_0 \cap \Pi_{\mathrm{v}} \neq \emptyset]\\ &\leq \sum_{u \in \Pi_{\mathrm{v}}} \mathbb{P}_p[u \in \mathcal{C}_0]\\ &= \sum_{e \in \Pi} p^{|e|}. \end{aligned} \tag{2.3.10}$$

This bound naturally leads to the next definition.

BRANCHING NUMBER

Definition 2.3.10 (Branching number). *The* branching number *of* $\mathscr{T}$ *is given by*

$$\mathrm{br}(\mathscr{T}) = \sup\left\{\lambda \geq 1 \colon \inf_{\mathit{cutset}\ \Pi} \sum_{e \in \Pi} \lambda^{-|e|} > 0\right\}. \tag{2.3.11}$$

Using the max-flow min-cut theorem (Theorem 1.1.15), the branching number can also be characterized in terms of a "flow to $+\infty$." We will not do this here. (But see Theorem 3.3.30.)

Equation (2.3.10) implies that $p_c(\mathscr{T}) \geq \frac{1}{\mathrm{br}(\mathscr{T})}$. Remarkably, this bound is tight. The proof is based on a "weighted" second moment method.

Claim 2.3.11 *For any rooted, locally finite tree* $\mathscr{T}$*,*

$$p_c(\mathscr{T}) = \frac{1}{\mathrm{br}(\mathscr{T})}.$$

Proof Suppose $p < \frac{1}{\mathrm{br}(\mathscr{T})}$. Then, $p^{-1} > \mathrm{br}(\mathscr{T})$ and the sum in (2.3.10) can be made arbitrarily small by definition of the branching number, that is, $\theta(p) = 0$. Hence we have shown that $p_c(\mathscr{T}) \geq \frac{1}{\mathrm{br}(\mathscr{T})}$.

To argue in the other direction, let $p > \frac{1}{\mathrm{br}(\mathscr{T})}$, $p^{-1} < \lambda < \mathrm{br}(\mathscr{T})$, and $\varepsilon > 0$ such that

$$\sum_{e \in \Pi} \lambda^{-|e|} \geq \varepsilon \tag{2.3.12}$$

for all cutsets Π. The existence of such an ε is guaranteed by the definition of the branching number. As in the proof of Claim 2.3.9, we use that $\theta(p)$ is the limit as $n \to +\infty$ of the probability that $\mathcal{C}_0$ reaches the nth level (i.e., the vertices at graph distance n from the root 0, which is necessarily a finite set in a locally finite tree). However, this time, we use a *weighted*

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

count on the nth level. Let $\mathscr{T}_n$ be the first n levels of $\mathscr{T}$ and, as before, let ∂_n be the vertices on the nth level. For a probability measure ν_n on ∂_n, we define the weighted count

$$X_n = \sum_{z\in\partial_n} \mathbf{1}_{\{z\in\mathcal{C}_0\}} \frac{\nu_n(z)}{\mathbb{P}_p[z\in\mathcal{C}_0]}.$$

The purpose of the denominator is normalization, that is,

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

Observe that, while $\nu_n(z)$ may be 0 for some zs (but not all), we still have that $X_n > 0, \forall n$ implies $\{|\mathcal{C}_0| = +\infty\}$, which is what we need to apply the second moment method.

Because of (2.3.12), a natural choice of ν_n follows from the max-flow min-cut theorem (Theorem 1.1.15) applied to $\mathscr{T}_n$ with source 0, sink ∂_n, and capacity constraint $|\phi(x,y)| \leq \kappa(e) := \varepsilon^{-1}\lambda^{-|e|}$ for all edges $e = \{x,y\}$. Indeed, for all cutsets Π in $\mathscr{T}_n$ separating 0 and ∂_n, we have $\sum_{e\in\Pi}\kappa(e) = \sum_{e\in\Pi}\varepsilon^{-1}\lambda^{-|e|} \geq 1$ by (2.3.12). That then guarantees by Theorem 1.1.15 the existence of a unit flow ϕ from 0 to ∂_n satisfying the capacity constraints. Define $\nu_n(z)$ to be the flow entering $z \in \partial_n$ under ϕ. In particular, because ϕ is a unit flow, ν_n defines a probability measure. It remains to bound the second moment of X_n under this choice. We have

$$\begin{aligned}
\mathbb{E}_p X_n^2 &= \mathbb{E}_p\left[\left(\sum_{z\in\partial_n} \mathbf{1}_{\{z\in\mathcal{C}_0\}} \frac{\nu_n(z)}{\mathbb{P}_p[z\in\mathcal{C}_0]}\right)^2\right] \\
&= \sum_{x,y\in\partial_n} \nu_n(x)\nu_n(y) \frac{\mathbb{P}_p[x,y\in\mathcal{C}_0]}{\mathbb{P}_p[x\in\mathcal{C}_0]\mathbb{P}_p[y\in\mathcal{C}_0]} \\
&= \sum_{m=0}^{n}\sum_{x,y\in\partial_n} \mathbf{1}_{\{x\wedge y\in\partial_m\}} \nu_n(x)\nu_n(y) \frac{p^m p^{2(n-m)}}{p^{2n}} \\
&= \sum_{m=0}^{n} p^{-m} \sum_{z\in\partial_m}\left(\sum_{x,y\in\partial_n} \mathbf{1}_{\{x\wedge y = z\}}\nu_n(x)\nu_n(y)\right).
\end{aligned}$$

In the expression in parentheses, for each x descendant of z, the sum over y is at most $\nu_n(x)\nu_n(z)$ by the definition of a flow; then the sum over those xs gives at most $\nu_n(z)^2$. So

$$\begin{aligned}
\mathbb{E}_p X_n^2 &\leq \sum_{m=0}^{n} p^{-m} \sum_{z\in\partial_m} \nu_n(z)^2 \\
&\leq \sum_{m=0}^{n} p^{-m} \sum_{z\in\partial_m} (\varepsilon^{-1}\lambda^{-m})\nu_n(z) \\
&\leq \varepsilon^{-1}\sum_{m=0}^{+\infty} (p\lambda)^{-m} \\
&= \frac{\varepsilon^{-1}}{1-(p\lambda)^{-1}} =: C_{\varepsilon,\lambda,p} < +\infty,
\end{aligned}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

where the second line follows from the capacity constraint, and we used $p\lambda > 1$ on the last line. From the second moment method (recalling that $\mathbb{E}_p X_n = 1$),

$$\theta(p) = \mathbb{P}_p[|\mathcal{C}_0| = +\infty] \geq \mathbb{P}_p[\forall n,\ X_n > 0] = \lim_n \mathbb{P}_p[X_n > 0] \geq C_{\varepsilon,\lambda,p}^{-1} > 0.$$

It follows that

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

and $p_c(\mathscr{T}) \leq \frac{1}{\mathrm{br}(\mathscr{T})}$. That concludes the proof. ∎

Note that Claims 2.3.9 and 2.3.11 imply that $\mathrm{br}(\mathbb{T}_d) = d - 1$. The next example is more striking and insightful.

Example 2.3.12 (The 3–1 tree). The 3–1 tree $\widehat{\mathscr{T}}_{3-1}$ is an infinite rooted tree. We give a planar description. The root ρ (level 0) is at the top. It has two children below it (level 1). Then on level n, for $n \geq 1$, the first 2^{n-1} vertices starting from the left have exactly 1 child and the next 2^{n-1} vertices have exactly 3 children. In particular, level n has 2^n vertices, which we denote by $u_{n,1}, \ldots, u_{n,2^n}$. For vertex $u_{n,j}$ we refer to $j/2^n$ as its *relative position* (on level n). So vertices have 1 or 3 children according to whether their relative position is $\leq 1/2$ or $> 1/2$.

RELATIVE POSITION

Because the level size is growing at rate 2, it is tempting to conjecture that the branching number is 2 – but that turns out to be way off.

Claim 2.3.13 $\mathrm{br}(\widehat{\mathscr{T}}_{3-1}) = 1$.

What makes this tree entirely different from the infinite 2-ary tree, despite having the same level growth, is that each infinite path from the root in $\widehat{\mathscr{T}}_{3-1}$ eventually "stops branching," with the sole exception of the rightmost path which we refer to as the *main path*. Indeed, let $\Gamma = v_0 \sim v_1 \sim v_2 \sim \cdots$ with $v_0 = \rho$ be an infinite path distinct from the main path. Let x_i be the relative position of v_i, $i \geq 1$. Let v_k be the first vertex of Γ *not* on the main path. It lies on the kth level.

MAIN PATH

Lemma 2.3.14 *Let v be a vertex that is not on the main path with relative position x and assume that $0 \leq x \leq \alpha < 1$. Let w be a child of v and denote by y its relative position. Then,*

$$y \leq \begin{cases} \frac{1}{2}x & \text{if } x \leq 1/2, \\ x - \frac{1}{2}(1-\alpha) & \text{otherwise.} \end{cases}$$

Proof Assume without loss of generality that $v = u_{n,j}$ for some n and $j < 2^n$. If $j \leq 2^{n-1}$, then by construction v has exactly one child with relative position

$$y = \frac{j}{2^{n+1}} = \frac{1}{2}x.$$

That proves the first claim.

If $j > 2^{n-1}$, then all vertices to the right of v have 3 children, all of whom are to the right of the children of v. Hence, the children of v have relative position at most

$$y \leq \frac{2^{n+1} - 3(2^n - j)}{2^{n+1}} = \frac{3j - 2^n}{2^{n+1}} = \frac{3}{2}x - \frac{1}{2}.$$

Subtracting x and using $x \leq \alpha$ gives the second claim. ∎

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

We now apply Lemma 2.3.14 to v_k as defined above and its descendants on Γ with $\alpha = 1 - 1/2^k$. We get that the relative position decreases from v_k by $1/2^{k+1}$ on each level until it falls below $1/2$ at which point it gets cut in half at each level. Once this last regime is reached, each vertex on Γ from then on has exactly one child – that is, there is no more branching.

We are now ready to prove the claim.

Proof of Claim 2.3.13 Take any $\lambda > 1$. From the definition of the branching number (Definition 2.3.10), it suffices to find a sequence of cutsets $(\Pi_n)_n$ such that

$$\sum_{e \in \Pi_n} \lambda^{-|e|} \to 0,$$

as $n \to +\infty$. What does *not* work is to choose $\Pi_n := \Lambda_n$ to be the edges between level $n-1$ and level n, since we then have

$$\sum_{e \in \Lambda_n} \lambda^{-|e|} = 2^n \lambda^{-n},$$

which diverges whenever $\lambda < 2$. Instead, we construct a new cutset Φ_n based on Λ_n as follows. We divide up Λ_n into the disjoint union $\Lambda_n^- \cup \Lambda_n^+$, where Λ_n^- are the edges whose endvertex on level n has relative position $\leq 1/2$ and Λ_n^+ are the rest of the edges. Start with $\Phi_n := \emptyset$.

Step 1. For each edge e in Λ_n^-, letting v be the endvertex of e on level n, add to Φ_n the edge $\{v', v''\}$ where v' and v'' are the unique descendants of v on level $m_n - 1$ and m_n, respectively. The value of $m_n \geq n$ is chosen so that

$$2^n \lambda^{-m_n} \leq \frac{1}{2n}. \tag{2.3.13}$$

Any infinite path from the root going through one of the edges in Λ_n^- has to go through the edge that replaced it in Φ_n since there is no branching below that point by Lemma 2.3.14.

Step 2. We also add to Φ_n the edge $\{w', w''\}$ on the main path where $w' = u_{\ell_n - 1, 2^{\ell_n - 1}}$ is on level $\ell_n - 1$ and $w'' = u_{\ell_n, 2^{\ell_n}}$ is on level ℓ_n. We mean for the value of ℓ_n to be such that any infinite path going through an edge in Λ_n^+ has to go through $\{w', w''\}$ first. That is, we need all vertices of level n with relative position $> 1/2$ to be a descendant of w''. The number of descendants of w'' on level $J > \ell_n$ is $3^{J-\ell_n}$ until the last J such that it is $\leq 2^{J-1}$, which we denote by J^*. A quick calculation gives

$$J^* = \left\lfloor \frac{\ell_n \log 3 - \log 2}{\log 3 - \log 2} \right\rfloor.$$

After level J^*, the leftmost descendant of w'' has relative position $\leq 1/2$ by Lemma 2.3.14. Therefore, we need $n > J^*$. Taking

$$\ell_n = \left\lfloor \frac{\log 3/2}{\log 3} n \right\rfloor, \tag{2.3.14}$$

will do for n large enough, say $n \geq n_0$.

Finishing up. By construction, Φ_n is a cutset for all $n \geq n_0$. Moreover,

$$\sum_{e \in \Phi_n} \lambda^{-|e|} = 2^{n-1} \lambda^{-m_n} + \lambda^{-\ell_n} < \frac{1}{n}$$

for n large enough, where we used (2.3.13) and (2.3.14). Taking $n \to +\infty$ gives the claim. ■

As a consequence of Claims 2.3.11 and 2.3.13, $|\mathcal{C}_\rho| < +\infty$ almost surely for all $p < 1$ on $\widehat{\mathscr{T}}_{3-1}$. ◀

2.4 Chernoff–Cramér Method

Chebyshev's inequality (Theorem 2.1.2) gives a bound on the concentration around its mean of a square integrable random variable. It is, in general, best possible. Indeed, take X to be $\mu + b\sigma$ or $\mu - b\sigma$ with probability $(2b^2)^{-1}$ each, and μ otherwise. Then $\mathbb{E}X = \mu$, $\mathrm{Var}X = \sigma^2$, and for $\beta = b\sigma$,

$$\mathbb{P}[|X - \mathbb{E}X| \geq \beta] = \mathbb{P}[|X - \mathbb{E}X| = \beta] = \frac{1}{b^2} = \frac{\mathrm{Var}X}{\beta^2}.$$

However, in many cases, much stronger bounds can be derived. For instance, if $X \sim \mathrm{N}(0,1)$, by the following lemma

$$\mathbb{P}[|X - \mathbb{E}X| \geq \beta] \sim \sqrt{\frac{2}{\pi}}\beta^{-1}\exp(-\beta^2/2) \ll \frac{1}{\beta^2}, \tag{2.4.1}$$

as $\beta \to +\infty$. Indeed:

Lemma 2.4.1 *For $x > 0$,*

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

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

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

For the other direction, by differentiation,

$$\int_x^{+\infty}(1 - 3y^{-4})\, e^{-y^2/2}\mathrm{d}y = (x^{-1} - x^{-3})\, e^{-x^2/2}. \qquad \blacksquare$$

In this section, we discuss the Chernoff–Cramér method, which produces *exponential* tail bounds, provided the moment-generating function (see Section 2.1.1) 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 2.1.1) produces Chebyshev's inequality (Theorem 2.1.2). This "boosting" can be pushed further when stronger integrability conditions hold.

Chernoff–Cramér We refer to (2.4.2) in the next lemma as the *Chernoff–Cramér bound*.

CHERNOFF–CRAMÉR BOUND

Lemma 2.4.2 (Chernoff–Cramér bound). *Assume X is a 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 $s > 0$,*

$$\mathbb{P}[X \geq \beta] \leq \exp\left[-\{s\beta - \Psi_X(s)\}\right], \tag{2.4.2}$$

where

$$\Psi_X(s) := \log M_X(s)$$

is the cumulant-generating function of X.

Proof Exponentiating within Markov's inequality gives for $s > 0$,

$$\mathbb{P}[X \geq \beta] = \mathbb{P}[e^{sX} \geq e^{s\beta}] \leq \frac{M_X(s)}{e^{s\beta}} = \exp\left[-\{s\beta - \Psi_X(s)\}\right]. \qquad \blacksquare$$

Returning to the Gaussian case, let $X \sim N(0, \nu)$, where $\nu > 0$ is the variance and note that

$$\begin{aligned} M_X(s) &= \int_{-\infty}^{+\infty} e^{sx} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{x^2}{2\nu}}\, \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} e^{\frac{s^2\nu}{2}} \frac{1}{\sqrt{2\pi\nu}} e^{-\frac{(x-s\nu)^2}{2\nu}}\, \mathrm{d}x \\ &= \exp\left(\frac{s^2\nu}{2}\right). \end{aligned}$$

By straightforward calculus, the optimal choice of s in (2.4.2) gives the exponent

$$\sup_{s>0}(s\beta - s^2\nu/2) = \frac{\beta^2}{2\nu}, \tag{2.4.3}$$

achieved at $s_\beta = \beta/\nu$. For $\beta > 0$, this leads to the bound

$$\mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right), \tag{2.4.4}$$

which is much sharper than Chebyshev's inequality for large β – compare to (2.4.1).

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

RADEMACHER VARIABLE

Lemma 2.4.3 (Chernoff bound for simple random walk on $\mathbb{Z}$). *Let* $Z_1, \dots, Z_n$ *be independent* Rademacher variables, *that is, they are* $\{-1, 1\}$*-valued random variables with* $\mathbb{P}[Z_i = 1] = \mathbb{P}[Z_i = -1] = 1/2$. *Let* $S_n = \sum_{i \leq n} Z_i$. *Then, for any* $\beta > 0$,

$$\mathbb{P}[S_n \geq \beta] \leq e^{-\beta^2/2n}. \tag{2.4.5}$$

Proof The moment-generating function of Z_1 can be bounded as follows

$$M_{Z_1}(s) = \frac{e^s + e^{-s}}{2} = \sum_{j \geq 0} \frac{s^{2j}}{(2j)!} \leq \sum_{j \geq 0} \frac{(s^2/2)^j}{j!} = e^{s^2/2}. \tag{2.4.6}$$

Taking $s = \beta/n$ in the Chernoff–Cramér bound (2.4.2), we get

$$\begin{aligned} \mathbb{P}[S_n \geq \beta] &\leq \exp\left(-s\beta + n\Psi_{Z_1}(s)\right) \\ &\leq \exp\left(-s\beta + ns^2/2\right) \\ &= e^{-\beta^2/2n}, \end{aligned}$$

which concludes the proof. ■

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Observe the similarity between (2.4.5) and the Gaussian bound (2.4.4) if one takes ν to be the variance of S_n, that is,

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

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 2.4.3 will be substantially extended in Section 2.4.2.

Example 2.4.4 (Set balancing). Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ be arbitrary non-zero vectors in $\{0,1\}^n$. Think of $\boldsymbol{v}_i = (v_{i,1}, \ldots, v_{i,n})$ as representing a subset 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 $\boldsymbol{v}_i$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 $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{-1,+1\}^n$ such that $B^* = \max_{i=1,\ldots,m} |\boldsymbol{x} \cdot \boldsymbol{v}_i|$ is as small as possible.

A simple random choice does well: select each x_i independently, uniformly at random in $\{-1,+1\}$. Fix $\varepsilon > 0$. We claim that

$$\mathbb{P}\left[B^* \geq \sqrt{2n(\log m + \log(2\varepsilon^{-1}))}\right] \leq \varepsilon. \tag{2.4.7}$$

Indeed, by (2.4.5) (considering only the non-zero entries of $\boldsymbol{v}_i$),

$$\begin{aligned}\mathbb{P}&\left[|\boldsymbol{x} \cdot \boldsymbol{v}_i| \geq \sqrt{2n(\log m + \log(2\varepsilon^{-1}))}\right] \\ &\leq 2\exp\left(-\frac{2n(\log m + \log(2\varepsilon^{-1}))}{2\|\boldsymbol{v}_i\|_1}\right) \\ &\leq \frac{\varepsilon}{m},\end{aligned}$$

where we used that $\|\boldsymbol{v}_i\|_1 \leq n$. Taking a union bound over the m vectors gives the result. In (2.4.7), the $\sqrt{n}$ term on the right-hand side of the inequality is to be expected since it is the standard deviation of $|\boldsymbol{x} \cdot \boldsymbol{v}_i|$ in the worst case. The power of the exponential tail bound (2.4.5) appears in the logarithmic terms, which would have been much larger if one had used Chebyshev's inequality instead. ◀

The Chernoff–Cramér bound is particularly useful for sums of independent random variables as the moment-generating function then factorizes; see (2.1.3). Let

$$\Psi_X^*(\beta) = \sup_{s \in \mathbb{R}_+}(s\beta - \Psi_X(s))$$

be the *Fenchel–Legendre dual* of the cumulant-generating function of X. FENCHEL–LEGENDRE DUAL

Theorem 2.4.5 (Chernoff–Cramér method). *Let $S_n = \sum_{i \leq n} X_i$, where the X_is are i.i.d. random variables. Assume $M_{X_1}(s) < +\infty$ on $s \in (-s_0, s_0)$ for some $s_0 > 0$. For any $\beta > 0$,*

$$\mathbb{P}[S_n \geq \beta] \leq \exp\left(-n\Psi_{X_1}^*\left(\frac{\beta}{n}\right)\right). \tag{2.4.8}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

In particular, in the large deviations regime, that is, when $\beta = bn$ for some $b > 0$, we have

$$-\limsup_n \frac{1}{n} \log \mathbb{P}[S_n \geq bn] \geq \Psi^*_{X_1}(b)\,. \tag{2.4.9}$$

Proof By independence, we get

$$\Psi^*_{S_n}(\beta) = \sup_{s>0}(s\beta - n\Psi_{X_1}(s)) = \sup_{s>0} n\left(s\left(\frac{\beta}{n}\right) - \Psi_{X_1}(s)\right) = n\Psi^*_{X_1}\left(\frac{\beta}{n}\right),$$

and then we optimize over s in (2.4.2). ■

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

POISSON

Poisson variables We start with the Poisson case. Let $Z \sim \mathrm{Poi}(\lambda)$ be Poisson with mean λ, where we recall that

$$\mathbb{P}[Z = k] = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{Z}_+.$$

Letting $X = Z - \lambda$,

$$\begin{aligned}\Psi_X(s) &= \log\left(\sum_{\ell\geq 0} e^{-\lambda}\frac{\lambda^\ell}{\ell!}e^{s(\ell-\lambda)}\right)\\ &= \log\left(e^{-(1+s)\lambda}\sum_{\ell\geq 0}\frac{(e^s\lambda)^\ell}{\ell!}\right)\\ &= \log\left(e^{-(1+s)\lambda}e^{e^s\lambda}\right)\\ &= \lambda(e^s - s - 1),\end{aligned}$$

so that straightforward calculus gives for $\beta > 0$,

$$\begin{aligned}\Psi^*_X(\beta) &= \sup_{s>0}(s\beta - \lambda(e^s - s - 1))\\ &= \lambda\left[\left(1+\frac{\beta}{\lambda}\right)\log\left(1+\frac{\beta}{\lambda}\right) - \frac{\beta}{\lambda}\right]\\ &=: \lambda\, h\left(\frac{\beta}{\lambda}\right),\end{aligned}$$

achieved at $s_\beta = \log\left(1 + \frac{\beta}{\lambda}\right)$, where h is defined as the expression in square brackets in the above display. Plugging $\Psi^*_X(\beta)$ into Theorem 2.4.5 leads for $\beta > 0$ to the bound

$$\mathbb{P}[Z \geq \lambda + \beta] \leq \exp\left(-\lambda\, h\left(\frac{\beta}{\lambda}\right)\right). \tag{2.4.10}$$

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

$$\mathbb{P}[Z \leq \lambda + \beta] \leq \exp\left(-\lambda\, h\left(\frac{\beta}{\lambda}\right)\right). \tag{2.4.11}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

If S_n is a sum of n i.i.d. Poi(λ) variables, then by (2.4.9) for $a > \lambda$,

$$\begin{aligned} -\limsup_n \frac{1}{n} \log \mathbb{P}[S_n \geq an] &\geq \lambda\, h\left(\frac{a-\lambda}{\lambda}\right) \\ &= a \log\left(\frac{a}{\lambda}\right) - a + \lambda \\ &=: I_\lambda^{\mathrm{Poi}}(a), \end{aligned} \tag{2.4.12}$$

and similarly for $a < \lambda$,

$$-\limsup_n \frac{1}{n} \log \mathbb{P}[S_n \leq an] \geq I_\lambda^{\mathrm{Poi}}(a). \tag{2.4.13}$$

In fact, these bounds follow immediately from (2.4.10) and (2.4.11) by noting that $S_n \sim$ Poi($n\lambda$) (see, for example, Exercise 6.7).

Binomial variables and Chernoff bounds Let $Z \sim \mathrm{Bin}(n,p)$ be a binomial random variable with parameters n and p. Recall that Z is a sum of i.i.d. indicators $Y_1, \ldots, Y_n$ equal to 1 with probability p. The Y_is are also known as Bernoulli random variables or Bernoulli trials, and their law is denoted by Ber(p). We also refer to p as the success probability. Letting $X_i = Y_i - p$ and $S_n = Z - np$,

BINOMIAL

BERNOULLI

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

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

$$\begin{aligned} \Psi_{X_1}^*(b) &= \sup_{s>0}(sb - (\log\left[pe^s + (1-p)\right] - ps)) \\ &= (1-a)\log\frac{1-a}{1-p} + a \log\frac{a}{p} =: D(a\|p), \end{aligned} \tag{2.4.14}$$

achieved at $s_b = \log\frac{(1-p)a}{p(1-a)}$. The function $D(a\|p)$ in (2.4.14) is the so-called *Kullback–Leibler divergence* or *relative entropy* between two Bernoulli variables with parameters a and p, respectively. By (2.4.8) for $\beta > 0$,

KULLBACK–LEIBLER DIVERGENCE

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

Applying the same argument to $Z' = n - Z$ gives a bound in the other direction.

Remark 2.4.6 *In the large deviations regime, it can be shown that the previous bound is tight in the sense that*

$$-\frac{1}{n}\log \mathbb{P}[Z \geq np + bn] \to D\left(p + b \| p\right) =: I_{n,p}^{\mathrm{Bin}}(b),$$

as $n \to +\infty$. The theory of large deviations provides general results of this type. See, for example, [Dur10, section 2.6]. Upper bounds will be enough for our purposes.

The following related bounds, proved in Exercise 2.7, are often useful.

Theorem 2.4.7 (Chernoff bounds for Poisson trials). *Let $Y_1, \ldots, Y_n$ be independent* $\{0,1\}$-*valued random variables with $\mathbb{P}[Y_i = 1] = p_i$ and $\mu = \sum_i p_i$. These are called* Poisson trials. *Let $Z = \sum_i Y_i$. Then:*

POISSON TRIALS

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

(i) ***Above the mean***

(a) *For any* $\delta > 0$*,*

$$\mathbb{P}[Z \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}.$$

(b) *For any* $0 < \delta \leq 1$*,*

$$\mathbb{P}[Z \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/3}.$$

(ii) ***Below the mean***

(a) *For any* $0 < \delta < 1$*,*

$$\mathbb{P}[Z \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}.$$

(b) *For any* $0 < \delta < 1$*,*

$$\mathbb{P}[Z \leq (1-\delta)\mu] \leq e^{-\mu\delta^2/2}.$$

2.4.2 Sub-Gaussian and Sub-Exponential Random Variables

The bounds in Section 2.4.1 were obtained by computing the moment-generating function explicitly (possibly with some approximations). This is not always possible. In this section, we give some important examples of tail bounds derived from the Chernoff–Cramér method for broad classes of random variables under natural conditions on their distributions.

Sub-Gaussian random variables

We begin with sub-Gaussian random variables which, as the name suggests, have a tail that is bounded by that of a Gaussian.

General case Here is our key definition.

SUB-GAUSSIAN VARIABLE

Definition 2.4.8 (Sub-Gaussian random variables). *We say that a random variable X with mean μ is* sub-Gaussian with variance factor ν *if*

$$\Psi_{X-\mu}(s) \leq \frac{s^2\nu}{2} \quad \forall s \in \mathbb{R} \tag{2.4.15}$$

for some $\nu > 0$. We use the notation $X \in \mathrm{s}\mathcal{G}(\nu)$.

Note that the right-hand side in (2.4.15) is the cumulant-generating function of a $\mathrm{N}(0,\nu)$. By the Chernoff–Cramér method and (2.4.3) it follows immediately that

$$\mathbb{P}[X-\mu \leq -\beta] \vee \mathbb{P}[X-\mu \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right), \tag{2.4.16}$$

where we used that $X \in \mathrm{s}\mathcal{G}(\nu)$ implies $-X \in \mathrm{s}\mathcal{G}(\nu)$. As a quick example, note that this is the approach we took in Lemma 2.4.3, that is, we showed that a uniform random variable in $\{-1,1\}$ (i.e., a Rademacher variable) is sub-Gaussian with variance factor 1.

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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Theorem 2.4.9 (General Hoeffding inequality). *Suppose $X_1, \ldots, X_n$ are independent random variables where, for each i, $X_i \in \mathrm{s}\mathcal{G}(\nu_i)$ with $0 < \nu_i < +\infty$. For $w_1, \ldots, w_n \in \mathbb{R}$, let $S_n = \sum_{i \leq n} w_i X_i$. Then*

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

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

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

Proof Assume the X_is are centered. By independence and (2.1.3),

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

Bounded random variables For bounded random variables, the previous inequality reduces to a standard bound.

Theorem 2.4.10 (Hoeffding's inequality for bounded variables). *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 2.4.9, it suffices to show that $X_i - \mathbb{E}X_i \in \mathrm{s}\mathcal{G}(\nu_i)$ 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_is 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,

SYMMETRIZATION

$$\begin{aligned}\mathbb{E}\left[e^{sX_i}\right] &= \mathbb{E}\left[e^{s\mathbb{E}[X_i - X_i' \mid X_i]}\right]\\ &\leq \mathbb{E}\left[\mathbb{E}\left[e^{s(X_i - X_i')} \,\middle|\, X_i\right]\right]\\ &= \mathbb{E}\left[e^{s(X_i - X_i')}\right],\end{aligned}$$

where the first line comes from the taking out what is known lemma (Lemma B.6.16) and the fact that X_i' is centered and independent of X_i, the second line follows from the conditional Jensen's inequality (Lemma B.6.12), and the third line uses the tower property (Lemma B.6.16). Observe that $X_i - X_i'$ is *symmetric*, that is, identically distributed to $-(X_i - X_i')$. Hence, using that Z_i is independent of both X_i and X_i', we get

$$\begin{aligned}\mathbb{E}\left[e^{s(X_i - X_i')}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{s(X_i - X_i')} \,\middle|\, Z_i\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[e^{sZ_i(X_i - X_i')} \,\middle|\, Z_i\right]\right]\\ &= \mathbb{E}\left[e^{sZ_i(X_i - X_i')}\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[e^{sZ_i(X_i - X_i')} \,\middle|\, X_i - X_i'\right]\right].\end{aligned}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

From (2.4.6) (together with Lemma B.6.15), the last line is

$$\begin{aligned} &\le \mathbb{E}\left[e^{(s(X_i-X_i'))^2/2}\right] \\ &\le e^{-4c_i^2s^2/2} \end{aligned}$$

since $|X_i|, |X_i'| \le c_i$. Putting everything together, we arrive at

$$\mathbb{E}\left[e^{sX_i}\right] \le e^{-4c_i^2s^2/2}.$$

That is, X_i is sub-Gaussian with variance factor $4c_i^2$. By Theorem 2.4.9, S_n is sub-Gaussian with variance factor $\sum_{i\le n} 4c_i^2$ and

$$\mathbb{P}[S_n \ge t] \le \exp\left(-\frac{t^2}{8\sum_{i\le n} c_i^2}\right).$$

Proof of Theorem 2.4.10 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 below (which we will use again in Chapter 3). First an observation:

Lemma 2.4.11 (Variance of bounded random variables). *For any random variable Z taking values in $[a, b]$ with $-\infty < a \le b < +\infty$, we have*

$$\mathrm{Var}[Z] \le \frac{1}{4}(b-a)^2.$$

Proof Indeed,

$$\left|Z - \frac{a+b}{2}\right| \le \frac{b-a}{2}$$

and

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

■

HOEFFDING'S LEMMA

Lemma 2.4.12 (Hoeffding's lemma). *Let X be a random variable taking values in $[a, b]$ for $-\infty < a \le b < +\infty$. Then, $X \in \mathrm{s}\mathcal{G}\left(\frac{1}{4}(b-a)^2\right)$.*

Proof Note first that $X - \mathbb{E}X \in [a - \mathbb{E}X, b - \mathbb{E}X]$ and $\frac{1}{4}((b-\mathbb{E}X)-(a-\mathbb{E}X))^2 = \frac{1}{4}(b-a)^2$. So without loss of generality we assume that $\mathbb{E}X = 0$. Because X is bounded, $M_X(s)$ is finite for all $s \in \mathbb{R}$. Hence, by (2.1.2),

$$\Psi_X(0) = \log M_X(0) = 0, \qquad \Psi_X'(0) = \frac{M_X'(0)}{M_X(0)} = \mathbb{E}X = 0,$$

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,

$$\Psi_X''(s) \le \frac{1}{4}(b-a)^2. \tag{2.4.17}$$

Note that

$$
\begin{aligned}
\Psi_X''(s) &= \frac{M_X''(s)}{M_X(s)} - \left(\frac{M_X'(s)}{M_X(s)}\right)^2 \\
&= \frac{1}{M_X(s)}\mathbb{E}\left[X^2 e^{sX}\right] - \left(\frac{1}{M_X(s)}\mathbb{E}\left[Xe^{sX}\right]\right)^2 \\
&= \mathbb{E}\left[X^2 \frac{e^{sX}}{M_X(s)}\right] - \left(\mathbb{E}\left[X\frac{e^{sX}}{M_X(s)}\right]\right)^2.
\end{aligned}
$$

The trick to conclude is to notice that $\frac{e^{sx}}{M_X(s)}$ defines a density 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 2.4.11. This establishes (2.4.17) and concludes the proof. ■

Remark 2.4.13 *The change of measure above is known as* tilting *and is a standard trick in large deviation theory. See, for example, [Dur10, section 2.6].*

Since we have shown that $X_i - \mathbb{E}X_i$ is sub-Gaussian with variance factor $\frac{1}{4}(b_i - a_i)^2$, Theorem 2.4.10 follows from Theorem 2.4.9. ■

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 χ^2-distributed and its moment-generating function can be computed explicitly. Using the change of variable $u = x\sqrt{1-2s}$, for $s < 1/2$,

$$
\begin{aligned}
M_W(s) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{sx^2} e^{-x^2/2}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{1-2s}} \times \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-u^2/2}\,\mathrm{d}u \\
&= \frac{1}{(1-2s)^{1/2}}.
\end{aligned}
\tag{2.4.18}
$$

When $s \geq 1/2$, however, we clearly have $M_W(s) = +\infty$. 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 function which does not prevent it from diverging.) Further confirming this observation, arguing as in (2.4.1), the upper tail of W decays as

$$
\begin{aligned}
\mathbb{P}[W \geq \beta] &= \mathbb{P}[X \geq \sqrt{\beta}] \\
&\sim \sqrt{\frac{1}{2\pi}}[\sqrt{\beta}]^{-1}\exp(-[\sqrt{\beta}]^2/2) \\
&\sim \sqrt{\frac{1}{2\pi\beta}}\exp(-\beta/2),
\end{aligned}
$$

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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

General case We now define a broad class of distributions which have such exponential tail decay.

SUB-EXPONENTIAL VARIABLE

Definition 2.4.14 (Sub-exponential random variable). *We say that a random variable X with mean μ is* sub-exponential with parameters (ν, α) *if*

$$\Psi_{X-\mu}(s) \leq \frac{s^2\nu}{2} \quad \forall |s| \leq \frac{1}{\alpha} \tag{2.4.19}$$

for some $\nu, \alpha > 0$. We write $X \in \mathrm{s}\mathcal{E}(\nu, \alpha)$.[1]

Observe that the key difference between (2.4.15) and (2.4.19) is the interval of s over which it holds. As we will see, the parameter α dictates the exponential decay rate of the tail. The specific form of the bound in (2.4.19) is natural once one notices that, as $|s| \to 0$, a centered random variable with variance ν should roughly satisfy

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

Returning to the χ^2 distribution, note that from (2.4.18) we have for $|s| \leq 1/4$:

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

Hence, $W \in \mathrm{s}\mathcal{E}(4,4)$.

Using the Chernoff–Cramér bound (Lemma 2.4.2), we obtain the following tail bound for sub-exponential variables.

Theorem 2.4.15 (Sub-exponential tail bound). *Suppose the random variable X with mean μ is sub-exponential with parameters (ν, α). Then, for all $\beta \in \mathbb{R}_+$,*

$$\mathbb{P}[X - \mu \geq \beta] \leq \begin{cases} \exp(-\frac{\beta^2}{2\nu}) & \text{if } 0 \leq \beta \leq \nu/\alpha, \\ \exp(-\frac{\beta}{2\alpha}) & \text{if } \beta > \nu/\alpha. \end{cases} \tag{2.4.20}$$

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

Proof of Theorem 2.4.15 We start by applying the Chernoff–Cramér bound. For any $\beta > 0$ and $|s| \leq 1/\alpha$,

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

[1] More commonly, "sub-exponential" refers to the case $\alpha = \sqrt{\nu}$.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

At this point, the proof diverges from the sub-Gaussian case because the optimal choice of s *depends on* β 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

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

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

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

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

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

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

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

Proof Assume the X_is are centered. By independence and (2.1.3),

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

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

$$|s| \leq \frac{1}{\max_i |w_i|\alpha_i}.$$

■

Bounded random variables: revisited We apply the previous result to bounded random variables.

Theorem 2.4.17 (Bernstein's inequality for bounded variables). *Let* $X_1, \ldots, X_n$ *be independent random variables, where, for each i,* X_i *has mean* μ_i, *variance* ν_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* $\beta > 0$,

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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

Proof We claim that $X_i \in \mathrm{s}\mathcal{E}(2\nu_i, 2c)$. To establish the claim, we derive a bound on all moments of X_i. Note that for all integers $k \geq 2$,

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

Hence, first applying the dominated convergence theorem (Proposition B.4.14) to establish the limit, we have for $|s| \leq \frac{1}{2c}$,

$$\begin{aligned}
\mathbb{E}[e^{s(X_i-\mu_i)}] &= \sum_{k=0}^{+\infty} \frac{s^k}{k!}\mathbb{E}[(X_i - \mu_i)^k] \\
&\leq 1 + s\,\mathbb{E}[(X_i - \mu_i)] + \sum_{k=2}^{+\infty} \frac{s^k}{k!}c^{k-2}\nu_i \\
&\leq 1 + \frac{s^2\nu_i}{2} + \frac{s^2\nu_i}{3!}\sum_{k=3}^{+\infty}(cs)^{k-2} \\
&= 1 + \frac{s^2\nu_i}{2}\left\{1 + \frac{1}{3}\frac{cs}{1-cs}\right\} \\
&\leq 1 + \frac{s^2\nu_i}{2}\left\{1 + \frac{1}{3}\frac{1/2}{1-1/2}\right\} \\
&\leq 1 + \frac{s^2}{2}2\nu_i \\
&\leq \exp\left(\frac{s^2}{2}2\nu_i\right).
\end{aligned}$$

Using the general Bernstein inequality (Theorem 2.4.16) 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 2.4.17 than we did in Theorem 2.4.10 – *when* β *is not too large*. That improvement stems from the use of the (actual) variance for moderate deviations. This is easier to appreciate through an example.

Example 2.4.18 (Erdős–Rényi: maximum degree). Let $G_n = (V_n, E_n) \sim \mathbb{G}_{n,p_n}$ be a random graph with n vertices and density p_n under the Erdős–Rényi model (Definition 1.2.2). 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, that is, $\delta(v) = |N(v)|$. 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 $\mathrm{Bin}(n-1, p_n)$ by independence of the edges. In particular, its expected degree is $(n-1)p_n$. To prove

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

a high-probability upper bound on the maximum D_n, we need to control the deviation of the degree of each vertex 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 2.4.19 *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 $\delta(v) = S_{n-1} \sim \mathrm{Bin}(n-1, p_n)$ 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 is a bounded random variable. The mean of X_i is p_n and its variance is $p_n(1-p_n)$. So in Bernstein's inequality (Theorem 2.4.17), 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. We choose β to be the smallest value 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, $\beta = o(np_n)$ (i.e., the deviation is much smaller than the expectation). 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 2.4.10) in the proof of Claim 2.4.19 we would have had to take $\beta = \sqrt{(1+\varepsilon)n\log n}$. 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. ◀

2.4.3 ▹ Probabilistic Analysis of Algorithms: Knapsack Problem

In a knapsack problem, we have n items. Item i has weight W_i and value V_i. Given a weight bound $\mathcal{W}$, we want to pack as valuable a collection of items in the knapsack under the constraint that the total weight is less than or equal to $\mathcal{W}$. Formally, we seek a solution to the optimization problem

$$Z^* = \max\left\{\sum_{j=1}^{n} x_j V_j : x_1, \ldots, x_n \in [0,1], \sum_{j=1}^{n} x_j W_j \leq \mathcal{W}\right\}. \tag{2.4.21}$$

This is the *fractional knapsack problem*, where we allow a fraction of an item to be added to the knapsack.

KNAPSACK PROBLEM

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

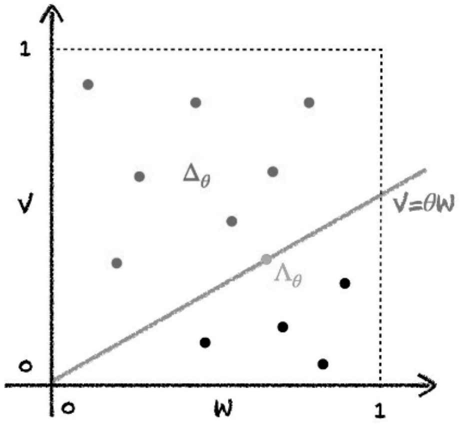

Figure 2.7 Vizualization of the greedy algorithm.

It is used as a computationally tractable relaxation of the 0-1 *knapsack problem*, which also includes the combinatorial constraint $x_j \in \{0,1\}, \forall j$. Indeed, it turns out that the optimization problem (2.4.21) is solved exactly by a simple greedy solution (see Exercise 2.8 for a formal proof of correctness): let π be a permutation of $\{1,\ldots,n\}$ that puts the items in decreasing order of value per unit weight

$$\frac{V_{\pi(1)}}{W_{\pi(1)}} \geq \frac{V_{\pi(2)}}{W_{\pi(2)}} \geq \cdots \geq \frac{V_{\pi(n)}}{W_{\pi(n)}};$$

add the items in that order until the first time the weight constraints is violated; include whatever fraction of that last item that will fit. This greedy algorithm has a natural geometric interpretation, depicted in Figure 2.7, that will be useful. We associate item j to a point $(W_j, V_j) \in [0,1]^2$ and keep only those items falling on or above a line with slope θ chosen to satisfy the total weight constraint. Specifically, let

$$\Delta_\theta = \left\{j \in [n]\colon V_j > \theta W_j\right\},$$

$$\Lambda_\theta = \left\{j \in [n] \,:\, V_j = \theta W_j\right\},$$

and

$$\Theta^* = \inf\left\{\theta \geq 0 \,:\, W_{\Delta_\theta} < \mathcal{W}\right\},$$

where, for a subset of items $J \subset [n]$, $W_J = \sum_{j\in J} W_j$.

We consider a stochastic version of the fractional knapsack problem where the weights and values are i.i.d. random variables picked uniformly at random in $[0,1]$. Characterizing Z^* (e.g., its moments or distribution) is not straightforward. Here we show that Z^* is highly concentrated around a natural quantity. Observe that, under our probabilistic model, almost surely $|\Lambda_\theta| \in \{0,1\}$ for any $\theta \geq 0$. Hence, there are two cases. Either $\Theta^* = 0$, in which case all items fit in the knapsack so that $Z^* = \sum_{j=1}^n W_j$, or $\Theta^* > 0$, in which case $|\Lambda_{\Theta^*}| = 1$ and

$$Z^* = V_{\Delta_{\Theta^*}} + (\mathcal{W} - W_{\Delta_{\Theta^*}})V_{\Lambda_{\Theta^*}}. \tag{2.4.22}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

One interesting regime is $\mathcal{W} = \tau n$ for some constant $\tau > 0$. Clearly, $\tau > 1$ is trivial. In fact, because

$$\mathbb{E}\left[\sum_{j=1}^{n} W_j\right] = n\,\mathbb{E}[W_1] = \frac{1}{2}n,$$

we assume that $\tau \le 1/2$. To further simplify the calculations, we restrict ourselves to the case $\tau \in (1/6, 1/2)$. (See Exercise 2.8 for the remaining case.) In this regime, we show that Z^* grows linearly with n and give a bound on its deviation.

Although Z^* is technically a sum of random variables, the choice of Θ^* correlates them and we cannot apply our concentration bounds directly. Instead, we show that Θ^* itself can be controlled well. It is natural to conjecture that Θ^* is approximately equal to a solution θ_τ of the expected constraint equation $\mathbb{E}[W_{\Delta_{\theta_\tau}}] = \mathcal{W}$, that is,

$$n\bar{w}_{\theta_\tau} = n\tau, \tag{2.4.23}$$

where $\bar{w}_\theta$ is defined through

$$\begin{aligned}
\mathbb{E}[W_{\Delta_\theta}] &= \mathbb{E}\left[\sum_{j\in\Delta_\theta} W_j\right]\\
&= \mathbb{E}\left[\sum_{j=1}^{n} \mathbf{1}\{V_j > \theta W_j\}\, W_j\right]\\
&= n\mathbb{E}\left[\mathbf{1}\{V_1 > \theta W_1\}\, W_1\right]\\
&=: n\bar{w}_\theta.
\end{aligned}$$

Similarly, we define

$$\bar{v}_\theta := \mathbb{E}\left[\mathbf{1}\{V_1 > \theta W_1\}\, V_1\right].$$

We see directly from the definitions that both $\bar{w}_\theta$ and $\bar{v}_\theta$ are monotone as functions of θ.

Our main claim is the following.

Claim 2.4.20 *There is a constant $c > 0$ such that for any $\delta > 0$,*

$$\mathbb{P}\left[|Z^* - n\bar{v}_{\theta_\tau}| \ge \sqrt{cn\log\delta^{-1}}\right] \le \delta$$

for all n large enough.

Proof Because all weights and values are in [0, 1], it follows from (2.4.22) that

$$V_{\Delta_{\Theta^*}} \le Z^* \le V_{\Delta_{\Theta^*}} + 1, \tag{2.4.24}$$

and it will suffice to work with $V_{\Delta_{\Theta^*}}$. The idea of the proof is to show that Θ^* is close to θ_τ by establishing that W_{Δ_θ} is highly likely to be less than τn when $\theta > \theta_\tau$, while the opposite holds when $\theta < \theta_\tau$. For this, we view W_{Δ_θ} as a sum of independent bounded random variables and use Hoeffding's inequality (Theorem 2.4.10).

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Controlling Θ^*. First, it will be useful to compute $\bar{w}_\theta$ and θ_τ analytically. By definition,

$$\begin{aligned}\bar{w}_\theta &= \mathbb{E}\left[\mathbf{1}\{V_1 > \theta W_1\} W_1\right] \\ &= \int_0^1 \int_0^1 \mathbf{1}\{y > \theta x\} x \,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^{1\wedge 1/\theta} (1-\theta x)x\,\mathrm{d}x \\ &= \begin{cases} \frac{1}{2} - \frac{1}{3}\theta & \text{if } \theta \le 1, \\ \frac{1}{6\theta^2} & \text{otherwise.}\end{cases}\end{aligned} \tag{2.4.25}$$

Plugging back into (2.4.23), we get the unique solution

$$\theta_\tau := 3\left(\frac{1}{2} - \tau\right) \in (0,1)$$

for the range $\tau \in (1/6, 1/2)$.

Now observe that, for each fixed θ, the quantity

$$W_{\Delta_\theta} = \sum_{j=1}^n \mathbf{1}\{V_j > \theta W_j\} W_j$$

is a sum of independent random variables taking values in $[0,1]$. Hence, for any $\beta > 0$, Hoeffding's inequality gives

$$\mathbb{P}\left[W_{\Delta_\theta} - n\bar{w}_\theta \ge \beta\right] \le \exp\left(-\frac{2\beta^2}{n}\right).$$

Using this inequality with $\theta = \theta_\tau - \frac{C}{\sqrt{n}}$ (with n large enough that $\theta < 1$) and $\beta = 3C\sqrt{n}$ gives

$$\mathbb{P}\left[W_{\Delta_{\theta_\tau - \frac{C}{\sqrt{n}}}} - 3n\left(\frac{1}{2} - \theta_\tau + \frac{C}{\sqrt{n}}\right) \ge 3C\sqrt{n}\right] \le \exp\left(-2(3C)^2\right),$$

where we used (2.4.25). After rearranging and using that $3n\left(\frac{1}{2} - \theta_\tau\right) = n\tau$ by (2.4.23) and (2.4.25), we get

$$\mathbb{P}\left[\Theta^* \ge \theta_\tau - \frac{C}{\sqrt{n}}\right] = \mathbb{P}\left[W_{\Delta_{\theta_\tau - \frac{C}{\sqrt{n}}}} \ge n\tau\right] \le \exp\left(-2(3C)^2\right).$$

Applying the same argument to $-W_{\Delta_\theta}$ with $\theta = \theta_\tau + \frac{C}{\sqrt{n}}$ and combining with the previous inequality gives

$$\mathbb{P}\left[|\Theta^* - \theta_\tau| > \frac{C}{\sqrt{n}}\right] \le 2\exp\left(-2(3C)^2\right). \tag{2.4.26}$$

Controlling Z^*. We conclude by applying Hoeffding's inequality to V_{Δ_θ}. Arguing as above with the same θ's and β, we obtain

$$\mathbb{P}\left[V_{\Delta_{\theta_\tau - \frac{C}{\sqrt{n}}}} - n\bar{v}_{\theta_\tau - \frac{C}{\sqrt{n}}} \ge 3C\sqrt{n}\right] \le \exp\left(-2(3C)^2\right) \tag{2.4.27}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

and

$$\mathbb{P}\left[V_{\Delta_{\theta_\tau+\frac{C}{\sqrt{n}}}} - n\bar{v}_{\theta_\tau+\frac{C}{\sqrt{n}}} \le -3C\sqrt{n}\right] \le \exp\left(-2(3C)^2\right). \tag{2.4.28}$$

Again, it will be useful to compute $\bar{v}_\theta$ analytically. By definition,

$$\begin{aligned}\bar{v}_\theta &= \mathbb{E}\left[\mathbf{1}\{V_1 > \theta W_1\}\, V_1\right]\\ &= \int_0^1\int_0^1 \mathbf{1}\{y > \theta x\} y\, \mathrm{d}x\, \mathrm{d}y\\ &= \int_0^{1\wedge\theta} \frac{y^2}{\theta}\,\mathrm{d}y + \int_{1\wedge\theta}^1 y\,\mathrm{d}y\\ &= \begin{cases}\frac{1}{2} - \frac{1}{6}\theta^2 & \text{if } \theta \le 1,\\ \frac{1}{3\theta} & \text{otherwise.}\end{cases}\end{aligned}$$

Assuming n is large enough that $\theta_\tau + C/\sqrt{n} < 1$ (recall that $\theta_\tau < 1$), we get

$$\bar{v}_{\theta_\tau} - \bar{v}_{\theta_\tau+\frac{C}{\sqrt{n}}} = \frac{1}{6}\left(2\frac{C}{\sqrt{n}}\theta_\tau + \frac{C^2}{n}\right) \le \frac{C}{\sqrt{n}}.$$

A quick check reveals that, similarly, $\bar{v}_{\theta_\tau-\frac{C}{\sqrt{n}}} - \bar{v}_{\theta_\tau} \le \frac{C}{\sqrt{n}}$. Plugging back into (2.4.27) and (2.4.28) gives

$$\mathbb{P}\left[V_{\Delta_{\theta_\tau-\frac{C}{\sqrt{n}}}} \ge n\bar{v}_{\theta_\tau} + 4C\sqrt{n}\right] \le \exp\left(-2(3C)^2\right) \tag{2.4.29}$$

and

$$\mathbb{P}\left[V_{\Delta_{\theta_\tau+\frac{C}{\sqrt{n}}}} \le n\bar{v}_{\theta_\tau} - 4C\sqrt{n}\right] \le \exp\left(-2(3C)^2\right). \tag{2.4.30}$$

Observe that the following monotonicity property holds almost surely

$$\theta_0 \le \theta_1 \le \theta_2 \implies V_{\Delta_{\theta_0}} \ge V_{\Delta_{\theta_1}} \ge V_{\Delta_{\theta_2}}. \tag{2.4.31}$$

Combining (2.4.24), (2.4.26), (2.4.29), (2.4.30), and (2.4.31), we obtain

$$\mathbb{P}\left[\left|Z^* - n\bar{v}_{\theta_\tau}\right| > 5C\sqrt{n}\right] \le 4\exp\left(-2(3C)^2\right)$$

for n large enough. Choosing C appropriately gives the claim. ∎

A similar bound is proved for the 0-1 knapsack problem in Exercise 2.9.

2.4.4 Epsilon-Nets and Chaining

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

$$\sup_{t\in\mathscr{T}} X_t,$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

where $\mathscr{T}$ is an arbitrary index set and the X_ts are real-valued random variables. To avoid measurability issues, we assume throughout that $\mathscr{T}$ is countable.[2] Note that t does not in general need to be a “time” index.

So far we have developed tools that can handle cases where $\mathscr{T}$ 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: $\mathscr{T}$ is finite and X_t is sub-Gaussian.

Theorem 2.4.21 (Maximal inequalities: sub-Gaussian case). *Let $\{X_t\}_{t\in\mathscr{T}}$ be a stochastic process with finite index set $\mathscr{T}$. Assume that there is $\nu > 0$ such that, for all t, $X_t \in s\mathcal{G}(\nu)$ and $\mathbb{E}[X_t] = 0$. Then,*

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

and, for all $\beta > 0$,

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

Proof For the expectation, we apply a variation on the Chernoff–Cramér method (Section 2.4). Naively, we could bound the supremum $\sup_{t\in\mathscr{T}} X_t$ by the sum $\sum_{t\in\mathscr{T}} |X_t|$, but that would lead to a bound growing linearly with the cardinality $|\mathscr{T}|$. Instead we first take an exponential, which tends to amplify the largest term and produces a much stronger bound. Specifically, by Jensen’s inequality (Theorem B.4.15), for any $s > 0$,

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

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

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

[2] Technically, it suffices to assume that there is a countable $\mathscr{T}_0 \subseteq \mathscr{T}$ such that $\sup_{t\in\mathscr{T}} X_t = \sup_{t\in\mathscr{T}_0} X_t$ almost surely.

The optimal choice of s (i.e., leading to the least bound) is when the two terms in the sum above are equal, that is, $s = \sqrt{2\nu^{-1} \log |\mathscr{T}|}$, which gives finally

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

as claimed.

For the tail inequality, we use a union bound and (2.4.16):

$$\begin{aligned}\mathbb{P}\left[\sup_{t \in \mathscr{T}} X_t \ge \sqrt{2\nu \log |\mathscr{T}|} + \beta\right] &\le \sum_{t \in \mathscr{T}} \mathbb{P}\left[X_t \ge \sqrt{2\nu \log |\mathscr{T}|} + \beta\right] \\ &\le |\mathscr{T}| \exp\left(-\frac{(\sqrt{2\nu \log |\mathscr{T}|} + \beta)^2}{2\nu}\right) \\ &\le \exp\left(-\frac{\beta^2}{2\nu}\right),\end{aligned}$$

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 ε-net. This notion requires that a pseudometric ρ (i.e., $\rho\colon \mathscr{T} \times \mathscr{T} \to \mathbb{R}_+$ is symmetric and satisfies the triangle inequality) be defined over $\mathscr{T}$.

Definition 2.4.22 (ε-net). *Let $\mathscr{T}$ be a subset of a pseudometric space (M, ρ) and let $\varepsilon > 0$. The collection of points $N \subseteq M$ is called an* ε-net of *$\mathscr{T}$ if*

ε-NET

$$\mathscr{T} \subseteq \bigcup_{t \in N} B_\rho(t, \varepsilon),$$

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

COVERING NUMBER

$$\mathcal{N}(\mathscr{T}, \rho, \varepsilon) = \inf\{|N| \colon N \textit{ is an } \varepsilon\textit{-net of } \mathscr{T}\}.$$

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

Definition 2.4.23 (ε-packing). *Let $\mathscr{T}$ be a subset of a pseudometric space (M, ρ) and let $\varepsilon > 0$. The collection of points $N \subseteq \mathscr{T}$ is called an* ε-packing of *$\mathscr{T}$ if*

$$t \notin B_\rho(t', \varepsilon) \quad \forall t \ne t' \in N,$$

that is, every pair of elements of N is at distance strictly greater than ε. The largest cardinality of an ε-packing of $\mathscr{T}$ is called the packing number

PACKING NUMBER

$$\mathcal{P}(\mathscr{T}, \rho, \varepsilon) = \sup\{|N| \colon N \textit{ is an } \varepsilon\textit{-packing of } \mathscr{T}\}.$$

Lemma 2.4.24 (Covering and packing numbers). *For any $\mathscr{T} \subseteq M$ and all $\varepsilon > 0$,*

$$\mathcal{N}(\mathscr{T}, \rho, \varepsilon) \le \mathcal{P}(\mathscr{T}, \rho, \varepsilon).$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Proof Observe that a maximal ε-packing N is an ε-net. Indeed, by maximality, any element of $\mathscr{T} \setminus N$ is at distance at most ε from an element of N. ■

Example 2.4.25 (Sphere in $\mathbb{R}^k$). We let $\mathbb{B}^k(x, \varepsilon)$ be the ball of radius ε around $x \in \mathbb{R}^k$ with the Euclidean metric. We let $S := \mathbb{S}^{k-1}$ be the sphere of radius 1 centered around the origin $\mathbf{0}$, that is, the surface of $\mathbb{B}^k(\mathbf{0}, 1)$. Let $0 < \varepsilon < 1$.

Claim 2.4.26

$$\mathcal{N}(S, \rho, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^k.$$

Proof Let N be any maximal ε-packing of S. We show that $|N| \leq (3/\varepsilon)^k$, which implies the claim by Lemma 2.4.24. The balls of radius $\varepsilon/2$ around points in N, $\{\mathbb{B}^k(x_i, \varepsilon/2) \colon x_i \in N\}$, 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 $\varepsilon/2$ is $\frac{\pi^{k/2}(\varepsilon/2)^k}{\Gamma(k/2+1)}$ and that of a ball of radius $3/2$ is $\frac{\pi^{k/2}(3/2)^k}{\Gamma(k/2+1)}$. Dividing one by the other proves the claim. ■

This bound will be useful later. ◀

LIPSCHITZ PROCESS

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

$$|X_t - X_s| \leq K\rho(s, t), \quad \forall s, t \in \mathscr{T}.$$

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

$$\sup_{t\in\mathscr{T}} X_t \leq \sup_{t\in\mathscr{T}} (X_t - X_{\pi(t)}) + \sup_{t\in\mathscr{T}} X_{\pi(t)} \leq K\varepsilon + \sup_{s\in N} X_s, \tag{2.4.32}$$

where we can use Theorem 2.4.21 to bound the last term. We give an example of this type of argument next (although we do not apply the above bound directly). Another example (where (2.4.32) is used this time) can be found in Section 2.4.5.

SPECTRAL NORM

Example 2.4.27 (Spectral norm of a random matrix). For an $m \times n$ matrix $A \in \mathbb{R}^{m\times n}$, the *spectral norm* (or *induced* 2-*norm*, or 2-*norm* for short) is defined as

$$\|A\|_2 := \sup_{\mathbf{x}\in\mathbb{R}^n\setminus\{0\}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x}\in\mathbb{S}^{n-1}} \|A\mathbf{x}\|_2 = \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{m-1}}} \langle A\mathbf{x}, \mathbf{y}\rangle, \tag{2.4.33}$$

where $\mathbb{S}^{n-1}$ is the sphere of Euclidean radius 1 around the origin in $\mathbb{R}^n$. The rightmost expression, which is central to our developments, is justified in Exercise 5.4.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

We will be interested in the case where A is a random matrix with independent entries. One key observation is that the quantity $\langle A\mathbf{x}, \mathbf{y}\rangle$ can then be seen as a linear combination of independent random variables

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

Hence we will be able to apply our previous tail bounds. *However,* we also need to deal with the supremum.

Theorem 2.4.28 (Upper tail of the spectral norm). *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix whose entries are centered, independent, and sub-Gaussian with variance factor ν. Then there exists a constant $0 < C < +\infty$ such that, for all $t > 0$,*

$$\|A\|_2 \leq C\sqrt{\nu}(\sqrt{m} + \sqrt{n} + t),$$

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

Without the independence assumption, the norm can be much larger in general (see Exercise 2.15).

Proof Fix $\varepsilon = 1/4$. By Claim 2.4.26, there is an ε-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 2.4.9) to control the deviations of the supremum in (2.4.33) *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 2.4.29 (Spectral norm: ε-net). *Let N and M be as above. For C large enough, for all $t > 0$,*

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

Lemma 2.4.30 (Spectral norm: Lipschitz constant). *For any ε-nets N and M of $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$, respectively, the following inequalities hold:*

$$\sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \leq \|A\|_2 \leq \frac{1}{1-2\varepsilon} \sup_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle.$$

Proof of Lemma 2.4.29 Recall that

$$\langle A\mathbf{x}, \mathbf{y}\rangle = \sum_{i,j} x_i y_j A_{ij}$$

is a linear combination of independent random variables. By the general Hoeffding inequality, $\langle A\mathbf{x}, \mathbf{y}\rangle$ is sub-Gaussian with variance factor

$$\sum_{i,j} (x_i y_j)^2\, \nu = \|\mathbf{x}\|_2^2\, \|\mathbf{y}\|_2^2\, \nu = \nu$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

for all $\mathbf{x} \in N$ and $\mathbf{y} \in M$. In particular, for all $\beta > 0$,

$$\mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\nu}\right).$$

Hence, by a union bound over N and M,

$$\begin{aligned}
&\mathbb{P}\left[\max_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2}C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right]\\
&\qquad \leq \sum_{\substack{\mathbf{x}\in N\\ \mathbf{y}\in M}} \mathbb{P}\left[\langle A\mathbf{x}, \mathbf{y}\rangle \geq \frac{1}{2}C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right]\\
&\qquad \leq |N||M| \exp\left(-\frac{1}{2\nu}\left\{\frac{1}{2}C\sqrt{\nu}(\sqrt{m}+\sqrt{n}+t)\right\}^2\right)\\
&\qquad \leq 12^{n+m} \exp\left(-\frac{C^2}{8}\left\{m+n+t^2\right\}\right)\\
&\qquad \leq e^{-t^2}
\end{aligned}$$

for $C^2/8 = \log 12 \geq 1$, where in the third inequality we ignored all cross-products since they are non-negative. ■

Proof of Lemma 2.4.30 The first inequality is immediate by definition of the spectral norm. For the second inequality, we will use the following observation:

$$\langle A\mathbf{x}, \mathbf{y}\rangle - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle = \langle A\mathbf{x}, \mathbf{y}-\mathbf{y}_0\rangle + \langle A(\mathbf{x}-\mathbf{x}_0), \mathbf{y}_0\rangle. \tag{2.4.34}$$

Fix $\mathbf{x} \in \mathbb{S}^{n-1}$ and $\mathbf{y} \in \mathbb{S}^{m-1}$ such that $\langle A\mathbf{x}, \mathbf{y}\rangle = \|A\|_2$ (which exist by compactness), and let $\mathbf{x}_0 \in N$ and $\mathbf{y}_0 \in M$ such that

$$\|\mathbf{x}-\mathbf{x}_0\|_2 \leq \varepsilon \qquad \text{and} \qquad \|\mathbf{y}-\mathbf{y}_0\|_2 \leq \varepsilon.$$

Then (2.4.34), Cauchy–Schwarz and the definition of the spectral norm imply

$$\|A\|_2 - \langle A\mathbf{x}_0, \mathbf{y}_0\rangle \leq \|A\|_2\|\mathbf{x}\|_2\|\mathbf{y}-\mathbf{y}_0\|_2 + \|A\|_2\|\mathbf{x}-\mathbf{x}_0\|_2\|\mathbf{y}_0\|_2 \leq 2\varepsilon\|A\|_2.$$

Rearranging gives the claim. ■

Putting the two lemmas together concludes the proof of Theorem 2.4.28. ■

We will give an application of this bound in Section 5.1.4. ◀

Chaining method

We go back to the inequality

$$\sup_{t\in\mathscr{T}} X_t \leq \sup_{t\in\mathscr{T}}(X_t - X_{\pi(t)}) + \sup_{t\in\mathscr{T}} X_{\pi(t)}. \tag{2.4.35}$$

Previously we controlled the first term on the right-hand side with a random Lipschitz constant and the second term with a maximal inequality for finite sets. Now 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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

stochastic process $\{X_t\}_{t\in\mathscr{T}}$ has sub-Gaussian increments on $(\mathscr{T},\rho)$ if there exists a *deterministic constant* $0 < \mathcal{K} < +\infty$ such that

SUB-GAUSSIAN INCREMENTS

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

Even with this assumption, in (2.4.35) the first term on the right-hand side remains a supremum over an infinite set. To control it, the *chaining method* repeats the argument above at progressively smaller scales, leading to the following inequality. The diameter of $\mathscr{T}$, denoted by $\text{diam}(\mathscr{T})$, is defined as

CHAINING METHOD

$$\text{diam}(\mathscr{T}) = \sup\{\rho(s,t)\colon s,t,\in\mathscr{T}\}.$$

Theorem 2.4.31 (Discrete Dudley inequality). *Let $\{X_t\}_{t\in\mathscr{T}}$ be a zero-mean stochastic process with sub-Gaussian increments on $(\mathscr{T},\rho)$ and assume* $\text{diam}(\mathscr{T}) \leq 1$. *Then*

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

for some constant $0 \leq C < +\infty$.

Proof Recall that we assume that $\mathscr{T}$ is countable. Let $\mathscr{T}_j \subseteq \mathscr{T}$, $j \geq 1$, be a sequence of finite sets such that $\mathscr{T}_j \uparrow \mathscr{T}$. By monotone convergence (Proposition B.4.14),

$$\mathbb{E}\left[\sup_{t\in\mathscr{T}} X_t\right] = \sup_{j\geq 1}\mathbb{E}\left[\sup_{t\in\mathscr{T}_j} X_t\right].$$

Moreover, $\mathcal{N}(\mathscr{T}_j,\rho,\varepsilon) \leq \mathcal{N}(\mathscr{T},\rho,\varepsilon)$ for any $\varepsilon > 0$ since $\mathscr{T}_j \subseteq \mathscr{T}$. Hence, it suffices to handle the case $|\mathscr{T}| < +\infty$.

ε-nets at all scales. For each $k \geq 0$, let N_k be an 2^{-k}-net of $\mathscr{T}$ with $|N_k| = \mathcal{N}(\mathscr{T},\rho,2^{-k})$ and projection map π_k. Because $\text{diam}(\mathscr{T}) \leq 1$, $N_0 = \{t_0\}$, where $t_0 \in \mathscr{T}$ can be taken arbitrarily. Moreover, because $\mathscr{T}$ is finite, there is $1 \leq \kappa < +\infty$ such that $N_k = \mathscr{T}$ for all $k \geq \kappa$. In particular, $\pi_\kappa(t) = t$ for all $t \in \mathscr{T}$. 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 then an expectation gives

$$\mathbb{E}\left[\sup_{t\in\mathscr{T}} X_t\right] \leq \sum_{k=0}^{\kappa-1}\mathbb{E}\left[\sup_{t\in\mathscr{T}}\left(X_{\pi_{k+1}(t)} - X_{\pi_k(t)}\right)\right], \tag{2.4.36}$$

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

Sub-Gaussian bound. We use the maximal inequality (Theorem 2.4.21) to bound the expectation in (2.4.36). For each k, the number of distinct elements in the supremum is at most

$$\begin{aligned}\left|\{(\pi_k(t),\pi_{k+1}(t))\colon t\in\mathscr{T}\}\right| &\leq |N_k \times N_{k+1}|\\ &= |N_k| \times |N_{k+1}|\\ &\leq (\mathcal{N}(\mathscr{T},\rho,2^{-k-1}))^2.\end{aligned}$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

For any $t \in \mathscr{T}$, 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

$$X_{\pi_{k+1}(t)} - X_{\pi_k(t)} \in \mathrm{s}\mathcal{G}(\mathcal{K}^2 2^{-2k+2})$$

for some $0 < \mathcal{K} < +\infty$ by the sub-Gaussian increments assumption. We can therefore apply Theorem 2.4.21 to get

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

for some constant $0 \leq C < +\infty$.

To finish the argument, we plug back into (2.4.36),

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

which implies the claim. ■

Using a similar argument, one can derive a tail inequality.

Theorem 2.4.32 (Chaining tail inequality). *Let $\{X_t\}_{t \in \mathscr{T}}$ be a zero-mean stochastic process with sub-Gaussian increments on $(\mathscr{T}, \rho)$ and assume* $\mathrm{diam}(\mathscr{T}) \leq 1$. *Then, for all $t_0 \in \mathscr{T}$ and $\beta > 0$,*

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

for some constant $0 \leq C < +\infty$.

We give an application of the discrete Dudley inequality in Section 2.4.6.

2.4.5 ▹ Data Science: Johnson–Lindenstrauss Lemma and Application to Compressed Sensing

In this section, we discuss an application of the Chernoff–Cramér method (Section 2.4.1) to dimension reduction in data science. We use once again an ε-net argument (Section 2.4.4).

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 roughly preserves the metric relationships of the points, that is, their distances. Remarkably, no structure is assumed on the original points and the result is *independent of the input dimension*. The method of proof simply involves performing a random projection.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Lemma 2.4.33 (Johnson–Lindenstrauss lemma). *For any set of points $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ in $\mathbb{R}^n$ and $\theta \in (0,1)$, there exists a mapping $f \colon \mathbb{R}^n \to \mathbb{R}^d$ with $d = \Theta(\theta^{-2} \log m)$ such that the following hold: for all i,j,*

$$(1-\theta)\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2 \le \|f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^{(j)})\|_2 \le (1+\theta)\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2. \tag{2.4.37}$$

We use the probabilistic method: we derive a "distributional" version of the result that, in turn, implies Lemma 2.4.33 by showing that a mapping with the desired properties exists with positive probability. Before stating this claim formally, we define the explicit random linear mapping we will employ. Let A be a $d \times n$ matrix whose entries are independent $\mathrm{N}(0,1)$. Note that, for any fixed $\boldsymbol{z} \in \mathbb{R}^n$,

$$\mathbb{E}\,\|A\boldsymbol{z}\|_2^2 = \mathbb{E}\left[\sum_{i=1}^{d}\left(\sum_{j=1}^{n} A_{ij} z_j\right)^2\right] = d\,\mathrm{Var}\left[\sum_{j=1}^{n} A_{1j} z_j\right] = d\|\boldsymbol{z}\|_2^2, \tag{2.4.38}$$

where we used the independence of the A_{ij}s (and, in particular, of the rows of A) and the fact that

$$\mathbb{E}\left[\sum_{j=1}^{n} A_{ij} z_j\right] = 0. \tag{2.4.39}$$

Hence, the normalized mapping

$$L = \frac{1}{\sqrt{d}} A$$

preserves the squared Euclidean norm "on average," that is, $\mathbb{E}\,\|L\boldsymbol{z}\|_2^2 = \|\boldsymbol{z}\|_2^2$. We use the Chernoff–Cramér method to prove a high-probability result.

Lemma 2.4.34 *Fix $\delta, \theta \in (0,1)$. Then the random linear mapping L above with $d = \Theta(\theta^{-2} \log \delta^{-1})$ is such that for any $\boldsymbol{z} \in \mathbb{R}^n$ with $\|\boldsymbol{z}\|_2 = 1$,*

$$\mathbb{P}\left[|\|L\boldsymbol{z}\|_2 - 1| \ge \theta\right] \le \delta. \tag{2.4.40}$$

Before proving Lemma 2.4.34, we argue that it implies the Johnson–Lindenstrauss lemma (Lemma 2.4.33). Simply take $\delta = 1/(2\binom{m}{2})$, apply the previous lemma to each normalized pairwise difference $\boldsymbol{z} = (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})/\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2$, and use a union bound over all $\binom{m}{2}$ such pairs. The probability that any of the inequalities (2.4.37) is not satisfied by the linear mapping $f(\boldsymbol{z}) = L\boldsymbol{z}$ is then at most $1/2$. Hence, a mapping with the desired properties exists for $d = \Theta(\theta^{-2} \log m)$.

Proof of Lemma 2.4.34 We prove one direction. Specifically, we establish

$$\mathbb{P}\left[\|L\,\boldsymbol{z}\|_2 \ge 1 + \theta\right] \le \exp\left(-\frac{3}{4} d\theta^2\right). \tag{2.4.41}$$

Note that the right-hand side is $\le \delta$ for $d = \Theta(\theta^{-2} \log \delta^{-1})$. An inequality in the other direction can be proved similarly by working with $-W$ below.

Recall that a sum of independent Gaussians is Gaussian (just compute the convolution and complete the squares). So

$$(A\,\boldsymbol{z})_k \sim \mathrm{N}(0, \|\boldsymbol{z}\|_2^2) = \mathrm{N}(0,1) \quad \forall k,$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

where we argued as in (2.4.38) to compute the variance. Hence,

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

is a sum of squares of independent Gaussians, that is, χ^2-distributed random variables. By (2.4.18) and independence,

$$M_W(s) = \frac{1}{(1-2s)^{d/2}}.$$

Applying the Chernoff–Cramér bound (2.4.2) with $s = \frac{1}{2}(1 - d/\beta)$ gives

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

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

$$\begin{aligned}
\mathbb{P}[\|Lz\|_2 \geq 1+\theta] &= \mathbb{P}[\|Az\|_2^2 \geq d(1+\theta)^2] \\
&= \mathbb{P}[W \geq \beta] \\
&\leq e^{d[1-(1+\theta)^2]/2}\left[(1+\theta)^2\right]^{d/2} \\
&= \exp\left(-d(\theta + \theta^2/2 - \log(1+\theta))\right) \\
&\leq \exp\left(-\frac{3}{4}d\theta^2\right),
\end{aligned}$$

where we used $\log(1+x) \leq x - x^2/4$ on $[0,1]$ (see Exercise 1.16). ■

Remark 2.4.35 *The Johnson–Lindenstrauss lemma is essentially optimal [Alo03, section 9]: any set of n points with all pairwise distances in* $[1-\theta, 1+\theta]$ *requires at least* $\Omega(\log n/(\theta^2 \log \theta^{-1}))$ *dimensions. Note, however, that it relies crucially on the use of the Euclidean norm [BC03].*

To give some further geometric insights into the proof, we make a series of observations:

1. The d 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 mean 0 and variance 1. When n is large, those two quantities are concentrated around their mean. Fix a unit vector z. Then $\frac{1}{\sqrt{n}}Az$ corresponds approximately to an orthogonal projection of z onto a uniformly chosen random subspace of dimension d.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

2. Now observe that projecting z on a uniform random subspace of dimension d can be done in the following way: first apply a uniformly chosen random rotation to z; and then project 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, note that a vector in $\mathbb{R}^n$ whose components are independent $\mathrm{N}(0,1)$, when divided by its norm, produces a uniform vector in $\mathbb{R}^n$. When d is large, the norm of the first d components of that vector is therefore a ratio whose numerator is concentrated around $\sqrt{d}$ and whose denominator is concentrated around $\sqrt{n}$ (by calculations similar to those in the first point).
3. Hence, $\|Lz\|_2 = \sqrt{\frac{n}{d}} \times \frac{1}{\sqrt{n}}\|Az\|_2$ should be concentrated around 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.

Compressed sensing

In the *compressed sensing* problem, one seeks to recover a signal $\boldsymbol{x} \in \mathbb{R}^n$ from a small number of linear measurements $(L\boldsymbol{x})_i$, $i = 1, \ldots, d$. In complete generality, one needs n such measurements to recover *any* unknown $\boldsymbol{x} \in \mathbb{R}^n$ as the *sensing matrix* L must be invertible (or, more precisely, injective). However, by imposing extra structure on the signal and choosing the sensing matrix appropriately, much better results can be obtained. Compressed sensing relies on sparsity.

SENSING MATRIX

Definition 2.4.36 (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 $\mathscr{S}_k^n$ be the set of k-sparse vectors in $\mathbb{R}^n$. Note that $\mathscr{S}_k^n$ is a union of $\binom{n}{k}$ linear subspaces, one for each support of the non-zero entries.*

K-SPARSE VECTOR

To solve the compressed sensing problem over k-sparse vectors, it suffices to find a sensing matrix L satisfying that all subsets of $2k$ columns are linearly independent. Indeed, if $\boldsymbol{x}, \boldsymbol{x}' \in \mathscr{S}_k^n$, then $\boldsymbol{x}-\boldsymbol{x}'$ has at most $2k$ non-zero entries. Hence, in order to have $L(\boldsymbol{x}-\boldsymbol{x}') = 0$, it must be that $\boldsymbol{x} - \boldsymbol{x}' = 0$ 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 proof of 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, in particular, computational efficiency. We describe such a computationally efficient approach.

The following definition will play a key role. Roughly speaking, a restricted isometry preserves enough of the metric structure of $\mathscr{S}_k^n$ to be invertible on its image.

Definition 2.4.37 (Restricted isometry property). *A $d \times n$ linear mapping L satisfies the* (k,θ)-restricted isometry property (RIP) *if for all $z \in \mathscr{S}_k^n$,*

RESTRICTED ISOMETRY PROPERTY

$$(1-\theta)\|z\|_2 \leq \|Lz\|_2 \leq (1+\theta)\|z\|_2. \tag{2.4.42}$$

We say that L is (k,θ)-RIP.

Given a (k,θ)-RIP matrix L, can we recover $z \in \mathscr{S}_k^n$ from Lz? And how small can d be? The next two claims answer these questions.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

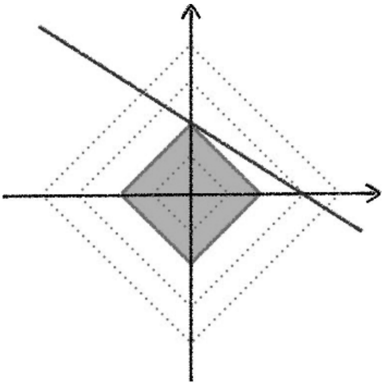

Figure 2.8 Because ℓ^1 balls (square) have corners, minimizing the ℓ^1 norm over a linear subspace (line) tends to produce sparse solutions.

Lemma 2.4.38 (Sensing matrix). *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 $(10k, 1/3)$-RIP with probability at least $1 - 1/n$.*

Lemma 2.4.39 (Sparse signal recovery). *Let L be $(10k, 1/3)$-RIP. Then for any $\boldsymbol{x} \in \mathscr{S}_k^n$, the unique solution to the following minimization problem*

$$\min_{\boldsymbol{z} \in \mathbb{R}^n} \|\boldsymbol{z}\|_1 \quad \textit{subject to} \quad L\boldsymbol{z} = L\boldsymbol{x} \tag{2.4.43}$$

is $\boldsymbol{z}^* = \boldsymbol{x}$.

It may seem that a more natural alternative approach to (2.4.43) is to instead minimize the *number of non-zero entries* in $\boldsymbol{z}$, that is, $\|\boldsymbol{z}\|_0$. However, the advantage of the ℓ^1 norm is that the problem can then be formulated as a linear program, that is, the minimization of a linear objective subject to linear inequalities (see Exercise 2.13). This permits much faster computation of the solution using standard techniques – while still leading to a sparse solution. See Figure 2.8 for some insights as to why ℓ^1 indeed promotes sparsity.

Putting the two lemmas together shows we obtain the next claim:

Claim 2.4.40 *Let L be as in Lemma 2.4.38 with $d = \Theta(k \log n)$. With probability $1 - o(1)$, any $\boldsymbol{x} \in \mathscr{S}_k^n$ can be recovered from the input $L\boldsymbol{x}$ by solving (2.4.43).*

Note that d can in general be much smaller than n and not far from the $2k$ bound we derived above.

ε-net argument We start with the proof of Lemma 2.4.38. The claim does *not* follow immediately from the (distributional) Johnson–Lindenstrauss lemma (i.e., Lemma 2.4.34). Indeed, that lemma implies that a (normalized) matrix with i.i.d. standard 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 $\mathscr{S}_k^n$, an uncountable space.

For a subset of indices $J \subseteq [n]$ and a vector $\boldsymbol{y} \in \mathbb{R}^n$, we let $\boldsymbol{y}_J$ be the vector $\boldsymbol{y}$ restricted to the entries in J, that is, the subvector $(y_j)_{j \in J}$. Fix a subset of indices $I \subseteq [n]$ of size $10k$.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

We need the RIP condition (Definition 2.4.37) to hold for all $z \in \mathbb{R}^n$ with non-zero entries in I (and all such I). The way to achieve this is to use an ε-net argument, as described in Section 2.4.4. Indeed, notice that, for $z \neq \mathbf{0}$, 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 $\partial B_I := \{z\colon z_{[n]\setminus I} = \mathbf{0},\ \|z\|_2 = 1\}$, and
2. is continuous on ∂B_I, so that it suffices to construct a fine enough covering of ∂B_I by a finite collection of balls (i.e., an ε-net) and apply Lemma 2.4.34 to the centers of those balls.

Proof of Lemma 2.4.38 Let $I \subseteq [n]$ be a subset of indices of size $k' := 10k$. There are $\binom{n}{k'} \leq n^{k'} = \exp(k' \log n)$ such subsets and we denote their collection by $\mathcal{I}(k', n)$. We let N_I be an ε-net of ∂B_I. By Claim 2.4.26, we can choose one of size at most $(3/\varepsilon)^{k'}$. We take

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

for a constant C' that will be determined below. The reason for this choice will become clear when we set C'. The union of all ε-nets has size

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

for some $C'' > 0$. Our goal is to show that

$$\sup_{z\in\cup_{I\in\mathcal{I}(k',n)}\partial B_I} \left|\|Lz\|_2 - 1\right| \leq \frac{1}{3}. \tag{2.4.44}$$

We seek to apply the inequality (2.4.32).

Applying Johnson–Lindenstrauss to the ε-nets: The first step is to control the supremum in (2.4.44) – restricted to the ε-nets. Lemma 2.4.34 is exactly what we need for this. Take $\theta = 1/6$, $\delta = 1/(2n|\cup_I N_I|)$, and

$$d = \Theta\left(\theta^{-2}\log(2n|\cup_I N_I|)\right) = \Theta(k' \log n),$$

as required by the lemma. Then, by a union bound over the N_Is, with probability $1 - 1/(2n)$, we have

$$\sup_{z\in\cup_I N_I} \left|\|Lz\|_2 - 1\right| \leq \frac{1}{6}. \tag{2.4.45}$$

Lipschitz continuity: The next step is to establish Lipschitz continuity of $\left|\|Lz\|_2 - 1\right|$. For vectors $\mathbf{y}, z \in \mathbb{R}^n$, by repeated applications of the triangle inequality, we have

$$\left|\left|\|Lz\|_2 - 1\right| - \left|\|L\mathbf{y}\|_2 - 1\right|\right| \leq \left|\|Lz\|_2 - \|L\mathbf{y}\|_2\right| \leq \|L(z - \mathbf{y})\|_2.$$

To bound the rightmost expression, we let A_* be the largest entry of A in absolute value and note that

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

$$
\begin{aligned}
\|L(\boldsymbol{z}-\boldsymbol{y})\|_2^2 &= \sum_{i=1}^{d}\left(\sum_{j=1}^{n} L_{ij}(z_j - y_j)\right)^2\\
&\leq \sum_{i=1}^{d}\left(\sum_{j=1}^{n} L_{ij}^2\right)\left(\sum_{j=1}^{n}(z_j - y_j)^2\right)\\
&\leq dn\left(\frac{1}{\sqrt{d}}A_*\right)^2 \|\boldsymbol{z}-\boldsymbol{y}\|_2^2\\
&\leq nA_*^2\|\boldsymbol{z}-\boldsymbol{y}\|_2^2,
\end{aligned}
$$

where we used Cauchy–Schwarz (Theorem B.4.8) on the second line. Taking the square root, we see that the (random) Lipschitz constant of $|\|L\boldsymbol{z}\|_2 - 1|$ (with respect to the Euclidean metric) is at most $K := \sqrt{n}A_*$.

Controlling the Lipschitz constant: So it remains to control A_*. For this we use the Chernoff–Cramér bound for Gaussians (see (2.4.4)), which implies by a union bound over the entries of A that

$$
\begin{aligned}
\mathbb{P}[A_* \geq C'\sqrt{\log n}] &\leq \mathbb{P}\left[\exists i,j,\ |A_{ij}| \geq C'\sqrt{\log n}\right]\\
&\leq n^2 \exp\left(-\frac{(C'\sqrt{\log n})^2}{2}\right)\\
&\leq \frac{1}{2n}
\end{aligned}
$$

for a $C' > 0$ large enough. Hence, with probability $1 - 1/(2n)$, we have $A_* < C'\sqrt{\log n}$ and

$$
K\varepsilon \leq \frac{1}{6} \tag{2.4.46}
$$

by the choice of ε made previously.

Putting everything together: We apply (2.4.32). Combining (2.4.45) and (2.4.46), with probability $1 - 1/n$, the claim (2.4.44) holds. That concludes the proof. ■

ℓ^1 minimization Finally we prove Lemma 2.4.39 (which can be skipped).

Proof of Lemma 2.4.39 Let $\boldsymbol{z}^*$ be a solution to (2.4.43) and note that such a solution exists because $\boldsymbol{z} = \boldsymbol{x}$ satisfies the constraint. Without loss of generality assume that only the first k entries of $\boldsymbol{x}$ are non-zero, that is, $\boldsymbol{x}_{[n]\setminus[k]} = \mathbf{0}$. Moreover, order the remaining entries of $\boldsymbol{x}$ so that the residual $\boldsymbol{r} = \boldsymbol{z}^* - \boldsymbol{x}$ has its entries $\boldsymbol{r}_{[n]\setminus[k]}$ in non-increasing order in absolute value. Our goal is to show that $\|\boldsymbol{r}\|_2 = 0$.

In order to leverage the RIP condition, we break up the vector $\boldsymbol{r}$ into $9k$-long subvectors. Let

$$
I_0 = [k], \quad I_i = \{(9(i-1)+1)k+1, \ldots, (9i+1)k\}\ \forall i \geq 1,
$$

and $\bar{I}_i = \bigcup_{j>i} I_j$. We will also need $I_{01} = I_0 \cup I_1$ and $\bar{I}_{01} = \bar{I}_1$.

We first use the optimality of $\boldsymbol{z}^*$. Note that $\boldsymbol{x}_{\bar{I}_0} = 0$ implies that

$$
\|\boldsymbol{z}^*\|_1 = \|\boldsymbol{z}^*_{I_0}\|_1 + \|\boldsymbol{z}^*_{\bar{I}_0}\|_1 = \|\boldsymbol{z}^*_{I_0}\|_1 + \|\boldsymbol{r}_{\bar{I}_0}\|_1
$$

and

$$\|\boldsymbol{x}\|_1 = \|\boldsymbol{x}_{I_0}\|_1 \leq \|\boldsymbol{z}^*_{I_0}\|_1 + \|\boldsymbol{r}_{I_0}\|_1$$

by the triangle inequality. Since $\|\boldsymbol{z}^*\|_1 \leq \|\boldsymbol{x}\|_1$ by optimality (and the fact that $\boldsymbol{x}$ satisfies the constraint), we then have

$$\|\boldsymbol{r}_{\bar{I}_0}\|_1 \leq \|\boldsymbol{r}_{I_0}\|_1. \tag{2.4.47}$$

On the other hand, the RIP condition gives a similar inequality in the other direction. Indeed, notice that $L\boldsymbol{r} = \boldsymbol{0}$ by the constraint in (2.4.43) or, put differently, $L\boldsymbol{r}_{I_{01}} = -\sum_{i\geq 2} L\boldsymbol{r}_{I_i}$. Then, by the RIP condition and the triangle inequality, we have

$$\frac{2}{3}\|\boldsymbol{r}_{I_{01}}\|_2 \leq \|L\boldsymbol{r}_{I_{01}}\|_2 \leq \sum_{i\geq 2}\|L\boldsymbol{r}_{I_i}\|_2 \leq \frac{4}{3}\sum_{i\geq 2}\|\boldsymbol{r}_{I_i}\|_2, \tag{2.4.48}$$

where we used the fact that by construction $\boldsymbol{r}_{I_{01}}$ is $10k$-sparse and each $\boldsymbol{r}_{I_i}$ is $9k$-sparse.

We note that by the ordering of the entries of $\boldsymbol{x}$,

$$\|\boldsymbol{r}_{I_{i+1}}\|_2^2 \leq 9k\left(\frac{\|\boldsymbol{r}_{I_i}\|_1}{9k}\right)^2 = \frac{\|\boldsymbol{r}_{I_i}\|_1^2}{9k}, \tag{2.4.49}$$

where we bounded $\boldsymbol{r}_{I_{i+1}}$ entrywise by the expression in parenthesis. Combining (2.4.47) and (2.4.49), and using that $\|\boldsymbol{r}_{I_0}\|_1 \leq \sqrt{k}\|\boldsymbol{r}_{I_0}\|_2$ by Cauchy–Schwarz, we have

$$\sum_{i\geq 2}\|\boldsymbol{r}_{I_i}\|_2 \leq \sum_{j\geq 1}\frac{\|\boldsymbol{r}_{I_j}\|_1}{\sqrt{9k}} = \frac{\|\boldsymbol{r}_{\bar{I}_0}\|_1}{3\sqrt{k}} \leq \frac{\|\boldsymbol{r}_{I_0}\|_1}{3\sqrt{k}} \leq \frac{\|\boldsymbol{r}_{I_0}\|_2}{3} \leq \frac{\|\boldsymbol{r}_{I_{01}}\|_2}{3}.$$

Plugging this back into (2.4.48) gives

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

which implies $\boldsymbol{r}_{I_{01}} = \boldsymbol{0}$. In particular, $\boldsymbol{r}_{I_0} = \boldsymbol{0}$ and, by (2.4.47), $\boldsymbol{r}_{\bar{I}_0} = \boldsymbol{0}$ as well. We have shown that $\boldsymbol{r} = \boldsymbol{0}$. Or, in other words, $\boldsymbol{z}^* = \boldsymbol{x}$. ■

Remark 2.4.41 *Lemma 2.4.39 can be extended to noisy measurements using a modification of* (2.4.43). *This provides some robustness to noise which is important in applications. See [CRT06b].*

2.4.6 ▷ Data Science: Classification, Empirical Risk Minimization, and VC Dimension

In the *binary classification* problem, one is given samples $\mathcal{S}_n = \{(X_i, C(X_i))\}_{i=1}^n$ 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 μ, and $C\colon \mathbb{R}^d \to \{0,1\}$ 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 to learn the function (or concept) C from the samples. BINARY CLASSIFICATION

More precisely, we seek to construct a *hypothesis* $h\colon \mathbb{R}^d \to \{0,1\}$ that is a good approximation to C in the sense that it predicts the label well on a new sample (from the same distribution). Formally, we want h to have small *true risk* (or *generalization error*): HYPOTHESIS TRUE RISK

$$R(h) = \mathbb{P}[h(X) \neq C(X)],$$

where $X \sim \mu$. Because we only have access to the distribution μ through the samples, it is natural to estimate the true risk of the hypothesis h using the samples as

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(X_i) \neq C(X_i)\},$$

EMPIRICAL RISK

EMPIRICAL RISK MINIMIZER

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*

$$h^* \in \mathrm{ERM}_{\mathcal{H}}(\mathcal{S}_n) = \{h \in \mathcal{H} \colon R_n(h) \leq R_n(h'), \forall h' \in \mathcal{H}\},$$

HYPOTHESIS CLASS

where $\mathcal{H}$, the *hypothesis class*, is a given collection of Boolean functions over $\mathbb{R}^d$. We assume further that h^* can be defined as a measurable function of the samples.

LEARNER

Overfitting Why restrict the hypothesis class? It turns out 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 follow the data, leads to overfitting: the learned hypothesis will fit the sampled data, but it may not generalize well to unseen examples. A *learner* $\mathcal{A}$ is a map from samples to measurable Boolean functions over $\mathbb{R}^d$, that is, for any n and any $\mathcal{S}_n \in (\mathbb{R}^d \times \{0,1\})^n$, the learner outputs a function $\mathcal{A}(\,\cdot\,, \mathcal{S}_n) \colon \mathbb{R}^d \to \{0,1\}$. The following theorem shows that any learner has fundamental limitations if all concepts are possible.

Theorem 2.4.42 (No free lunch). *For any learner $\mathcal{A}$ and any finite $\mathcal{X} \subseteq \mathbb{R}^d$ of even size $|\mathcal{X}| =: 2m > 4$, there exist a concept $C \colon \mathcal{X} \to \{0,1\}$ and a distribution μ over $\mathcal{X}$ such that*

$$\mathbb{P}[R(\mathcal{A}(\,\cdot\,, \mathcal{S}_m)) \geq 1/8] \geq 1/8, \tag{2.4.50}$$

where $\mathcal{S}_m = \{(X_i, C(X_i))\}_{i=1}^m$ with independent $X_i \sim \mu$.

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.

Proof of Theorem 2.4.42 We let μ be uniform over $\mathcal{X}$. To prove the existence of a concept satisfying (2.4.50), 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_xs are i.i.d. uniform in $\{0,1\}$.

We first bound $\mathbb{E}[R(\mathcal{A}(\,\cdot\,, \mathcal{S}_m))]$, where the expectation runs over both random labels $\{Y_x\}_{x \in \mathcal{X}}$ and the samples $\mathcal{S}_m = \{(X_i, C(X_i))\}_{i=1}^m$. For an additional independent sample $X \sim \mu$, we will need the event that the learner, given samples $\mathcal{S}_m$, makes an incorrect prediction on X

$$B = \{\mathcal{A}(X, \mathcal{S}_m)) \neq Y_X\},$$

and the event that X is observed in the samples $\mathcal{S}_m$

$$O = \{X \in \{X_1, \ldots, X_m\}\}.$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

By the tower property (Lemma B.6.16),

$$\begin{aligned}\mathbb{E}[R(\mathcal{A}(\,\cdot\,,\mathcal{S}_m))] &= \mathbb{P}[B]\\ &= \mathbb{E}[\mathbb{P}[B \mid \mathcal{S}_m]]\\ &= \mathbb{E}\left[\mathbb{P}[B \mid O,\mathcal{S}_m]\mathbb{P}[O \mid \mathcal{S}_m] + \mathbb{P}[B \mid O^c,\mathcal{S}_m]\mathbb{P}[O^c \mid \mathcal{S}_m]\right]\\ &\geq \mathbb{E}\left[\mathbb{P}[B \mid O^c,\mathcal{S}_m]\mathbb{P}[O^c \mid \mathcal{S}_m]\right]\\ &\geq \frac{1}{2}\times\frac{1}{2},\end{aligned}$$

where we used that

- $\mathbb{P}[O^c \mid \mathcal{S}_m] \geq 1/2$ because $|\mathcal{X}| = 2m$ and μ is uniform, and
- $\mathbb{P}[B \mid O^c,\mathcal{S}_m] = 1/2$ because for any $x \notin \{X_1,\ldots,X_m\}$ the prediction $\mathcal{A}(x,\mathcal{S}_m) \in \{0,1\}$ is independent of Y_x and the latter is uniform.

Conditioning over the concept, we have proved that

$$\mathbb{E}\left[\mathbb{E}[R(\mathcal{A}(\,\cdot\,,\mathcal{S}_m)) \mid \{Y_x\}_{x\in\mathcal{X}}]\right] \geq \frac{1}{4}.$$

Hence, by the first moment principle (Theorem 2.2.1),

$$\mathbb{P}[\mathbb{E}[R(\mathcal{A}(\,\cdot\,,\mathcal{S}_m)) \mid \{Y_x\}_{x\in\mathcal{X}}] \geq 1/4] > 0,$$

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

$$\mathbb{E}[R(\mathcal{A}(\,\cdot\,,\mathcal{S}_m)) \mid \{Y_x = y_x\}_{x\in\mathcal{X}}] \geq 1/4. \tag{2.4.51}$$

Finally, to prove (2.4.50), we use a variation on Markov's inequality (Theorem 2.1.1) 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 (2.4.51), we obtain

$$\mathbb{P}\left[R(\mathcal{A}(\,\cdot\,,\mathcal{S}_m)) \geq \frac{1}{8}\,\middle|\,\{Y_x = y_x\}_{x\in\mathcal{X}}\right] \geq \frac{1}{8},$$

establishing the claim. ■

The way out is to "limit the complexity" of the hypotheses. For instance, we could restrict ourselves to half-spaces

$$\mathcal{H}_{\mathrm{H}} = \left\{h(x) = \mathbf{1}\{x^T u \geq \alpha\} : u \in \mathbb{R}^d, \alpha \in \mathbb{R}\right\},$$

or axis-aligned boxes

$$\mathcal{H}_{\mathrm{B}} = \{h(x) = \mathbf{1}\{x_i \in [\alpha_i,\beta_i],\ \forall i\} : -\infty \leq \alpha_i \leq \beta_i \leq \infty,\ \forall i\}.$$

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

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

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

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\{R_n(h)-R(h)\}\right]=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(h,X_i)-\mathbb{E}[\ell(h,X)]\right\}\right] \tag{2.4.52}$$

in terms of a measure of complexity of the class $\mathcal{H}$, where we defined the loss $\ell(h,x)=\mathbf{1}\{h(x)\neq C(x)\}$ to simplify the notation. We assume that $\mathcal{H}$ is countable. (Observe for instance that, for $\mathcal{H}_{\mathrm{H}}$ and $\mathcal{H}_{\mathrm{B}}$, nothing is lost by assuming that the parameters defining the hypotheses are rational-valued.)

Controlling deviations uniformly over $\mathcal{H}$ as in (2.4.52) allows one to provide guarantees on the empirical risk minimizer. Indeed, for any $h'\in\mathcal{H}$,

$$\begin{aligned}
R(h^*) &= R_n(h^*)+\{R(h^*)-R_n(h^*)\}\\
&\le R_n(h^*)+\sup_{h\in\mathcal{H}}\{R(h)-R_n(h)\}\\
&\le R_n(h')+\sup_{h\in\mathcal{H}}\{R(h)-R_n(h)\}\\
&= R(h')+\left\{R_n(h')-R(h')\right\}+\sup_{h\in\mathcal{H}}\{R(h)-R_n(h)\}\\
&\le R(h')+\sup_{h\in\mathcal{H}}\{R_n(h)-R(h)\}+\sup_{h\in\mathcal{H}}\{R(h)-R_n(h)\},
\end{aligned}$$

where, on the third line, we used the definition of the empirical risk minimizer. Taking an infimum over h', then an expectation over the samples, and rearranging gives

$$\begin{aligned}
&\mathbb{E}[R(h^*)]-\inf_{h'\in\mathcal{H}}R(h')\\
&\qquad\le \mathbb{E}\left[\sup_{h\in\mathcal{H}}\{R_n(h)-R(h)\}\right]+\mathbb{E}\left[\sup_{h\in\mathcal{H}}\{R(h)-R_n(h)\}\right].
\end{aligned} \tag{2.4.53}$$

This inequality allows us to relate two quantities of interest: the expected true risk of the empirical risk minimizer (i.e., $\mathbb{E}[R(h^*)]$) and the best possible true risk (i.e., $\inf_{h'\in\mathcal{H}}R(h')$). The first term on the right-hand side is (2.4.52) and the second one can be bounded in a similar fashion as we argue below. Observe that the suprema are inside the expectations 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 $\mathcal{H}$.

To bound (2.4.52), we use the methods of Section 2.4.4. As a first step, we apply the symmetrization trick, which we introduced in Section 2.4.2 to give a proof of Hoeffding's lemma (Lemma 2.4.12). Let $(\varepsilon_i)_{i=1}^n$ be i.i.d. uniform random variables in $\{-1,+1\}$ (i.e., Rademacher variables) and let $(X_i')_{i=1}^n$ be an independent copy of $(X_i)_{i=1}^n$. Then,

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{h\in\mathcal{H}}\{R_n(h)-R(h)\}\right]\\
&=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(h,X_i)-\mathbb{E}[\ell(h,X)]\right\}\right]\\
&=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}[\ell(h,X_i)-\mathbb{E}[\ell(h,X_i')\,|\,(X_j)_{j=1}^{n}]]\right\}\right]\\
&=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[\ell(h,X_i)-\ell(h,X_i')]\,\middle|\,(X_j)_{j=1}^{n}\right]\right]\\
&\le\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}[\ell(h,X_i)-\ell(h,X_i')]\right\}\right],
\end{aligned}
$$

where on the fourth line we used taking it out what is known (Lemma B.6.13) and on the fifth line we used $\sup_h \mathbb{E}Y_h \le \mathbb{E}[\sup_h Y_h]$ and the tower property. Next we note that $\ell(h,X_i)-\ell(h,X_i')$ is symmetric and independent of ε_i (which is also symmetric) to deduce that the last line above is

$$
\begin{aligned}
&=\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i[\ell(h,X_i)-\ell(h,X_i')]\right\}\right]\\
&\le\mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(h,X_i)+\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}(-\varepsilon_i)\ell(h,X_i')\right]\\
&=2\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(h,X_i)\right].
\end{aligned}
$$

The exact same argument also applies to the second term on the right-hand side of (2.4.53), so

$$
\mathbb{E}[R(h^*)]-\inf_{h'\in\mathcal{H}}R(h')\le 4\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(h,X_i)\right]. \tag{2.4.54}
$$

Changing the normalization slightly, we define the process

$$
Z_n(h)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\ell(h,X_i),\qquad h\in\mathcal{H}. \tag{2.4.55}
$$

Our task reduces to upper bounding

$$
\mathbb{E}\left[\sup_{h\in\mathcal{H}}Z_n(h)\right]. \tag{2.4.56}
$$

Note that we will not compute the best possible true risk (which in general could be “bad,” i.e., large) – only how close the empirical risk minimizer gets to it.

VC dimension We make two observations about $Z_n(h)$.

1. It is centered. Also, as a weighted sum of independent random variables in $[-1,1]$, it is sub-Gaussian with variance factor 1 by the general Hoeffding inequality (Theorem 2.4.9) and Hoeffding’s lemma (Lemma 2.4.12).

2. It depends only on the values of the hypothesis h at a finite number of points, $X_1, \dots, X_n$. Hence, while the supremum in (2.4.56) 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 hs to $(X_i)_{i=1}^n$.

A naive application of the maximal inequality in Lemma 2.4.21, together with the two observations above, gives

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} Z_n(h)\right] \leq \sqrt{2\log 2^n} = \sqrt{2n\log 2}.$$

Unfortunately, plugging this back into (2.4.54) gives an upper bound, which fails to converge to 0 as $n \to +\infty$.

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 2.4.43 (Shattering). *Let $\Lambda = \{\ell_1, \dots, \ell_n\} \subseteq \mathbb{R}^d$ be a finite set and let $\mathcal{H}$ be a class of Boolean functions on $\mathbb{R}^d$. The restriction of $\mathcal{H}$ to Λ is*

$$\mathcal{H}_\Lambda = \{(h(\ell_1), \dots, h(\ell_n)) \,:\, h \in \mathcal{H}\}.$$

SHATTERING

We say that Λ is shattered *by $\mathcal{H}$ if $|\mathcal{H}_\Lambda| = 2^{|\Lambda|}$, that is, if all Boolean functions over Λ can be obtained by restricting a function in $\mathcal{H}$ to the points in Λ.*

VC DIMENSION

Definition 2.4.44 (VC dimension). *Let $\mathcal{H}$ be a class of Boolean functions on $\mathbb{R}^d$. The* VC dimension *of $\mathcal{H}$, denoted* $\mathrm{vc}(\mathcal{H})$*, is the maximum cardinality of a set shattered by $\mathcal{H}$.*

We prove the following combinatorial lemma at the end of this section.

Lemma 2.4.45 (Sauer's lemma). *Let $\mathcal{H}$ be a class of Boolean functions on $\mathbb{R}^d$. For any finite set $\Lambda = \{\ell_1, \dots, \ell_n\} \subseteq \mathbb{R}^d$,*

$$|\mathcal{H}_\Lambda| \leq \left(\frac{en}{\mathrm{vc}(\mathcal{H})}\right)^{\mathrm{vc}(\mathcal{H})}.$$

That is, the number of distinct restrictions of $\mathcal{H}$ to any n points grows at most as $\propto n^{\mathrm{vc}(\mathcal{H})}$.

Returning to $\mathbb{E}[\sup_{h\in\mathcal{H}} Z_n(h)]$, we get the following inequality.

Lemma 2.4.46 *There exists a constant $0 < C < +\infty$ such that, for any countable class of measurable Boolean functions $\mathcal{H}$ over $\mathbb{R}^d$,*

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} Z_n(h)\right] \leq C\sqrt{\mathrm{vc}(\mathcal{H})\log n}. \tag{2.4.57}$$

Proof Recall that $Z_n(h) \in \mathrm{s}\mathcal{G}(1)$. Since the supremum over $\mathcal{H}$, when seen as restricted to $\{X_1, \dots, X_n\}$, is in fact a supremum over at most $\left(\frac{en}{\mathrm{vc}(\mathcal{H})}\right)^{\mathrm{vc}(\mathcal{H})}$ functions by Sauer's lemma (Lemma 2.4.45), we have by Lemma 2.4.21,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} Z_n(h)\right] \leq \sqrt{2\log\left[\left(\frac{en}{\mathrm{vc}(\mathcal{H})}\right)^{\mathrm{vc}(\mathcal{H})}\right]}.$$

That proves the claim. ■

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Returning to (2.4.54), the previous lemma finally implies

$$\mathbb{E}[R(h^*)] - \inf_{h' \in \mathcal{H}} R(h') \leq 4C\sqrt{\frac{\mathrm{vc}(\mathcal{H}) \log n}{n}}.$$

For hypothesis classes with finite VC dimension, the bound goes to 0 as $n \to +\infty$.

We give some examples.

Example 2.4.47 (VC dimension of half-spaces). Consider the class of half-spaces.

Claim 2.4.48

$$\mathrm{vc}(\mathcal{H}_{\mathrm{H}}) = d + 1.$$

We only prove the case $d = 1$, where $\mathcal{H}_{\mathrm{H}}$ 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 $\mathcal{H}_{\mathrm{H}}$. On the other hand, for any $\Lambda = \{\ell_1, \ell_2, \ell_3\}$ with $\ell_1 < \ell_2 < \ell_3$, any half-line containing ℓ_1 and ℓ_3 necessarily includes ℓ_2 as well. Hence, no set of size 3 is shattered by $\mathcal{H}_{\mathrm{H}}$. ◀

Example 2.4.49 (VC dimension of boxes). Consider the class of axis-aligned boxes.

Claim 2.4.50

$$\mathrm{vc}(\mathcal{H}_{\mathrm{B}}) = 2d.$$

We only prove the case $d = 2$, where $\mathcal{H}_{\mathrm{B}}$ reduces to rectangles. The four-point set $\Lambda = \{(-1, 0), (1, 0), (0, -1), (0, 1)\}$ is shattered by $\mathcal{H}_{\mathrm{B}}$. Indeed, the rectangle $[-1, 1] \times [-1, 1]$ contains Λ, with each side of the rectangle containing one of the points. Moving any side inward by $\varepsilon < 1$ removes the corresponding point from the rectangle without affecting the other ones. Hence, any subset of Λ can be obtained by this procedure.

On the other hand, let $\Lambda = \{\ell_1, \ldots, \ell_5\} \subseteq \mathbb{R}^2$ be any set of five distinct points. If the points all lie on the same axis-aligned line, then an argument similar to the half-line case in Claim 2.4.48 shows that Λ is not shattered. Otherwise consider the axis-aligned rectangle with smallest area containing Λ. For each side of the rectangle, choose one point of Λ that lies on it. These necessarily exist (otherwise the rectangle could be made even smaller) and denote them by x_{N} for the highest, x_{E} for the rightmost, x_{S} for the lowest, and x_{W} for the leftmost. Note that they may not be distinct, but in any case at least one point in Λ, say ℓ_5 without loss of generality, is not in the list. Now observe that any axis-aligned rectangle containing $x_{\mathrm{N}}, x_{\mathrm{E}}, x_{\mathrm{S}}, x_{\mathrm{W}}$ must also contain ℓ_5 since its coordinates are sandwiched between the bounds defined by those points. Hence, no set of size 5 is shattered. That proves the claim. ◀

These two examples also provide insights 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 $\mathcal{H}_{\mathrm{B}}$ to these sample points.

Application of chaining It turns out that the $\sqrt{\log n}$ factor in (2.4.57) is not optimal. We use chaining (Section 2.4.4) to improve the bound.

We claim that the process $\{Z_n(h)\}_{h \in \mathcal{H}}$ has sub-Gaussian increments under an appropriately defined pseudometric. Indeed, conditioning on $(X_i)_{i=1}^n$, by the general Hoeffding inequality

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

(Theorem 2.4.9) and Hoeffding's lemma (Lemma 2.4.12), we have that the increment (as a function of the ε_is which have variance factor 1)

$$Z_n(g) - Z_n(h) = \sum_{i=1}^{n} \varepsilon_i \frac{\ell(g, X_i) - \ell(h, X_i)}{\sqrt{n}}$$

is sub-Gaussian 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

$$\rho_n(g, h) = \left[\frac{1}{n} \sum_{i=1}^{n} [\ell(g, X_i) - \ell(h, X_i)]^2 \right]^{1/2} = \left[\frac{1}{n} \sum_{i=1}^{n} [g(X_i) - h(X_i)]^2 \right]^{1/2},$$

where we used that $\ell(h, x) = \mathbf{1}\{h(x) \neq C(x)\}$ by definition. It 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 η over $\mathbb{R}^d$, define

$$\|g - h\|_{L^2(\eta)}^2 = \int_{\mathbb{R}^d} (f(x) - g(x))^2 \mathrm{d}\eta(x).$$

EMPIRICAL MEASURE

Let μ_n be the empirical measure

$$\mu_n = \mu_{(X_i)_{i=1}^n} := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \qquad (2.4.58)$$

where δ_x is the probability measure that puts mass 1 on x. Then, we can rewrite

$$\rho_n(g, h) = \|g - h\|_{L^2(\mu_n)}.$$

Hence we have shown that, conditioned on the samples, the process $\{Z_n(h)\}_{h \in \mathcal{H}}$ 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. Though, 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 inequality (Theorem 2.4.31), we obtain the following bound.

Lemma 2.4.51 *There exists a constant* $0 < C < +\infty$ *such that, for any countable class of measurable Boolean functions* $\mathcal{H}$ *over* $\mathbb{R}^d$,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} Z_n(h) \right] \leq C\, \mathbb{E}\left[\sum_{k=0}^{+\infty} 2^{-k} \sqrt{\log \mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(\mu_n)}, 2^{-k})} \right],$$

where μ_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 inequality conditioned on $(X_i)_{i=1}^n$. Then we take an expectation over the samples. ■

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Our use of the symmetrization trick is more intuitive than it may have appeared 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 to cancel out and that, in the limit, the variance alone characterizes the overall behavior. The ε_is in some sense explicitly capture the canceling part of this phenomenon, while ρ_n captures the scale of the resulting global fluctuations in the increments.

Our final task is to bound the covering numbers $\mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(\mu_n)}, 2^{-k})$.

Theorem 2.4.52 (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* η *over* $\mathbb{R}^d$*, and any* $\varepsilon \in (0, 1)$*,*

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(\eta)}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{C\,\mathrm{vc}(\mathcal{H})}.$$

Before proving Theorem 2.4.52, we derive its implications for uniform deviations. Compare the following bound to Lemma 2.4.46.

Lemma 2.4.53 *There exists a constant* $0 < C < +\infty$ *such that, for any countable class of measurable Boolean functions* $\mathcal{H}$ *over* $\mathbb{R}^d$*,*

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} Z_n(h)\right] \leq C\sqrt{\mathrm{vc}(\mathcal{H})}.$$

Proof By Lemma 2.4.51,

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{H}} Z_n(h)\right] &\leq C\,\mathbb{E}\left[\sum_{k=0}^{+\infty} 2^{-k}\sqrt{\log \mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(\mu_n)}, 2^{-k})}\right]\\ &\leq C\,\mathbb{E}\left[\sum_{k=0}^{+\infty} 2^{-k}\sqrt{\log\left(\frac{2}{2^{-k}}\right)^{C'\,\mathrm{vc}(\mathcal{H})}}\right]\\ &= C\sqrt{\mathrm{vc}(\mathcal{H})}\,\mathbb{E}\left[\sum_{k=0}^{+\infty} 2^{-k}\sqrt{k+1}\sqrt{C'\log 2}\right]\\ &\leq C''\sqrt{\mathrm{vc}(\mathcal{H})}\end{aligned}$$

for some $0 < C'' < +\infty$. ■

It remains to prove Theorem 2.4.52.

Proof of Theorem 2.4.52 Let $\mathcal{G} = \{g_1, \ldots, g_N\} \subseteq \mathcal{H}$ be a maximal ε-packing of $\mathcal{H}$ with $N \geq \mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(\eta)}, \varepsilon)$, which exists by Lemma 2.4.24. We use the probabilistic method (Section 2.2) and Hoeffding's inequality for bounded variables (Theorem 2.4.10) 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_is. Then we use Sauer's lemma (Lemma 2.4.45) to conclude.

1. **Restriction.** By construction, the collection $\mathcal{G}$ satisfies

$$\|g_i - g_j\|_{L^2(\eta)} > \varepsilon, \quad \forall i \neq j.$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

For an integer m that we will choose as small as possible below, let $\mathbb{X} = \{X_1, \ldots, X_m\}$ be i.i.d. samples from η and let $\mu_{\mathbb{X}}$ be the corresponding empirical measure (as defined in (2.4.58)). Observe that, for any $i \neq j$,

$$\mathbb{E}\left[\|g_i - g_j\|^2_{L^2(\mu_{\mathbb{X}})}\right] = \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}[g_i(X_k) - g_j(X_k)]^2\right] = \|g_i - g_j\|^2_{L^2(\eta)}.$$

Moreover, $[g_i(X_k) - g_j(X_k)]^2 \in [0, 1]$. Hence, by Hoeffding's inequality there exists a constant $0 < C < +\infty$ and an $m \leq C\varepsilon^{-4}\log N$ such that

$$\begin{aligned}
&\mathbb{P}\left[\|g_i - g_j\|^2_{L^2(\eta)} - \|g_i - g_j\|^2_{L^2(\mu_{\mathbb{X}})} \geq \frac{3\varepsilon^2}{4}\right] \\
&= \mathbb{P}\left[m\|g_i - g_j\|^2_{L^2(\eta)} - \sum_{k=1}^{m}[g_i(X_k) - g_j(X_k)]^2 \geq m\frac{3\varepsilon^2}{4}\right] \\
&\leq \exp\left(-\frac{2(m \cdot 3\varepsilon^2/4)^2}{m}\right) \\
&= \exp\left(-\frac{9}{8}m\varepsilon^4\right) \\
&< \frac{1}{N^2}.
\end{aligned}$$

This implies that, for this choice of m,

$$\mathbb{P}\left[\|g_i - g_j\|_{L^2(\mu_{\mathbb{X}})} > \frac{\varepsilon}{2}\ \forall i \neq j\right] > 0,$$

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

$$\|g_i - g_j\|_{L^2(\mu_{\mathcal{X}})} > \frac{\varepsilon}{2}\ \forall i \neq j. \tag{2.4.59}$$

2. **VC bound.** In particular, by (2.4.59), the functions in $\mathcal{G}$ restricted to $\mathcal{X}$ are distinct. By Sauer's lemma (Lemma 2.4.45),

$$N = |\mathcal{G}_{\mathcal{X}}| \leq |\mathcal{H}_{\mathcal{X}}| \leq \left(\frac{em}{\mathrm{vc}(\mathcal{H})}\right)^{\mathrm{vc}(\mathcal{H})} \leq \left(\frac{eC\varepsilon^{-4}\log N}{\mathrm{vc}(\mathcal{H})}\right)^{\mathrm{vc}(\mathcal{H})}. \tag{2.4.60}$$

Using that $\frac{1}{2D}\log N = \log N^{1/2D} \leq N^{1/2D}$, where $D = \mathrm{vc}(\mathcal{H})$, we get

$$\left(\frac{eC\varepsilon^{-4}\log N}{\mathrm{vc}(\mathcal{H})}\right)^{\mathrm{vc}(\mathcal{H})} \leq \left(C'\varepsilon^{-4}\right)^{\mathrm{vc}(\mathcal{H})} N^{1/2}, \tag{2.4.61}$$

where $C' = 2eC$. Plugging (2.4.61) back into (2.4.60) and rearranging gives

$$N \leq \left(C'\varepsilon^{-4}\right)^{2\,\mathrm{vc}(\mathcal{H})}.$$

That concludes the proof. ■

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Proof of Sauer's lemma Recall from Appendix A (see also Exercise 1.4) that for integers $0 < d \leq n$,

$$\sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^{d}. \tag{2.4.62}$$

Sauer's lemma (Lemma 2.4.45) follows from the following claim.

Lemma 2.4.54 (Pajor). *Let $\mathcal{H}$ be a class of Boolean functions on $\mathbb{R}^d$ and let $\Lambda = \{\ell_1, \ldots, \ell_n\} \subseteq \mathbb{R}^d$ be any finite subset. Then* PAJOR'S LEMMA

$$|\mathcal{H}_\Lambda| \leq |\{S \subseteq \Lambda\colon S \textit{ is shattered by } \mathcal{H}\}|,$$

where the right-hand side includes the empty set.

Going back to Sauer's lemma, by Lemma 2.4.54 we have the upper bound

$$|\mathcal{H}_\Lambda| \leq |\{S \subseteq \Lambda\colon S \text{ is shattered by } \mathcal{H}\}|.$$

By definition of the VC-dimension (Definition 2.4.44), the subsets $S \subseteq \Lambda$ that are shattered by $\mathcal{H}$ have size at most $\mathrm{vc}(\mathcal{H})$. So the right-hand side is bounded above by the total number of subsets of size at most $d = \mathrm{vc}(\mathcal{H})$ of a set of size n. By (2.4.62), this gives

$$|\mathcal{H}_\Lambda| \leq \left(\frac{en}{\mathrm{vc}(\mathcal{H})}\right)^{\mathrm{vc}(\mathcal{H})},$$

which establishes Sauer's lemma.

So it remains to prove Lemma 2.4.54.

Proof of Lemma 2.4.54 We prove the claim by induction on the size n of Λ. The result is trivial for $n = 1$. Assume the result is true for any $\mathcal{H}$ and any subset of size $n-1$. To apply induction, for $\iota = 0, 1$ we let

$$\mathcal{H}^\iota = \{h \in \mathcal{H}\colon h(\ell_n) = \iota\},$$

and we set

$$\Lambda' = \{\ell_1, \ldots, \ell_{n-1}\}.$$

It will be convenient to introduce the following notation:

$$\mathcal{S}(\Lambda; \mathcal{H}) = |\{S \subseteq \Lambda\colon S \text{ is shattered by } \mathcal{H}\}|.$$

Because $|\mathcal{H}_\Lambda| = |\mathcal{H}^0_{\Lambda'}| + |\mathcal{H}^1_{\Lambda'}|$ and the induction hypothesis implies $\mathcal{S}(\Lambda'; \mathcal{H}^\iota) \geq |\mathcal{H}^\iota_{\Lambda'}|$ for $\iota = 0, 1$, it suffices to show that

$$\mathcal{S}(\Lambda; \mathcal{H}) \geq \mathcal{S}(\Lambda'; \mathcal{H}^0) + \mathcal{S}(\Lambda'; \mathcal{H}^1). \tag{2.4.63}$$

There are two types of sets that contribute to the right-hand side.

- *One but not both.* Let $S \subseteq \Lambda'$ be a set that contributes to one of $\mathcal{S}(\Lambda'; \mathcal{H}^0)$ or $\mathcal{S}(\Lambda'; \mathcal{H}^1)$ but not both. Then, S is a subset of the larger set Λ and it is certainly shattered by the larger collection $\mathcal{H}$. Hence, it also contributes to the left-hand side of (2.4.63).

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

- *Both.* Let $S \subseteq \Lambda'$ be a set that contributes to both $\mathcal{S}(\Lambda';\mathcal{H}^0)$ and $\mathcal{S}(\Lambda';\mathcal{H}^1)$. Hence, it contributes two to the right-hand side of (2.4.63). As in the previous point, it is also included in $\mathcal{S}(\Lambda;\mathcal{H})$, *but it only contributes one to the left-hand side of* (2.4.63). It turns out that there is another set that contributes one to the left-hand side but zero to the right-hand side: the subset $S \cup \{\ell_n\}$. Indeed, by definition of $\mathcal{H}^\iota$, the subset $S \cup \{\ell_n\}$ cannot be shattered by it since all functions in it take the same value on ℓ_n. On the other hand, any Boolean function h on $S \cup \{\ell_n\}$ with $h(\ell_n) = \iota$ is realized in $\mathcal{H}^\iota$ since S itself is shattered by $\mathcal{H}^\iota$.

That concludes the proof. ■

Exercises

Exercise 2.1 (Moments of non-negative random variables). Prove (B.5.1). (Hint: Use Fubini's Theorem to compute the integral.)

Exercise 2.2 (Bonferroni inequalities). Let $A_1, \ldots, A_n$ be events and $B_n := \cup_i A_i$. Define

$$S^{(r)} := \sum_{1 \le i_1 < \cdots < i_r \le n} \mathbb{P}[A_{i_1} \cap \cdots \cap A_{i_r}]$$

and

$$X_n := \sum_{i=1}^{n} \mathbf{1}_{A_i}.$$

(i) Let $x_0 \le x_1 \le \cdots \le x_s \ge x_{s+1} \ge \cdots \ge x_m$ 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 ≥ 0 for even ℓ and ≤ 0 for odd ℓ.

(ii) Show that, for all r,

$$\sum_{1 \le i_1 < \cdots < i_r \le n} \mathbf{1}_{A_{i_1}} \mathbf{1}_{A_{i_2}} \cdots \mathbf{1}_{A_{i_r}} = \binom{X_n}{r}.$$

(iii) Use (i) and (ii) to show that when $\ell \in [n]$ is odd

$$\mathbb{P}[B_n] \le \sum_{r=1}^{\ell} (-1)^{r-1} S^{(r)}$$

and when $\ell \in [n]$ is even

$$\mathbb{P}[B_n] \ge \sum_{r=1}^{\ell} (-1)^{r-1} S^{(r)}.$$

These inequalities are called *Bonferroni inequalities*. The case $\ell = 1$ is Boole's inequality.

Exercise 2.3 (Percolation on $\mathbb{Z}^2$: a better bound). 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 $p > 2/3$.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Exercise 2.4 (Percolation on $\mathbb{Z}^d$: existence of critical threshold). Consider bond percolation on $\mathbb{L}^d$.

(i) Show that $p_c(\mathbb{L}^d) > 0$. (Hint: Count self-avoiding paths.)
(ii) Show that $p_c(\mathbb{L}^d) < 1$. (Hint: Use the result for $\mathbb{L}^2$.)

Exercise 2.5 (Sums of uncorrelated variables). Centered random variables $X_1, X_2, \ldots$ are *uncorrelated* if

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

(i) Assume further that $\mathrm{Var}[X_r] \leq C < +\infty$ for all r. Show that

$$\mathbb{P}\left[\frac{1}{n}\sum_{r \leq n} X_r \geq \beta\right] \leq \frac{C^2}{\beta^2 n}.$$

(ii) Use (i) to prove Theorem 2.1.6.

Exercise 2.6 (Pairwise independence: lack of concentration). Let $\boldsymbol{U} = (U_1, \ldots, U_\ell)$ be uniformly distributed over $\{0,1\}^\ell$. Let $n = 2^\ell - 1$. For all $\boldsymbol{v} \in \{0,1\}^\ell \backslash \mathbf{0}$, define

$$X_{\boldsymbol{v}} = (\boldsymbol{U} \cdot \boldsymbol{v}) \mod 2.$$

(i) Show that the random variables $X_{\boldsymbol{v}}$, $\boldsymbol{v} \in \{0,1\}^\ell \backslash \mathbf{0}$, are uniformly distributed in $\{0,1\}$ and pairwise independent.
(ii) Show that for any event A measurable with respect to $\sigma(X_{\boldsymbol{v}}, \boldsymbol{v} \in \{0,1\}^\ell \backslash \mathbf{0})$, $\mathbb{P}[A]$ is either 0 or $\geq 1/(n+1)$.

Exercise 2.5 shows that pairwise independence implies "polynomial concentration" of the average of square-integrable $X_{\boldsymbol{v}}$s. On the other hand, the current exercise suggests that in general pairwise independence cannot imply "exponential concentration."

Exercise 2.7 (Chernoff bound for Poisson trials). Using the Chernoff–Cramér method, prove part (i) of Theorem 2.4.7. Show that part (ii) follows from part (i).

Exercise 2.8 (Stochastic knapsack: some details). Consider the stochastic fractional knapsack problem in Section 2.4.3.

(i) Prove that the greedy algorithm described there gives an optimal solution to problem (2.4.21).
(ii) Prove Claim 2.4.20 for $\tau \in (0, 1/6)$.

Exercise 2.9 (Stochastic knapsack: 0-1 version). Consider the stochastic fractional knapsack problem in Section 2.4.3.

(i) Adapt the greedy algorithm for the 0-1 knapsack problem and show that it is not optimal in general. (Hint: Construct a counter-example with two items.)
(ii) Prove Claim 2.4.20 for the greedy solution of (i).

Exercise 2.10 (A proof of Pólya's theorem). Let (S_t) be simple random walk on $\mathbb{L}^d$ started at the origin 0.

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

(i) For $d = 1$, use Stirling's formula (see Appendix A) to show that $\mathbb{P}[S_{2n} = 0] = \Theta(n^{-1/2})$.

(ii) For $j = 1, \ldots, d$, let $N_t^{(j)}$ be the number of steps in the jth coordinate by time t. Show that

$$\mathbb{P}\left[N_n^{(j)} \in \left[\frac{n}{2d}, \frac{3n}{2d}\right], \forall j\right] \geq 1 - \exp(-\kappa_d n)$$

for some constant $\kappa_d > 0$.

(iii) Use (i) and (ii) to show that, for any $d \geq 3$, $\mathbb{P}[S_{2n} = 0] = O(n^{-d/2})$.

Exercise 2.11 (Maximum degree). Let $G_n = (V_n, E_n) \sim \mathbb{G}_{n,p_n}$ 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 $\varepsilon > 0$,

$$\mathbb{P}\left[D_n \geq (n-1)p_n + \max\{C, 4(1+\varepsilon)\} \log n\right] \to 0,$$

as $n \to +\infty$.

Exercise 2.12 (RIP versus orthogonality). Show that a $(k, 0)$-RIP matrix with $k \geq 2$ is orthogonal, that is, its columns are orthonormal.

Exercise 2.13 (Compressed sensing: linear programming formulation). Formulate (2.4.43) as a linear program, that is, the minimization of a linear objective subject to linear inequalities.

Exercise 2.14 (Compressed sensing: almost sparse case). By adapting the proof of Lemma 2.4.39, show the following "almost sparse" version. Let L be $(10k, 1/3)$-RIP. Then, for any $\boldsymbol{x} \in \mathbb{R}^n$, the solution to (2.4.43) satisfies $\|\boldsymbol{z}^* - \boldsymbol{x}\|_2 = O(\eta(\boldsymbol{x})/\sqrt{k})$, where $\eta(\boldsymbol{x}) := \min_{\boldsymbol{x}' \in \mathscr{S}_k^n} \|\boldsymbol{x} - \boldsymbol{x}'\|_1$.

Exercise 2.15 (Spectral norm without independence). Give an example of a random matrix $A \in \mathbb{R}^{n \times n}$ whose entries are bounded, but not independent, such that the spectral norm is $\Omega(n)$ with high probability.

Exercise 2.16 (Spectral norm: symmetric matrix). Let $A \in \mathbb{R}^{n \times n}$ be a symmetric random matrix. We assume that entries on and above the diagonal $A_{i,j}$, $i \leq j$, are centered, independent, and sub-Gaussian with variance factor ν. Each entry below the diagonal is equal to the corresponding entry above it. Prove an analogue of Theorem 2.4.28 for A. (Hint: Mimic the proof of Theorem 2.4.28.)

Exercise 2.17 (Chaining tail inequality). Prove Theorem 2.4.32.

Exercise 2.18 (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

$$X_n := \sum_{i=1}^{n} A_i.$$

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press

Assume that there is $\mu > 0$ such that, for all r,

$$S^{(r)} \to \frac{\mu^r}{r!}.$$

Use Exercise 2.2 and a Taylor expansion of $e^{-\mu}$ to show that

$$\mathbb{P}[X_n = 0] \to e^{-\mu}.$$

In fact, $X_n \xrightarrow{\mathrm{d}} \mathrm{Poi}(\mu)$ (no need to prove this). This is a special case of the *method of moments*.

Exercise 2.19 (Connectivity: critical window). Using Exercise 2.18 show that, when $p_n = \frac{\log n + s}{n}$, the probability that an Erdős–Rényi graph $G_n \sim \mathbb{G}_{n,p_n}$ contains no isolated vertex converges to $e^{-e^{-s}}$.

Bibliographic Remarks

Section 2.1 For more on moment-generating functions, see [Bil12, section 21].

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 [Rom15], from which the material in Section 2.2.3 is taken. The contour lemma, Lemma 2.2.14, 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, and so on, see, for example, [JLR11, chapter 3]. For corresponding results for larger subgraphs, such as cycles or matchings, see, for example, [Bol01]. The connectivity threshold in Section 2.3.2 is also due to the same authors [ER59]. The presentation here follows [vdH17, section 5.2]. For more on the method of moments, see, for example, [Dur10, section 3.3.5] or [JLR11, section 6.1]. Claim 2.3.11 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, for example, [BLM13]. The basics of large deviation theory are covered in [Dur10, section 2.6]. See also [RAS15] and [DZ10]. Section 2.4.2 is based partly on [Ver18] and [Lug, section 3.2]. Section 2.4.3 is based on [FR98, section 5.3]. Very insightful, and much deeper, treatment of the material in Section 2.4.4 can be found in [Ver18, vH16]. The presentation in Section 2.4.5 is inspired by [Har, Lectures 6 and 8] and [Tao]. 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, for example, [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]. Lemma 2.4.39 is due to Candés, Romberg, and Tao [CRT06b]. The proof of Lemma 2.4.38 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 material in Section 2.4.2 can be found in [BLM13, chapter 2]. Hoeffding's lemma and inequality are due to Hoeffding [Hoe63]. Section 2.4.6 borrows from [Ver18, vH16, SSBD14, Haz16]. The proof of Sauer's lemma follows [Ver18, section 8.3.3]. For a proof of Claim 2.4.48 in general dimension d, see, for example, [SSBD14, section 9.1.3].

https://doi.org/10.1017/9781009305129.003 Published online by Cambridge University Press