Chapter 2

Moments and tails

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

As a quick reminder, let X be a random variable with $\mathbb{E}|X|^k < +\infty$ for $k \in \mathbb{N}$. Recall that the quantities $\mathbb{E}[X^k]$ and $\mathbb{E}[(X - \mathbb{E}X)^k]$ are called respectively the *k-th moment* and *k-th central moment* of X. The first moment and the second central moment are of course 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.

moments

The *moment-generating function* of X is the function

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

moment-generating function

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

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

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

tail probabilities

Version: October 3, 2014

2.1 First moment method

We begin with bounds based on the first moment. First recall that the expectation of a random variable has an elementary, yet very handy property: *linearity*. If random variables $X_1, \ldots, X_n$ defined on a joint probability space have finite first moments, *without any further assumption*,

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]. \tag{2.1}$$

In particular linearity holds whether or not the X_is are independent.

2.1.1 The probabilistic method

We first illustrate the usefulness of (2.1) on a key technique in probabilistic combinatorics, *the probabilistic method*. The idea behind this technique is that one can establish the existence of an object—say a graph—satisfying a certain property—say being 3-colorable—*without having to construct one explicitly*. One instead argues that a randomly chosen object exhibits the given property with positive probability. This is easier to understand on an example.

Example 2.1 (Balancing vectors)**.** Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be arbitrary unit vectors in $\mathbb{R}^n$. How small can we make the norm of the 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

$$\begin{aligned}
\mathbb{E}\|X_1\boldsymbol{v}_1 + \cdots + X_n\boldsymbol{v}_n\|^2 &= \mathbb{E}\left[\sum_{i,j} X_iX_j\boldsymbol{v}_i \cdot \boldsymbol{v}_j\right] \\
&= \sum_{i,j} \mathbb{E}[X_iX_j\boldsymbol{v}_i \cdot \boldsymbol{v}_j] \\
&= \sum_{i,j} \boldsymbol{v}_i \cdot \boldsymbol{v}_j\mathbb{E}[X_iX_j] \\
&= \sum_{i} \|\boldsymbol{v}_i\|^2 \\
&= n,
\end{aligned} \tag{2.2}$$

where, of course, we used the linearity of expectation on (2.2). But note that a discrete random variable $Z = \|X_1\boldsymbol{v}_1 + \cdots + X_n\boldsymbol{v}_n\|^2$ with expectation $\mathbb{E}Z = n$ must take a value $\leq n$ with positive probability. In other words, there must be 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 (Independent sets)**.** Let $G = (V, E)$ be a d-regular graph with n vertices and $m = nd/2$ edges, where $d \geq 1$. Our goal is derive a lower bound on the size $\alpha(G)$ of the largest independent set in G. Again, at first sight, this may seem like a rather complicated graph theory problem. But an appropriate random choice gives a non-trivial bound. Specifically, we claim that $\alpha(G) \geq n/2d$.

The proof proceeds in two steps:

1. We first prove the existence of a subset S of vertices with relatively few edges.

2. We remove vertices from S to obtain an independent set.

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

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

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

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

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

$$|S| - |\{\{i, j\} \in E \,:\, i, j \in S\}| \geq n/2d. \tag{2.3}$$

For each edge e connecting two vertices in S, remove one of the end-vertices of e. Then, by (2.3) the resulting subset of vertices has at least $n/2d$ vertices, with no edge between them. This proves the claim.

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

Remark 2.3. *The previous result can be strengthened to*

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

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

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

2.1.2 Markov's inequality

Our first bound on the tail of a random variable is *Markov's inequality*, which says: the heavier the tail, the larger the expectation. This simple inequality is in fact a key ingredient in more sophisticated tail bounds.

Theorem 2.4 (Markov's inequality)**.** *Let $X \geq 0$ be a non-negative random variable. Then, for all $b > 0$,*

$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E}X}{b}. \tag{2.4}$$

Proof.

$$\mathbb{E}X \geq \mathbb{E}[X \mathbb{1}_{X \geq b}] \geq b\, \mathbb{E}[\mathbb{1}_{X \geq b}] = b\, \mathbb{P}[X \geq b].$$

See Figure 2.1.2. ■

Note that this inequality is non-trivial only when $b > \mathbb{E}X$. The following corollary of Markov's inequality is sometimes referred to as the *first moment method*.

Corollary 2.5 (First moment method)**.** *If X is a non-negative, integer-valued random variable, then*

$$\mathbb{P}[X > 0] \leq \mathbb{E}X. \tag{2.5}$$

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

In particular, when X is a sum of indicators, the first moment method reduces to *Boole's inequality* or *union bound*.

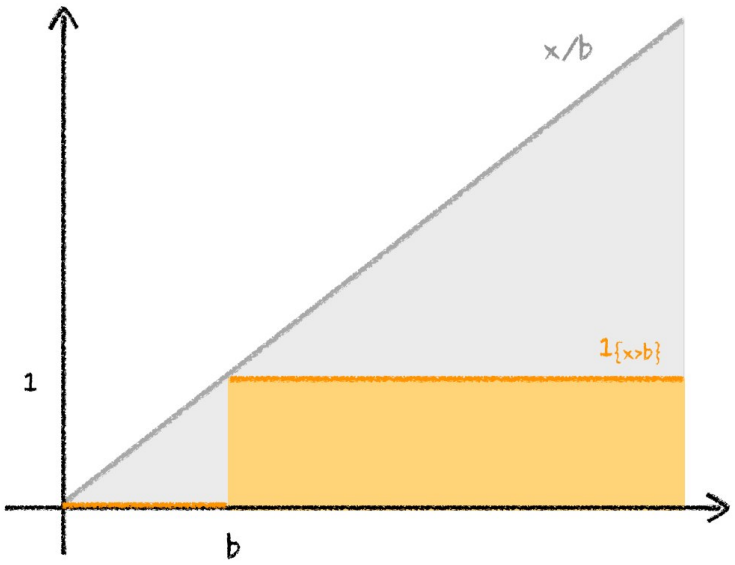

Figure 2.1: Proof of Markov's inequality: taking expectations of the two functions depicted above yields the inequality.

Corollary 2.6 (Union bound)**.** *Let $B_m = A_1 \cup \cdots \cup A_m$, where $A_1, \ldots, A_m$ is a collection of events. Then, letting*

$$\mu_m := \sum_i \mathbb{P}[A_i],$$

we have

$$\mathbb{P}[B_m] \leq \mu_m.$$

In particular, if $\mu_m \to 0$ then $\mathbb{P}[B_m] \to 0$.

Proof. Take $X = \sum_i \mathbb{1}_{A_i}$ in the first moment method. ■

An important generalization of the union bound is given in Exercise 2.1. Next we give three examples of applications of the first moment method.

2.1.3 ▷ *Random permutations: longest increasing subsequence*

In our first application of the first moment method, we bound the expected length of a longest increasing subsequence in a random permutation. Let σ_n be a uniformly random permutation of $[n] := \{1, \ldots, n\}$ and let L_n be the length of a longest increasing subsequence of σ_n.

Claim 2.7.

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

Proof. We first prove that

$$\limsup_{n\to\infty} \frac{\mathbb{E}L_n}{\sqrt{n}} \le e,$$

which implies half of the claim. Bounding the expectation of L_n is not straightforward as it is the expectation of a *maximum*. A natural way to proceed is to find a value ℓ for which $\mathbb{P}[L_n \ge \ell]$ is "small." More formally, we bound the expectation as follows

$$\mathbb{E}L_n \le \ell\,\mathbb{P}[L_n < \ell] + n\,\mathbb{P}[L_n \ge \ell] \le \ell + n\,\mathbb{P}[L_n \ge \ell], \tag{2.6}$$

for an ℓ chosen below. To bound the probability on the r.h.s., we appeal to the first moment method by letting X_n be the number of increasing subsequences of length ℓ. We also use the indicator trick, i.e., we think of X_n as a sum of indicators over subsequences (not necessarily increasing) of length 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,

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

Note that, in order for this bound to go to 0, we need $\ell > e\sqrt{n}$. The first claim follows by taking $\ell = (1+\delta)e\sqrt{n}$ in (2.6) for any $\delta > 0$.

For the other half of the claim, we show that

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

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

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

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

Remark 2.8. *It has been shown that in fact*

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

where $c = -1.77...$ *[BDJ99].*

2.1.4 ▷ *Constraint satisfaction: bound on random k-SAT threshold*

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

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

Studying such *threshold phenomena* or *phase transitions* is a major theme of modern discrete probability. Using the first moment method, we give an upper bound on this conjectured threshold.

Claim 2.9.

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

Proof. How to start 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

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

There are 2^n possible assignments, each of which satisfies $\Phi_{n,r}$ with probability $(1 - 2^{-k})^{rn}$. Indeed note that the rn clauses are independent and each clause literal picked is satisfied with probability $1/2$. Therefore, by the assumption on r, for some $\varepsilon > 0$

$$\begin{aligned} \mathbb{E}X_n &= 2^n(1-2^{-k})^{rn} \\ &< 2^n(1-2^{-k})^{(2^k \ln 2)(1+\varepsilon)n} \\ &< 2^n e^{-(\ln 2)(1+\varepsilon)n} \\ &= 2^{-\varepsilon n} \\ &\to 0. \end{aligned}$$

■

Remark 2.10. *For $k \geq 3$, it has been shown that if $r < 2^k \ln 2 - k$*

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

See [ANP05]. For the $k = 2$ case, the conjecture above has been established and $r_2^* = 1$ [CR92].

2.1.5 ▷ *Percolation on $\mathbb{Z}^d$: existence of a phase transition*

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

Percolation on $\mathbb{Z}^2$ Consider bond percolation on the two-dimensional lattice $\mathbb{L}^2$ with density p and let $\mathbb{P}_p$ denote the corresponding measure. Writing $x \Leftrightarrow y$ if $x, y \in \mathbb{L}^2$ are connected by an open path, recall that the open cluster of x is

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

The *percolation function* is defined as

percolation function

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

i.e., $\theta(p)$ is the probability that the origin is connected by open paths to infinitely many vertices. It is intuitively clear that the function $\theta(p)$ is non-decreasing. Indeed

consider the following alternative representation of the percolation process: to each edge e, assign a uniform $[0,1]$ random variable U_e and declare the edge open if $U_e \leq p$. Using the same U_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 a lot 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

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

the point at which the probability that the origin is contained in an infinite open cluster becomes positive. Note that by a union bound over all vertices, when $\theta(p) = 0$, we have that $\mathbb{P}_p[\exists x,\ |\mathcal{C}_x| = +\infty] = 0$. Conversely, because $\{\exists x,\ |\mathcal{C}_x| = +\infty\}$ is a tail event, by Kolmogorov's 0-1 law (e.g. [Dur10, Theorem 2.5.1]) it holds that $\mathbb{P}_p[\exists x,\ |\mathcal{C}_x| = +\infty] = 1$ when $\theta(p) > 0$.

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

Claim 2.11.

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

Proof. We first show that, for any $p < 1/3$, $\theta(p) = 0$. We observe that an infinite $\mathcal{C}_0$ must contain an open self-avoiding path starting at 0 of *infinite* length and, as a result, of *all* lengths. To apply the first moment method, we let X_n be the number of open self-avoiding paths of length n starting at 0. Then, by monotonicity,

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

We note that

$$\mathbb{E}X_n \leq 4(3^{n-1})p^n, \tag{2.7}$$

where we bounded the number of self-avoiding paths by noting that they cannot backtrack, giving 4 choices at the first step, and 3 choices at each subsequent step. The bound (2.7) goes to 0 when $p < 1/3$, which proves the first part of the claim.

For the other direction, we show that $\theta(p) > 0$ for p close enough to 1. This proof uses a standard *contour argument*, also known as a *Peierls' argument*, which is based on the following construction. Consider the *dual lattice* $\widehat{\mathbb{L}}^2$ whose vertices are $\mathbb{Z}^2 + (1/2, 1/2)$ and whose edges connect vertices u, v with $\|u - v\|_1 = 1$. See Figure 2.2. 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.

dual lattice

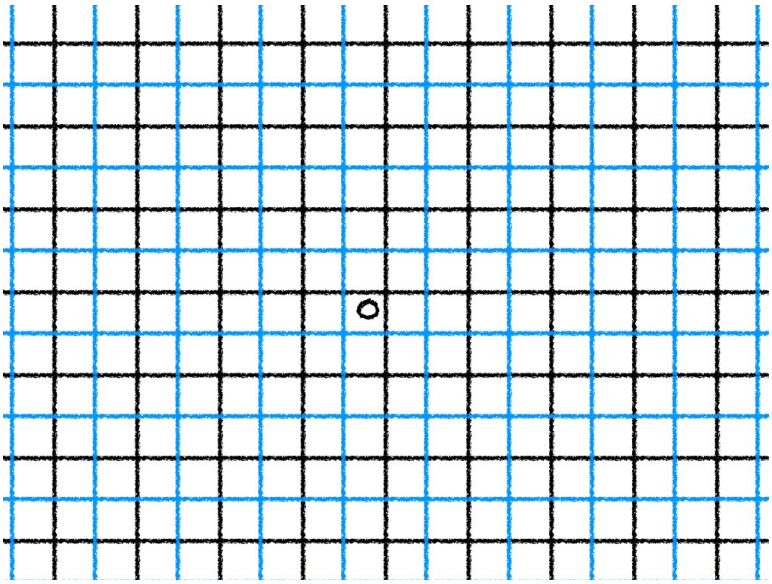

Figure 2.2: Primal (black) and dual (blue) lattices.

Lemma 2.12 (Contour lemma)**.** *If $|\mathcal{C}_0| < +\infty$, then there is a closed self-avoiding cycle around the origin in the dual lattice $\widehat{\mathbb{L}}^2$.*

To prove that $\theta(p) > 0$ for p close enough to 1, the idea is to use the first moment method with X_n equal to the number of closed self-avoiding dual cycles of length n surrounding the origin. Namely,

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

when $p > 2/3$ (where the first term comes from differentiating the geometric se-

ries). This expression can in fact be taken smaller than 1 if we let $p \to 1$. Above we bounded 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. We have shown that $\theta(p) > 0$ for p close enough to 1, and that concludes the proof. (Exercise 2.2 sketches a proof that $\theta(p) > 0$ for $p > 2/3$.) ■

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

Proof of the contour lemma We conclude this section by sketching the proof of the contour lemma which relies on topological arguments.

Proof of Lemma 2.12. 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 the highest vertex in $\mathcal{C}_0$ along the y-axis and let v_0 be the dual vertex corresponding to the upper left corner of the square of u_0. Because u_0 is highest, it must be that the square above it is red. Walk along the dual edge $\{v_0, v_1\}$ separating the squares of u_0 and $u_0 + (0, 1)$ from v_0 to v_1. Notice that this edge satisfies what we call the *red-green property*: a red square sits on your left and a green square is on your right. Proceed further by iteratively walking along an incident dual edge with the following rule. Choose an edge satisfying the red-green property, with the edges to your left, straight ahead, and to your right in decreasing order of priority (i.e., choose the highest priority edge that satisfies the red-green property). Stop when a previously visited dual vertex is reached. The claim is that this procedure constructs the desired cycle. Let $v_0, v_1, v_2, \ldots$ be the dual vertices visited. By construction $\{v_{i-1}, v_i\}$ is a dual edge for all i.

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

assumption the number of green squares is finite, this procedure must eventually terminate when a previously visited dual vertex is reached, forming a cycle.

- *(The origin lies within the cycle)* The inside of a cycle in the plane is well-defined by the Jordan curve theorem. So the dual cycle produced above has its adjacent green squares either on the inside (negative orientation) or on the outside (positive orientation). In the former case, the origin must lie inside the cycle as otherwise the vertices corresponding to the green squares on the inside would not be in $\mathcal{C}_0$. So it remains to consider the latter case where, for similar reasons, the origin is outside the cycle.

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

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

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

That concludes the proof. ■

2.2 Second moment method

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

So we turn to the second moment. Intuitively the basis for the so-called second moment method is that, if the expectation of X_n is large *and* its variance is rela-

tively 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.

2.2.1 Chebyshev's and Paley-Zygmund inequalities

We recall two classical tail inequalities involving the second moment of a random variable. The first one is an application of Markov's inequality to the random variable $|X - \mathbb{E}X|^2$.

Theorem 2.13 (Chebyshev's inequality)**.** *Let X be a random variable with $\mathbb{E}X^2 < +\infty$. Then, for all $\beta > 0$,*

$$\mathbb{P}[|X - \mathbb{E}X| > \beta] \leq \frac{\mathrm{Var}[X]}{\beta^2}. \tag{2.8}$$

Proof. This follows immediately by applying (2.4) 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.

Example 2.14. Let X be a Gaussian random variable with mean 0 and variance σ^2. A direct computation shows that $\mathbb{E}|X| = \sigma\sqrt{\frac{2}{\pi}}$. Hence Markov's inequality gives

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

while Chebyshev's inequality gives

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

For b large enough, Chebyshev's inequality produces a stronger bound. See Figure 2.3 for some insight. ◀

Example 2.15 (Coupon collector's problem)**.** Let (X_i) be i.i.d. uniform random variables over $[n]$. Let $T_{n,k}$ be the first time that k elements of $[n]$ have been picked, i.e.,

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

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

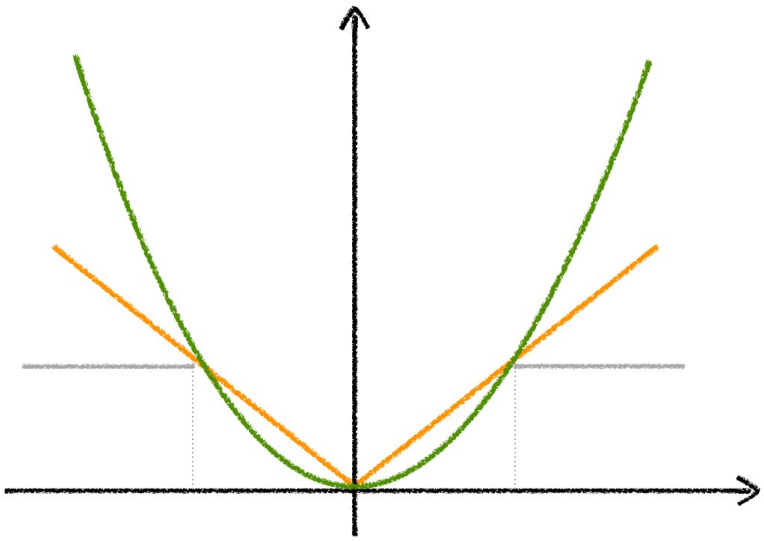

Figure 2.3: Comparison of Markov's and Chebyshev's inequalities: the squared deviation from the mean (in green) gives a better approximation of the indicator function (in grey) close to the mean (here 0) than the absolute deviation (in orange).

Claim 2.16.

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

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

$$\mathbb{E}[T_{n,n}] = \sum_{i=1}^{n}\left(1 - \frac{i-1}{n}\right)^{-1} = n\sum_{j=1}^{n} j^{-1} \sim n\log n, \tag{2.9}$$

and

$$\mathrm{Var}[T_{n,n}] \leq \sum_{i=1}^{n}\left(1 - \frac{i-1}{n}\right)^{-2} = n^2\sum_{j=1}^{n} j^{-2} \leq n^2\sum_{j=1}^{+\infty} j^{-2} = \Theta(n^2). \tag{2.10}$$

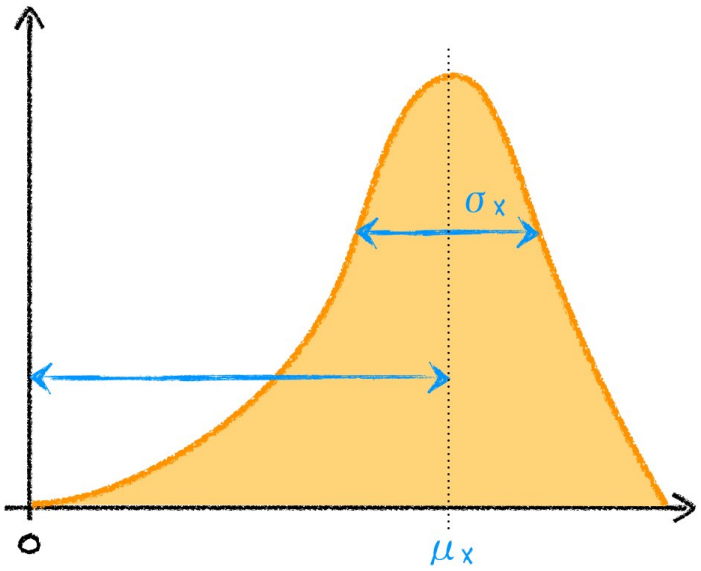

Figure 2.4: Second moment method: if the standard deviation σ_X of X is less than its expectation μ_X, then the probability that X is 0 is bounded away from 1.

So by Chebyshev's inequality

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

by (2.9) and (2.10). ◀

As an immediate corollary of Chebyshev's inequality, 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.

Theorem 2.17 (Second moment method)**.** *Let X be a non-negative, integer-valued random variable (not identically zero). Then*

$$\mathbb{P}[X > 0] \geq 1 - \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2}. \tag{2.11}$$

Proof. By (2.8),

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

■

The following tail inequality, a simple application of Cauchy-Schwarz, can be seen as an *anti-concentration* bound, i.e., it gives a lower bound on the tail of a random variable.

Theorem 2.18 (Paley-Zygmund inequality)**.** *Let X be a non-negative random variable. For all $0 < \theta < 1$,*

$$\mathbb{P}[X \geq \theta \mathbb{E}X] \geq (1 - \theta)^2 \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}. \tag{2.12}$$

Proof. We have

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

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

An immediate application of the Paley-Zygmund inequality gives a slightly improved (but perhaps less intuitive) version of the second moment method.

Corollary 2.19. *Let X be a non-negative random variable (not identically zero). Then*

$$\mathbb{P}[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} = 1 - \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2 + \mathrm{Var}[X]}. \tag{2.13}$$

Proof. Take $\theta \downarrow 0$ in (2.12). ■

We typically apply the second moment method to a sequence of random variables (X_n). The previous corollary 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 used in combination with sums of indicators (but see Section 2.2.4 for a weighted version).

Corollary 2.20 (Second moment method for sums of indicators)**.** *Let $B_m = A_1 \cup \cdots \cup A_m$, where $A_1, \ldots, A_m$ is a collection of events. Write $i \sim j$ if $i \neq j$ and A_i and A_j are* not *independent. Then, letting*

$$\mu_m := \sum_i \mathbb{P}[A_i], \qquad \gamma_m := \sum_{i \sim j} \mathbb{P}[A_i \cap A_j],$$

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

Proof. Take $X := \sum_i \mathbb{1}_{A_i}$ in the second moment method. Note that

$$\mathrm{Var}[X] = \sum_i \mathrm{Var}[\mathbb{1}_{A_i}] + \sum_{i \neq j} \mathrm{Cov}[\mathbb{1}_{A_i}, \mathbb{1}_{A_j}],$$

where

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

and, if A_i and A_j are independent,

$$\mathrm{Cov}[\mathbb{1}_{A_i}, \mathbb{1}_{A_j}] = 0,$$

whereas, if $i \sim j$,

$$\mathrm{Cov}[\mathbb{1}_{A_i}, \mathbb{1}_{A_j}] = \mathbb{E}[\mathbb{1}_{A_i}\mathbb{1}_{A_j}] - \mathbb{E}[\mathbb{1}_{A_i}]\mathbb{E}[\mathbb{1}_{A_j}] \leq \mathbb{P}[A_i \cap A_j].$$

Hence

$$\frac{\mathbb{E}[X^2]}{(\mathbb{E}X)^2} = 1 + \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2} \leq 1 + \frac{\mu_m + \gamma_m}{\mu_m^2}.$$

Applying Corollary 2.19 gives the result. ∎

We give applications of Theorem 2.17 and Corollary 2.20 in Sections 2.2.2 and 2.2.3.

2.2.2 ▷ *Erdös-Rényi graphs: small subgraph containment*

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

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

where, under $\mathbb{P}_{n,p_n}$, $G_n \sim \mathbb{G}_{n,p_n}$. is an Erdös-Rényi graph with n vertices and density p_n. In this section, we first illustrate this definition on the clique number of an Erdös-Rényi graph, then we consider the more general subgraph containment problem.

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

clique number

Claim 2.21. *The property $\omega(G) \geq 4$ has threshold function $n^{-2/3}$.*

Proof. Let X_n be the number of 4-cliques in the Erdös-Rényi 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 gives one direction.

For the other direction, we apply the second moment method for sums of indicators. For an enumeration $S_1, \ldots, S_m$ of the 4-tuples of vertices in G_n, let $A_1, \ldots, A_m$ be the events that the corresponding 4-cliques are present. By the calculation above we have $\mu_m = \Theta(n^4 p_n^6)$ which goes to $+\infty$ when $p_n \gg n^{-2/3}$. Also $\mu_m^2 = \Theta(n^8 p_n^{12})$ so it suffices to show that $\gamma_m = o(n^8 p_n^{12})$. Note that two 4-cliques with disjoint edge sets (but possibly sharing one vertex) are independent. Suppose S_i and S_j share 3 vertices. Then

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

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

$$\begin{aligned}
\gamma_m &= \sum_{i \sim j} \mathbb{P}_{n,p_n}[A_j]\, \mathbb{P}_{n,p_n}[A_i \mid A_j] \\
&= \binom{n}{4} p_n^6 \left[\binom{4}{3}(n-4) p_n^3 + \binom{4}{2}\binom{n-4}{2} p_n^5 \right] \\
&= O(n^5 p_n^9) + O(n^6 p_n^{11}) \\
&= O\left(\frac{n^8 p_n^{12}}{n^3 p_n^3}\right) + O\left(\frac{n^8 p_n^{12}}{n^2 p_n}\right) \\
&= o(n^8 p_n^{12}) \\
&= o(\mu_m^2),
\end{aligned}$$

where we used that $p_n \gg n^{-2/3}$ (so that for example $n^3 p_n^3 \gg 1$). Corollary 2.20 gives the result.

■

Roughly speaking the reason why the first and second moments suffice to pinpoint the threshold in this case is that the indicators in X_n are "mostly" pairwise independent and, as a result, the sum is concentrated around its mean. Note in particular that γ_m in Corollary 2.20 controls the extent of the dependence. We will come back to the type of "local" dependence encountered here in Chapter **??**.

General subgraphs The methods of Claim 2.21 can be applied to more general subgraphs. However the situation is somewhat more complicated than it is for cliques. For a graph H_0, let v_{H_0} and e_{H_0} be the number of vertices and edges of H_0 respectively. Let X_n be the number of copies of H_0 in $G_n \sim \mathbb{G}_{n,p_n}$. By the first moment method,

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

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.) From the proof of Claim 2.21, one might guess that the threshold function is $n^{-v_{H_0}/e_{H_0}}$. That is not the case in general. To see what can go wrong, consider the graph of Figure 2.5 whose *edge density* is $\frac{e_{H_0}}{v_{H_0}} = \frac{6}{5}$. When $p_n \gg n^{-5/6}$, the expected number of copies of H_0 tends to $+\infty$. But observe that the subgraph H of H_0 has the *higher* density $5/4$ and, hence, when $n^{-5/6} \ll p_n \ll n^{-4/5}$ the expected number of copies of H tends to 0. By the first moment method, the probability that a copy of H_0—and therefore H—is present in that regime is asymptotically negligible despite its diverging expectation. This leads to the following definition

edge density

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

Assume H_0 has at least one edge.

Claim 2.22. *"Having a copy of H_0" has threshold $n^{-1/r_{H_0}}$.*

Proof. We proceed as in Claim 2.21. Let H_0^* be a subgraph of H_0 achieving r_{H_0}. When $p_n \ll n^{-1/r_{H_0}}$, the probability that a copy of H_0^* is in G_n tends to 0 by the argument above. Therefore the same conclusion holds for H_0 itself.

Assume $p_n \gg n^{-1/r_{H_0}}$. Let $I_1, \ldots, I_m$ be an enumeration of the copies of H_0 in a complete graph on the vertices of G_n. Let A_i be the event that $I_i \subseteq G_n$. Using the notation of Corollary 2.20,

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

where

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

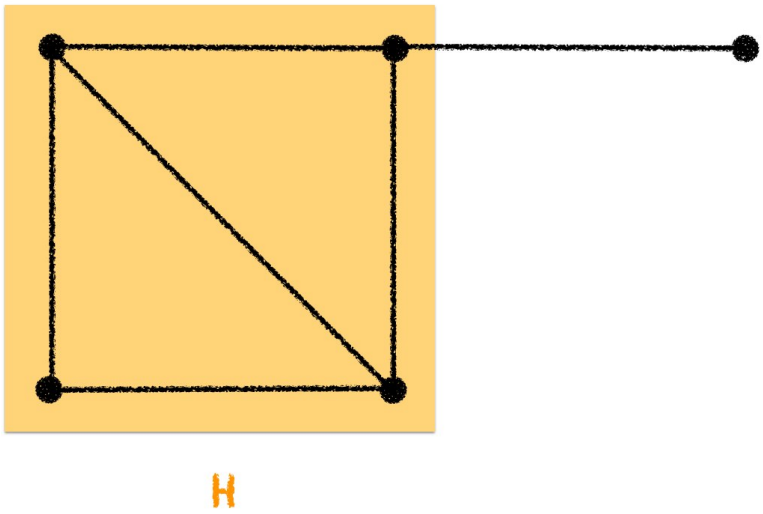

Figure 2.5: Graph H_0 and subgraph H.

Note that $\Phi_{H_0} \to +\infty$. The events A_i and A_j are independent if I_i and I_j share no edge. Otherwise we write $i \sim j$. Note that there are $\Theta(n^{v_H} n^{2(v_{H_0}-v_H)})$ pairs I_i, I_j whose intersection is isomorphic to H. The probability that both elements of such a pair are present in G_n is $\Theta(p_n^{e_H} p_n^{2(e_{H_0}-e_H)})$. Hence

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

The result follows from Corollary 2.20. ■

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

2.2.3 ▷ *Erdös-Rényi graphs: connectivity threshold*

In this section, we use the second moment method to show that the threshold function for connectivity in Erdös-Rényi graphs 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 works the other way too in the following sense: the two thresholds actually *coincide*.

Isolated vertices We begin with isolated vertices.

Claim 2.23. *"Not having an isolated vertex" has threshold function* $\frac{\log n}{n}$.

Proof. Let X_n be the number of isolated vertices in the Erdös-Rényi graph $G_n \sim \mathbb{G}_{n,p_n}$. Using $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$,

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

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

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

$$\mu_n = \sum_i \mathbb{P}_{n,p_n}[A_i] = n(1-p_n)^{n-1} \geq e^{\log n - np_n - np_n^2}, \tag{2.14}$$

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

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

so that

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

Because γ_n is not a $o(\mu_n^2)$, we cannot apply Corollary 2.20. Instead we use Corollary 2.19. We have

$$\begin{aligned} \frac{\mathbb{E}_{n,p_n}[X_n^2]}{(\mathbb{E}_{n,p_n}[X_n])^2} &= \frac{\mu_n + \gamma_n}{\mu_n^2} \\ &\leq \frac{n(1-p_n)^{n-1} + n^2(1-p_n)^{2n-3}}{n^2(1-p_n)^{2n-2}} \\ &\leq \frac{1}{n(1-p_n)^{n-1}} + \frac{1}{1-p_n}, \end{aligned} \tag{2.15}$$

which is $1 + o(1)$ when $p_n \ll \frac{\log n}{n}$. ∎

Connectivity We use Claim 2.23 to study the threshold for connectivity.

Claim 2.24. *Connectivity has threshold function* $\frac{\log n}{n}$.

Proof. We start with the easy direction. If $p_n \ll \frac{\log n}{n}$, Claim 2.23 implies that the graph has isolated vertices, and therefore is disconnected, with probability going to 1 as $n \to +\infty$.

Assume that $p_n \gg \frac{\log n}{n}$. Let $\mathcal{D}_n$ be the event that G_n is disconnected. To bound $\mathbb{P}_{n,p_n}[\mathcal{D}_n]$, for $k \in \{1, \dots, n/2\}$, we consider Y_k, the number of subsets of k vertices that are disconnected from all other vertices in the graph. 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 estimate. Using that $k \leq n/2$ and $k! \geq (k/e)^k$,

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

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

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

where we used that the geometric series is dominated asymptotically by its first term.

■

A more detailed picture 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 (but sub-optimal) 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, the graph is composed of a large connected component together with some isolated vertices.

Claim 2.25.

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

Proof. We first consider the isolated vertices. From (2.14) and (2.15),

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

as $n \to +\infty$.

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

$$\mathbb{P}_{n,p_n}\left[\sum_{k=2}^{n/2} Z_k > 0\right] \leq \mathbb{P}_{n,p_n}\left[\sum_{k=2}^{n/2} Z'_k > 0\right] \leq \sum_{k=2}^{n/2} \mathbb{E}_{n,p_n}[Z'_k]. \tag{2.16}$$

To estimate the expectation of Z'_k, we use Cayley's theorem (e.g. [LP, Corollary 4.5]) which implies that there are k^{k-2} labelled trees on a set of k vertices. Recall further that a tree on k vertices has $k-1$ edges. Hence,

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

where (a) is the number of trees of size k, (b) is the probability that such a tree is present in the graph, and (c) is the probability that this tree is disconnected from every other vertex in the graph. Using that $k! \geq (k/e)^k$,

$$\mathbb{E}_{n,p_n}[Z'_k] \leq \frac{n^k}{k!} k^{k-2} p_n^{k-1} (1-p_n)^{k(n-k)} \leq n \left(e c_n e^{-\left(1-\frac{k}{n}\right)c_n} \right)^k.$$

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

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

Plugging this back into (2.16) concludes the proof. ∎

The limit of $\mathbb{P}_{n,p_n}[Z_1 > 0]$ can be computed explicitly using the method of moments. See Exercise 2.12.

2.2.4 ▷ *Percolation on trees: branching number, weighted second moment method, and application to the critical value*

Consider bond percolation 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 connected component containing 0. In this section we illustrate the use of the first and second moment methods on the identification of the critical value

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

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.

Theorem 2.26.

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

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

$$\theta(p) \le \mathbb{P}_p[X_n > 0] \le \mathbb{E}_p X_n = d(d-1)^{n-1} p^n \to 0, \tag{2.17}$$

when $p < \frac{1}{d-1}$. Here we used that there is a unique path between 0 and any vertex in the tree to deduce that $\mathbb{P}_p[x \in \mathcal{C}_0] = p^n$ for $x \in \partial_n$. Equation (2.17) implies $p_c(\mathbb{T}_d) \ge \frac{1}{d-1}$.

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*, i.e., 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, i.e., the furthest vertex from 0 that lies on both the path from 0 to x and the path from 0 to y; see Figure 2.6. Letting $\mu := \mathbb{E}_p[X_n]$

descendant

most recent common ancestor

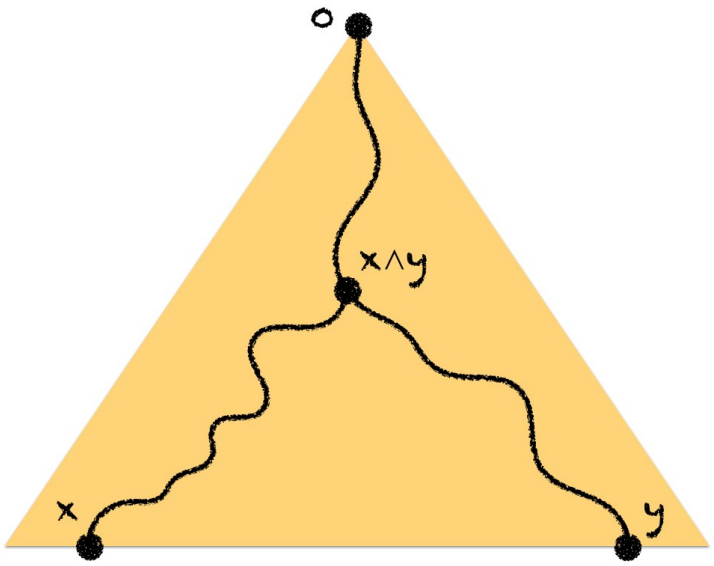

Figure 2.6: Most recent common ancestor of x and y.

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

where, on the third line, we used that all vertices on the n-th level are equivalent and that, for a fixed x, the set $\{y : x\wedge y \in \partial_m\}$ 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 $(\mathbb{E}_p X_n)^2 = \mu_n^2 \to +\infty$, we get

$$\begin{aligned} \frac{\mathbb{E}_p[X_n^2]}{(\mathbb{E}_p X_n)^2} &\leq \frac{1}{\mu_n} + \frac{b-1}{b+1} \cdot \frac{1}{1-(bp)^{-1}} \\ &\leq 1 + \frac{b-1}{b+1} \cdot \frac{1}{1-(bp)^{-1}} \\ &=: C_{b,p}, \end{aligned} \tag{2.18}$$

for n large enough. By the second moment method (version of Corollary 2.19) and monotonicity,

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

which concludes the proof. (Note that the version of the second moment method in Theorem 2.17 does not work here. Subtract 1 in (2.18) and take p close to $1/b$.) ■

The argument above relies crucially on the fact that, in a tree, any two vertices are connected by a unique path. For instance, estimating $\mathbb{P}_p[x \in \mathcal{C}_0]$ 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*, where X_n represents the population size at generation n. The qualitative behavior of a branching process is governed by its expectation: when the mean number of offsprings bp exceeds 1, the process grows exponentially on average and "explodes" with positive probability. We will come back to this point of view in Chapter 5 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. A *cutset* separating 0 and $+\infty$ is a set of edges Π such that all infinite self-avoiding paths starting at 0 go through Π. For a cutset Π, we let $\Pi_{\mathrm{v}} := \{v_e : e \in \Pi\}$. Repeating the argument in (2.17), for any cutset Π, by the first moment method

cutset

$$\theta(p) \leq \mathbb{P}_p[\mathcal{C}_0 \cap \Pi_{\mathrm{v}} \neq \emptyset] \leq \sum_{u \in \Pi_{\mathrm{v}}} \mathbb{P}_p[u \in \mathcal{C}_0] = \sum_{e \in \Pi} p^{|e|}. \tag{2.19}$$

This bound naturally leads to the following definition.

Definition 2.27 (Branching number)**.** *The* branching number *of $\mathscr{T}$ is given by*

branching number

$$\mathrm{br}(\mathscr{T}) = \sup\left\{\lambda \geq 1 : \inf_{\text{cutset } \Pi} \sum_{e \in \Pi} \lambda^{-|e|} > 0\right\}. \tag{2.20}$$

Remark 2.28. *For locally finite trees, it suffices to consider finite cutsets. See the proof of [LP, Theorem 3.1].*

Equation (2.19) 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*.

Theorem 2.29. *For any rooted, locally finite tree $\mathscr{T}$,*

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

Proof. As we argued above, $p_c(\mathscr{T}) \geq \frac{1}{\mathrm{br}(\mathscr{T})}$ follows from (2.19) and (2.20).

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

$$\sum_{e \in \Pi} \lambda^{-|e|} \geq \varepsilon \tag{2.21}$$

for all cutsets Π. The existence of such an ε is guaranteed by the definition of the branching number. As in the proof of Theorem 2.26, we use that $\theta(p)$ is the limit as $n \to +\infty$ of the probability that $\mathcal{C}_0$ reaches the n-th level. However, this time, we use a *weighted* count on the n-th level. Let $\mathscr{T}_n$ be the first n levels of $\mathscr{T}$ and, as before, let ∂_n be the vertices on the n-th level. For a probability measure ν on ∂_n, we define the weighted count

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

The purpose of the denominator is normalization:

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

(Note that multiplying all weights by a constant does not affect the event $\{X_n > 0\}$ and that the constant cancels out in the ratio of $(\mathbb{E}_p X_n)^2$ and $\mathbb{E}_p X_n^2$ in the second moment method.) Because of (2.21), a natural choice of ν follows from the max-flow min-cut theorem which guarantees the existence of a unit flow ϕ from 0 to ∂_n satisfying the capacity constraint $\phi(e) \leq \varepsilon^{-1} \lambda^{-|e|}$, for all edges e in $\mathscr{T}_n$. Define $\nu(z)$ to be the flow entering $z \in \mathscr{T}_n$ under ϕ. In particular, because ϕ is a unit flow, ν restricted to ∂_n defines a probability measure. It remains to bound the second

moment of X_n under this choice. We have

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

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

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

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

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

and $p_c(\mathscr{T}) \le \frac{1}{\mathrm{br}(\mathscr{T})}$. ∎

2.3 Chernoff-Cramér method

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

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

However in many special cases much stronger bounds can be derived. For instance, if X is $N(0,1)$, then by standard inequalities (e.g., [Dur10, Theorem 1.2.3])

$$\mathbb{P}[|X - \mathbb{E}X| \geq \beta] \sim \sqrt{\frac{2}{\pi}}\beta^{-1}\exp(-\beta^2/2) \ll \frac{1}{\beta^2}, \tag{2.22}$$

as $\beta \to +\infty$. In this section we discuss the Chernoff-Cramér method, which gives *exponential* tail inequalities—provided the moment-generating function is finite in an interval around 0.

2.3.1 Tail bounds via the moment-generating function

Under a finite variance, taking a square inside Markov's inequality produces Chebyshev's inequality. This "boosting" can be pushed further when stronger integrability conditions hold. Assume X is a centered random variable such that $M_X(s) < +\infty$ for $s \in (-s_0, s_0)$ for some $s_0 > 0$. Taking an exponential inside Markov's inequality instead gives, for any $\beta > 0$ and $s > 0$,

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

For any $s < s_0$, this observation already leads to an exponential concentration bound on X. But a better bound may be obtained by optimizing the choice of s. We give a few examples below.

Theorem 2.30 (Chernoff-Cramér method)**.** *Let X be a centered random variable such that $M_X(s) < +\infty$ on $s \in (-s_0, s_0)$ for some $s_0 > 0$. Then, for any $\beta > 0$,*

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

where

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

is the so-called Fenchel-Legendre dual of the cumulant-generating function

cumulant-generating function

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

We sometimes refer to the bound

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

as the *Chernoff-Cramér bound*.

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

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

so that

$$\Psi_X^*(\beta) = \sup_{s>0}(s\beta - s^2\nu/2) = \frac{\beta^2}{2\nu}, \tag{2.23}$$

achieved at $s_\beta = \beta/\nu$. Plugging $\Psi_X^*(\beta)$ into Theorem 2.30 leads to the bound

$$\mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right), \tag{2.24}$$

which is much sharper than Chebyshev's inequality—compare to (2.22).

Poisson variables Let Z be Poisson with mean μ and recall that, letting $X = Z - \mu$,

$$\Psi_X(s) = \mu(e^s - s - 1),$$

so that

$$\Psi_X^*(\beta) = \sup_{s>0}(s\beta - \mu(e^s - s - 1)) = \mu\left[\left(1 + \frac{\beta}{\mu}\right)\log\left(1 + \frac{\beta}{\mu}\right) - \frac{\beta}{\mu}\right] := \mu\, h\left(\frac{\beta}{\mu}\right),$$

achieved at $s_\beta = \log\left(1 + \frac{\beta}{\mu}\right)$. Plugging $\Psi_X^*(\beta)$ into Theorem 2.30 leads to the bound

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

Binomial variables and Chernoff bounds The Chernoff-Cramér method is particularly useful for sums of independent random variables as the moment-generating function then factorizes. Let Z be a binomial random variable with parameters n and p. Recall that Z is a sum of i.i.d. indicators $Z_1, \ldots, Z_n$ and, letting $X_i = Z_i - p$ and $X = Z - np$,

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

so that

$$\Psi_X^*(\beta) = \sup_{s>0}(s\beta - n\Psi_{X_1}(s)) = \sup_{s>0} n\left(s\left(\frac{\beta}{n}\right) - \Psi_{X_1}(s)\right) = n\Psi_{X_1}^*\left(\frac{\beta}{n}\right).$$

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

$$\begin{aligned}\Psi_{X_1}^*(b) &= \sup_{s>0}(sb - (\log[pe^s + (1-p)] - ps)) \\ &= (1-a)\log\frac{1-a}{1-p} + a\log\frac{a}{p} =: D(a\|p), \end{aligned} \tag{2.25}$$

achieved at $s_b = \log\frac{(1-p)a}{p(1-a)}$. The function $D(a\|p)$ in (2.25) is the so-called *Kullback-Leibler divergence* or *relative entropy* of Bernoulli variables with parameters a and p respectively. Plugging $\Psi_X^*(\beta)$ into Theorem 2.30 leads to the bound

Kullback-Leibler divergence

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

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

Remark 2.31. *For large deviations, i.e. when $\beta = bn$ for some $b > 0$, the previous bound is tight in the sense that*

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

as $n \to +\infty$. The theory of large deviations, which deals with asymptotics of probabilities of rare events, provides general results along those lines. See e.g. [Dur10, Section 2.6]. Upper bounds will be enough for our purposes.

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

Theorem 2.32 (Chernoff bounds for Poisson trials)**.** *Let $Z_1, \ldots, Z_n$ be independent $\{0,1\}$-valued random variables with $\mathbb{P}[Z_i = 1] = p_i$ and $\mu = \sum_i p_i$. These are called* Poisson trials*. Let $Z = \sum_i Z_i$. Then:*

Poisson trials

1. ***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 $0 < \delta \leq 1$,

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

2. ***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 $0 < \delta < 1$*,*

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

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

2.3.2 ▷ *Randomized algorithms: Johnson-Lindenstrauss, ε-nets, and application to compressed sensing*

In this section we discuss an application of the Chernoff-Cramér method to dimensionality reduction.

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

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

$$\mathbb{E}\|A\boldsymbol{z}\|_2^2 = \mathbb{E}\left[\sum_{i=1}^{d}\left(\sum_{j=1}^{n} A_{ij}z_j\right)^2\right] = d\,\mathrm{Var}\left[\sum_{j=1}^{n} A_{1j}z_j\right] = d\|\boldsymbol{z}\|_2^2, \quad (2.26)$$

where we used the independence of the A_{1j}s and

$$\mathbb{E}\left[\sum_{j=1}^{n} A_{ij}z_j\right] = 0. \quad (2.27)$$

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

Theorem 2.33 (Johnson-Lindenstrauss)**.** *Let $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}$ be arbitrary points in $\mathbb{R}^n$. Fix $\delta, \theta > 0$ and $d \geq \frac{8}{3}\theta^{-2}(\log m + \frac{1}{2}\log \delta^{-1})$. Let A be a $d \times n$ matrix whose entries are independent $N(0,1)$ and consider the linear mapping $L = \frac{1}{\sqrt{d}}A$. Then the following hold with probability at least $1 - \delta$:*

$$(1-\theta)\|\boldsymbol{x}^{(i)}\|_2 \leq \|L\boldsymbol{x}^{(i)}\|_2 \leq (1+\theta)\|\boldsymbol{x}^{(i)}\|_2, \qquad \forall i$$

and

$$(1-\theta)\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2 \leq \|L\boldsymbol{x}^{(i)} - L\boldsymbol{x}^{(j)}\|_2 \leq (1+\theta)\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2, \qquad \forall i, j.$$

Proof. Fix a $\boldsymbol{z} \in \mathbb{R}^n$ with $\|\boldsymbol{z}\|_2 = 1$. To prove the theorem it suffices to show that $\|L\boldsymbol{z}\|_2 \notin [1-\theta, 1+\theta]$ with probability at most $\delta/(m + \binom{m}{2})$ and use a union bound over all points $\boldsymbol{z} = \boldsymbol{x}^{(i)}/\|\boldsymbol{x}^{(i)}\|_2$ and pairwise differences $\boldsymbol{z} = (\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)})/\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_2$.

Recall that a sum of independent Gaussians is Gaussian. So

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

where we used (2.26) and (2.27) to compute the variance. Hence $W = \|A\boldsymbol{z}\|_2^2$ is a sum of squares of independent Gaussians (also known as χ^2-distribution with d degrees of freedom or $\Gamma(d/2, 2)$). Using independence and the change of variable $u = x\sqrt{1-2s}$,

$$M_W(s) = \left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{sx^2}e^{-x^2/2}\,\mathrm{d}x\right)^d = \frac{1}{(1-2s)^{d/2}}, \qquad s < 1/2.$$

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

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

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

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

where we used that $\log(1+x) \leq x - x^2/4$ on $[0,1]$. (To check the inequality, note that the two sides are equal at 0 and compare the derivatives on $[0,1]$.) Our choice of d gives that the r.h.s. is less than δ/m^2. The other direction is similar. ∎

Remark 2.34. *The Johnson-Lindenstrauss lemma is essentially optimal [Alo03]. Note however that it relies crucially on the use of the Euclidean norm [BC03].*

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

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

sensing matrix

Definition 2.35 (Sparse vectors)**.** *We say that a vector $\boldsymbol{z} \in \mathbb{R}^n$ 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 nonzero entries.*

k-sparse vector

To solve the compressed sensing problem in this case, it suffices to find a sensing matrix L satisfying that all subsets of $2k$ columns are linearly independent. Indeed in that case, if $x, x' \in \mathscr{S}_k^n$, then $x - x'$ has at most $2k$ nonzero entries. Hence in order to have $L(x - x') = 0$ it must be that $x - x' = 0$. That implies the required injectivity. The implication goes in the other direction as well. Observe for instance that the matrix used in the Johnson-Lindenstrauss lemma satisfies this property as long as $d \geq 2k$. Because of the continuous density of its entries, the probability that $2k$ of its columns are linearly dependent is 0 when $d \geq 2k$.

For practical applications, however, other requirements must be met: computational efficiency and robustness to measurement noise as well as to the sparsity assumption. We discuss the first two issues (and see Remark 2.40 for the last one which can be dealt with along the same lines). The following definition will play a key role.

Definition 2.36 (Restricted isometry property)**.** *A $d \times n$ linear mapping L satisfies the* (k, θ)-restricted isometry property (RIP) *if for all $\boldsymbol{z} \in \mathscr{S}_k^n$*

restricted isometry property

$$(1 - \theta)\|\boldsymbol{z}\|_2 \leq \|L\boldsymbol{z}\|_2 \leq (1 + \theta)\|\boldsymbol{z}\|_2. \tag{2.28}$$

We say that L is (k, θ)-RIP.

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

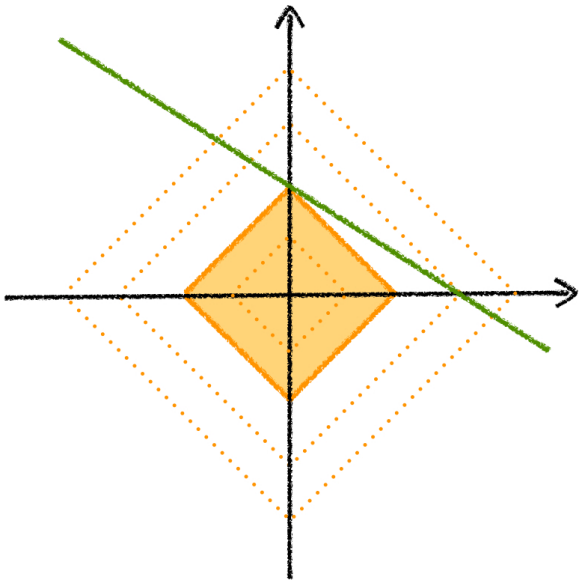

Figure 2.7: Because ℓ^1 balls (in orange) have corners, minimizing the ℓ^1 norm over a linear subspace (in green) tends to produce sparse solutions.

Claim 2.37. *Let L be $(10k, 1/3)$-RIP. Then:*

1. *(Sparse case) For any $\boldsymbol{x} \in \mathscr{S}_k^n$, the unique solution to the following minimization problem*

$$\min_{\boldsymbol{z} \in \mathbb{R}^n} \|\boldsymbol{z}\|_1 \quad \textit{subject to} \quad L\boldsymbol{z} = L\boldsymbol{x}, \tag{2.29}$$

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

2. *(Almost sparse case) For any $\boldsymbol{x} \in \mathbb{R}^n$ with $\eta(\boldsymbol{x}) := \min_{\boldsymbol{x}' \in \mathscr{S}_k^n} \|\boldsymbol{x} - \boldsymbol{x}'\|_1$, the solution to (2.29) satisfies $\|\boldsymbol{z}^* - \boldsymbol{x}\|_2 = O(\eta(\boldsymbol{x})/\sqrt{k})$.*

Claim 2.38. *Let A be a $d \times n$ matrix whose entries are i.i.d. $N(0,1)$ and let $L = \frac{1}{\sqrt{d}} A$. There is a constant $0 < C < +\infty$ such that if $d \geq Ck \log n$ then L is $(k, 1/3)$-RIP with probability at least $1 - 1/n$.*

Roughly speaking, a restricted isometry preserves enough of the structure of $\mathscr{S}_k^n$ to be invertible on its image. The purpose of the ℓ^1 minimization in (2.29) is to promote sparsity. See Figure 2.7 for an illustration. It may seem that a more natural approach is to minimize the number of non-zero entries in $\boldsymbol{z}$. However the advantage of ℓ^1 minimization is that it can be formulated as a linear program, i.e., the

minimization of a linear objective subject to linear inequalities. This permits much faster computation of the solution using standard techniques. See Exercise 2.9.

In Claim 2.38, note that d is much smaller than n and not far from the $2k$ bound we derived above. Note also that Claim 2.38 does *not* follow immediately from the Johnson-Lindenstrauss lemma. Indeed that lemma shows that a matrix with i.i.d. Gaussian entries is an approximate isometry *on a finite set of points*. Here we need a linear mapping that is an approximate isometry for *all* vectors in $\mathscr{S}_k^n$. The proof of this stronger statement uses an important trick, a so-called *ε-net argument.*

ε-nets We start with the proof of Claim 2.38. 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, i.e., the sub-vector $(y_i)_{i \in I}$. Fix a subset of indices $I \subseteq [n]$ of size k. As we mentioned above, the Johnson-Lindenstrauss lemma only applies to a finite collection of vectors. However we need the RIP condition to hold for all $\boldsymbol{z} \in \mathbb{R}^n$ with non-zero entries in I (and all such I). The way out is to notice that, for $\boldsymbol{z} \neq \boldsymbol{0}$, the function $\|L\boldsymbol{z}\|_2 / \|\boldsymbol{z}\|_2$

1. does not depend on the norm of $\boldsymbol{z}$, so that we can restrict ourselves to the compact set $\partial B_I := \{\boldsymbol{z} : \boldsymbol{z}_{[n]\setminus I} = \boldsymbol{0},\ \|\boldsymbol{z}\|_2 = 1\}$, 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 and apply the Johnson-Lindenstrauss lemma to the centers of these balls.

To elaborate on the last point, which is known as an *ε-net argument*, we make the following definition.

Definition 2.39 (ε-net)**.** *Let S be a subset of a metric space (M, ρ) and let $\varepsilon > 0$. The collection of points $N \subseteq S$ is called an* ε-net *of S if all distinct pairs of points in N are at distance at least ε and N is maximal by inclusion in S. In particular for all $z \in S$, $\inf_{y \in N} \rho(z, y) \leq \varepsilon$.* *$\varepsilon$-net*

For $0 < \varepsilon < 1$, let N be an ε-net of the sphere of radius 1 in $\mathbb{R}^k$. We can construct an ε-net by starting with $N = \emptyset$ and successivley adding a point to N at distance at least ε from all other previous points until that is not possible. We claim that $|N| \leq (3/\varepsilon)^k$. Indeed the balls of radius $\varepsilon/2$ around points in N are disjoint by definition and are included in the ball of radius $3/2$ around the origin. The volume of the former is $\frac{\pi^{k/2}(\varepsilon/2)^k}{\Gamma(k/2+1)}$ and that of the latter is $\frac{\pi^{k/2}(3/2)^k}{\Gamma(k/2+1)}$. Dividing one by the other proves the claim.

Proof of Claim 2.38. Let $I \subseteq [n]$ be a subset of indices of size k. There are $\binom{n}{k} \leq n^k = \exp(k \log n)$ such subsets and we denote their collection by $\binom{[n]}{k}$. We let N_I be an ε-net of ∂B_I. The proof strategy is to apply the Johnson-Lindenstrauss lemma to each N_I, $I \in \binom{[n]}{k}$, and use a continuity argument to extend the RIP property to all k-sparse vectors.

- *Continuity argument.* We first choose an appropriate value for ε. Let A^* be the largest entry of A in absolute value. For all $\boldsymbol{y} \in N_I$ and all $\boldsymbol{z} \in \partial B_I$ within distance ε of $\boldsymbol{y}$, by the triangle inequality, we have $\|L\boldsymbol{z}\|_2 \leq \|L\boldsymbol{y}\|_2 + \|L(\boldsymbol{z}-\boldsymbol{y})\|_2$ and $\|L\boldsymbol{z}\|_2 \geq \|L\boldsymbol{y}\|_2 - \|L(\boldsymbol{z}-\boldsymbol{y})\|_2$. Moreover

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

  where we used Cauchy-Schwarz. It remains to bound A^*. For this we use the Chernoff-Cramér bound for Gaussians in (2.24) which implies

$$\mathbb{P}\left[\exists i, j,\ |A_{ij}| \geq C\sqrt{\log n} \right] \leq n^2 e^{-(\sqrt{C \log n})^2/2} \leq \frac{1}{2n}, \tag{2.30}$$

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

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

  and, assuming that the event in (2.30) does not hold, we get

$$|\|L\boldsymbol{z}\|_2 - \|L\boldsymbol{y}\|_2| \leq \frac{1}{6}. \tag{2.31}$$

- *Applying Johnson-Lindenstrauss to the ε-net.* By the argument above,

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

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

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

Then

$$\frac{5}{6} = \frac{5}{6}\|\boldsymbol{y}\|_2 \le \|L\boldsymbol{y}\|_2 \le \frac{7}{6}\|\boldsymbol{y}\|_2 = \frac{7}{6}, \qquad \forall I,\ \forall \boldsymbol{y} \in N_I. \tag{2.32}$$

Assuming (2.31) and (2.32) hold, an event of probability at least $1-2(1/2n) = 1 - 1/n$, we finally get

$$\frac{2}{3} \le \|L\boldsymbol{z}\|_2 \le \frac{4}{3}, \qquad \forall I,\ \forall \boldsymbol{z} \in \partial B_I.$$

That concludes the proof. ■

ℓ^1 minimization It remains to prove Claim 2.37.

Proof of Claim 2.37. We only prove the sparse case $\boldsymbol{x} \in \mathscr{S}_k^n$. For the almost sparse case, see Exercise 2.10. Let $\boldsymbol{z}^*$ be a solution to (2.29) and note that such a solution exists because $\boldsymbol{z} = \boldsymbol{x}$ satisfies the constraint in (2.29). W.l.o.g. assume that only the first k entries of $\boldsymbol{x}$ are non-zero, i.e., $\boldsymbol{x}_{[n]\setminus[k]} = \boldsymbol{0}$. Moreover order the remaining entries of $\boldsymbol{x}$ so that the residual $\boldsymbol{r} = \boldsymbol{z}^* - \boldsymbol{x}$ is so that the entries $\boldsymbol{r}_{[n]\setminus[k]}$ are non-increasing 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 sub-vectors. Let

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

and let $I_{01} = I_0 \cup I_1$, $\bar{I}_i = \bigcup_{j>i} I_j$ and $\bar{I}_{01} = \bar{I}_1$.

We first use the optimality of $\boldsymbol{z}^*$. 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

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

by the triangle inequality. Since $\|\boldsymbol{z}^*\|_1 \le \|\boldsymbol{x}\|_1$ we then have

$$\|\boldsymbol{r}_{\bar{I}_0}\|_1 \le \|\boldsymbol{r}_{I_0}\|_1. \tag{2.33}$$

On the other hand, the RIP condition gives a similar inequality in the other direction. Indeed notice that $L\boldsymbol{r} = \mathbf{0}$ or $L\boldsymbol{r}_{I_{01}} = -\sum_{i\geq 2} L\boldsymbol{r}_{I_i}$ by the constraint in (2.29). Then, by the RIP condition and the triangle inequality, we have that

$$\frac{2}{3}\|\boldsymbol{r}_{I_{01}}\|_2 \leq \|L\boldsymbol{r}_{I_{01}}\|_2 \leq \sum_{i\geq 2} \|L\boldsymbol{r}_{I_i}\|_2 \leq \frac{4}{3}\sum_{i\geq 2} \|\boldsymbol{r}_{I_i}\|_2. \tag{2.34}$$

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

$$\|\boldsymbol{r}_{I_{i+1}}\|_2^2 \leq 9k \left(\frac{\|\boldsymbol{r}_{I_i}\|_1}{9k}\right)^2, \tag{2.35}$$

where we bounded $\boldsymbol{r}_{I_{i+1}}$ entrywise by the expression in parenthesis. Combining (2.33) and (2.35), 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 (2.34) 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 $\boldsymbol{r}_{I_{01}} = \mathbf{0}$. In particular $\boldsymbol{r}_{I_0} = \mathbf{0}$ and, by (2.33), $\boldsymbol{r}_{\bar{I}_0} = \mathbf{0}$ as well. We have shown that $\boldsymbol{r} = \mathbf{0}$ or $\boldsymbol{z}^* = \boldsymbol{x}$.

■

Remark 2.40. *Claim 2.37 can be extended to noisy measurements (using a slight modification of* (2.29)*). See [CRT06b].*

2.3.3 ▷ *Information theory: Shannon's theorem*

To be written. See [AS11, Section 14.1] or [MU05, Section 9.5].

2.3.4 ▷ *Markov chains: Varopoulos-Carne, and a bound on mixing*

For simple random walk on $\mathbb{Z}$, the Chernoff-Cramér method gives the following bound.

Theorem 2.41 (Chernoff bound for simple random walk on $\mathbb{Z}$)**.** *Let $Z_1, \ldots, Z_n$ be independent $\{-1, 1\}$-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 $\beta > 0$,*

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

Proof. Rather than using our result for binomial variables, we argue directly. The moment-generating function of Z_1 can be bounded as follows

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

Taking $s = \beta/n$ in the Chernoff-Cramér bound we get

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

which concludes the proof. ■

Example 2.42 (Set balancing)**.** This is a variant of the balancing vectors problem of Example 2.1. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ be arbitrary non-zero vectors in $\{0,1\}^n$. Think of the $\boldsymbol{v}_i$s as representing subsets of a set S of m elements. We want to partition S into two groups such that the subsets corresponding to the $\boldsymbol{v}_i$s are as balanced as possible. That is, 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. Once again we select each x_i independently, uniformly at random in $\{-1,+1\}$. Fix $\varepsilon > 0$. We claim that

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

By Theorem 2.41 (considering only the non-zero entries of $\boldsymbol{v}_i$),

$$\mathbb{P}\left[|\boldsymbol{x} \cdot \boldsymbol{v}_i| \geq \sqrt{4n \log m}\right] \leq 2 \exp\left(-\frac{2n(\log m + \log(2\varepsilon^{-1}))}{2\|\boldsymbol{v}_i\|_1}\right) \leq \frac{\varepsilon}{m},$$

where we used $\|\boldsymbol{v}_i\|_1 \leq n$. Taking a union bound over the m vectors gives the result. ◀

If (S_t) is simple random walk on $\mathbb{Z}$, then Theorem 2.41 implies that for any $x, y \in \mathbb{Z}$

$$P^t(x,y) \leq e^{-|x-y|^2/2t}, \tag{2.36}$$

where P is the transition matrix of (S_t).

Varopoulos-Carne Interestingly a bound similar to (2.36) holds for any reversible Markov chain. And Theorem 2.41 plays an unexpected role in its proof. An application to mixing times is discussed below.

Theorem 2.43 (Varopoulos-Carne bound)**.** *Let P be the transition matrix of an irreducible Markov chain (X_t) on the countable state space V. Assume further that*

P is reversible with respect to the stationary measure π and that the corresponding network $\mathcal{N}$ is locally finite. Then the following hold

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

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

For a sanity check before proving the theorem, note that if the chain is aperiodic and π is a stationary distribution then

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

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

Proof of Theorem 2.43. The idea of the proof is to show that

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

where S_t is simple random walk on $\mathbb{Z}$ started at 0, and use Theorem 2.41.

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

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

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

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

Chebyshev polynomials

$$\cos((k+1)\theta) + \cos((k-1)\theta) = 2\cos\theta\cos(k\theta)$$

implies the recursion

$$T_{k+1}(\xi) + T_k(\xi) = 2\xi\, T_k(\xi),$$

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

Lemma 2.44.

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

Proof. It suffices to prove

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

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

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

where we used that the probability of $S_t = -t + 2\ell = (+1)\ell + (-1)(t-\ell)$ is the probability of making ℓ steps to the right and $t - \ell$ steps to the left. (Put differently, $(\cos\theta)^t$ is the characteristic function of S_t.) Take real parts and use $\cos(k\theta) = \cos(-k\theta)$. ∎

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

Lemma 2.45.

$$T_k(P)(x, y) = 0, \qquad \forall k < \rho(x, y).$$

and

$$T_k(P)(x, y) \le \sqrt{\frac{\pi(y)}{\pi(x)}}, \qquad \forall k \ge \rho(x, y).$$

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

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

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

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

$$\|T_k(P)f\|_\pi^2 = \sum_{i=1}^n \alpha_i^2 \, T_k(\lambda_i)^2 \langle f_i, f_i\rangle_\pi \le \sum_{i=1}^n \alpha_i^2 \langle f_i, f_i\rangle_\pi = \|f\|_\pi^2, \tag{2.37}$$

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

Let δ_x denote the point mass at x. By Cauchy-Schwarz and (2.37),

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

for $k \ge \rho(x, y)$. ∎

Combining the two lemmas gives the result. ∎

Remark 2.46. *The local finiteness assumption is made for simplicity only. The result holds for any countable-space, reversible chain. See [LP, Section 13.2].*

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

$$\mathrm{t}_{\mathrm{mix}}(\varepsilon) = \min\{t \,:\, d(t) \le \varepsilon\},$$

where

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

It is intuitively clear that $\mathrm{t}_{\mathrm{mix}}(\varepsilon)$ is at least of the order of the "diameter" of the transition graph of P. For $x, y \in V$, let $\rho(x, y)$ be the graph distance between x and y on the undirected version of the transition graph, i.e., ignoring the orientation of the edges. With this definition, a shortest directed path from x to y contains at least $\rho(x, y)$ edges. Here we define the *diameter* of the transition graph as $\Delta := \max_{x,y \in V} \rho(x, y)$. Let x_0, y_0 be a pair of vertices achieving the diameter. Then we claim that $P^{\lfloor(\Delta-1)/2\rfloor}(x_0, \cdot)$ and $P^{\lfloor(\Delta-1)/2\rfloor}(y_0, \cdot)$ are supported on disjoint sets. To see this let

diameter

$$A = \{z \in S \,:\, \rho(x_0, z) < \rho(y_0, z)\}.$$

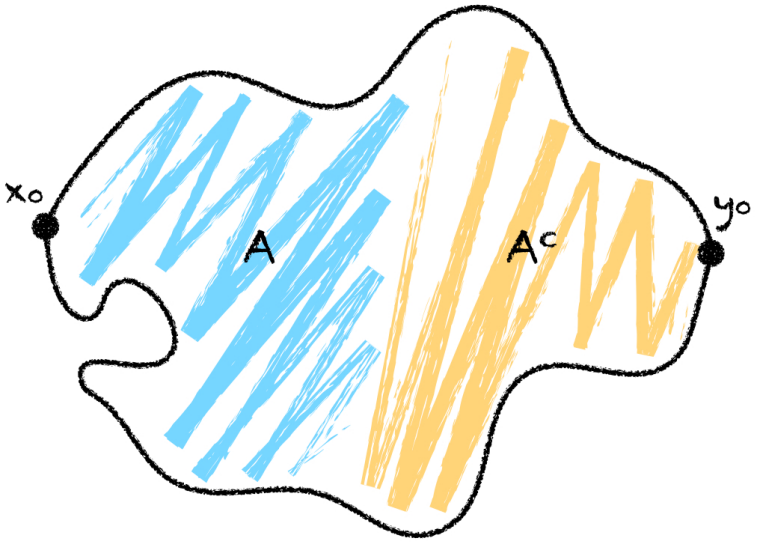

Figure 2.8: The supports of $P^{\lfloor(\Delta-1)/2\rfloor}(x_0, \cdot\,)$ and $P^{\lfloor(\Delta-1)/2\rfloor}(y_0, \cdot\,)$ are contained in A and A^c respectively.

See Figure 2.3.4. By the triangle for ρ, any z such that $\rho(x_0, z) \leq \lfloor(\Delta-1)/2\rfloor$ is in A, otherwise we would have $\rho(y_0, z) \leq \rho(x_0, z) \leq \lfloor(\Delta-1)/2\rfloor$ and hence $\rho(x_0, y_0) \leq \rho(x_0, z) + \rho(y_0, z) \leq 2\lfloor(\Delta-1)/2\rfloor < \Delta$. Similarly, if $\rho(y_0, z) \leq \lfloor(\Delta-1)/2\rfloor$, then $z \in A^c$. By the triangle inequality for the total variation distance,

$$\begin{aligned} d(\lfloor(\Delta-1)/2\rfloor) &\geq \frac{1}{2}\left\| P^{\lfloor(\Delta-1)/2\rfloor}(x_0, \cdot\,) - P^{\lfloor(\Delta-1)/2\rfloor}(y_0, \cdot\,)\right\|_{\mathrm{TV}} \\ &\geq \frac{1}{2}\left\{ P^{\lfloor(\Delta-1)/2\rfloor}(x_0, A) - P^{\lfloor(\Delta-1)/2\rfloor}(y_0, A)\right\} \\ &= \frac{1}{2}\{1-0\} = \frac{1}{2}, \end{aligned} \tag{2.38}$$

so that:

Claim 2.47.

$$\mathrm{t}_{\mathrm{mix}}(\varepsilon) \geq \frac{\Delta}{2}.$$

This bound is often far from the truth. Consider for instance simple random walk on a cycle of size n. The diameter is $\Delta = n/2$. But Theorem 2.41 suggests that it takes time of order Δ^2 to reach the antipode of the starting point. More generally, when P is reversible, we use the Varopoulos-Carne bound to show that

the mixing time does indeed scale at least as the *square* of the diameter. Assume that P is reversible with respect to π and has diameter Δ. Letting $n = |V|$ and $\pi_{\min} = \min_{x \in V} \pi(x)$, we have the following.

Claim 2.48.

$$\mathrm{t}_{\mathrm{mix}}(\varepsilon) \geq \frac{\Delta^2}{12 \log n + 4|\log \pi_{\min}|}, \qquad \textit{if } n \geq \frac{16}{(1-2\varepsilon)^2}.$$

Proof. The proof is based on the same argument we used to derive our first diameter-based bound, except that the Varopoulos-Carne bound gives a better dependence on the diameter. Namely, let x_0, y_0, and A be as above. By the Varopoulos-Carne bound,

$$P^t(x_0, A^c) = \sum_{z \in A^c} P^t(x_0, z) \leq \sum_{z \in A^c} 2\sqrt{\frac{\pi(z)}{\pi(x_0)}} e^{-\frac{\rho^2(x_0,z)}{2t}} \leq 2n\pi_{\min}^{-1/2} e^{-\frac{\Delta^2}{8t}},$$

where we used that $|A^c| \leq n$ and $\rho(x_0, z) \geq \frac{\Delta}{2}$ for $z \in A^c$. For any

$$t < \frac{\Delta^2}{12 \log n + 4|\log \pi_{\min}|},$$

we get that

$$P^t(x_0, A^c) \leq 2n\pi_{\min}^{-1/2} \exp\left(-\frac{3\log n + |\log \pi_{\min}|}{2}\right) = \frac{2}{\sqrt{n}}.$$

Similarly, $P^t(y_0, A) \leq \frac{2}{\sqrt{n}}$ so that arguing as in (2.38)

$$d(\lfloor (\Delta - 1)/2 \rfloor) \geq \frac{1}{2}\left\{1 - \frac{2}{\sqrt{n}} - \frac{2}{\sqrt{n}}\right\} = \frac{1}{2} - \frac{2}{\sqrt{n}} \geq \varepsilon,$$

for n as in the statement. ∎

Remark 2.49. *The dependence on Δ and $\pi_{\min}$ in Claim 2.48 cannot be improved. See [LP, Section 13.3].*

2.3.5 Hoeffding's and Bennett's inequalities

The bounds in Section 2.3 were obtained by computing the moment-generating function explicitly. This is seldom possible. In this section, we give a few more examples of concentration inequalities derived from the Chernoff-Cramér method for broad classes of random variables under natural conditions on their distributions.

Sub-Gaussian variables We say that a centered random variable X is *sub-Gaussian with variance factor* $\nu > 0$ if for all $s \in \mathbb{R}$

$$\Psi_X(s) \leq \frac{s^2\nu}{2},$$

which is denoted by $X \in \mathcal{G}(\nu)$. Note that the r.h.s. is the cumulant-generating function of a $N(0,\nu)$. By the Chernoff-Cramér method and (2.23) it follows immediately that

$$\mathbb{P}\left[X \leq -\beta\right] \vee \mathbb{P}\left[X \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\nu}\right), \tag{2.39}$$

where we used that $X \in \mathcal{G}(\nu)$ implies $-X \in \mathcal{G}(\nu)$.

Hoeffding's inequality For bounded random variables, the following tail inequality holds.

Theorem 2.50 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables where, for each i, X_i takes values in $[a_i, b_i]$ with $-\infty < a_i \leq b_i < +\infty$. Let $S_n = \sum_{i \leq n} X_i$. For all $t > 0$,*

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

We first give a quick proof of a weaker version. Suppose the X_is are centered and satisfy $|X_i| \leq c_i$ for some $c_i > 0$. To estimate the moment-generating function of X_i, observe that on $x \in [-c, c]$

$$e^{sx} \leq \frac{e^{sc} + e^{-sc}}{2} + \frac{e^{sc} - e^{-sc}}{2}\left(\frac{x}{c}\right),$$

as the r.h.s. is the line through e^{-sc} and e^{sc}. So, by a Taylor expansion and using $\mathbb{E}X_i = 0$, we have the bound

$$\mathbb{E}\left[e^{sX_i}\right] \leq \cosh(sc_i) = \sum_{k\geq 0}\frac{(sc_i)^{2k}}{(2k)!} \leq \sum_{k\geq 0}\frac{((sc_i)^2)^k}{2^k k!} = e^{(sc_i)^2/2},$$

that is, X_i is sub-Gaussian with variance factor c_i^2. By independence, S_n is sub-Gaussian with variance factor $\sum_{i\leq n} c_i^2$. Finally, by (2.39),

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

Proof of Theorem 2.50. The idea of the proof is to establish that $S_n - \mathbb{E}S_n$ is sub-Gaussian with variance factor $\frac{1}{4}\sum_{i\le n}(b_i - a_i)^2$. The tail bound then follows from (2.39). Because

$$\Psi_{S_n - \mathbb{E}S_n}(s) = \sum_{i\le n} \Psi_{X_i - \mathbb{E}X_i}(s),$$

it suffices in fact to show that $X_i - \mathbb{E}X_i$ is sub-Gaussian with variance factor $\frac{1}{4}(b_i - a_i)^2$. This follows from the next lemma.

Lemma 2.51 (Hoeffding's lemma)**.** *Let X be a random variable taking values in $[a, b]$ for $-\infty < a \le b < +\infty$. Then $X - \mathbb{E}X \in \mathcal{G}\left(\frac{1}{4}(b-a)^2\right)$.*

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 w.l.o.g. we assume $\mathbb{E}X = 0$. Because X is bounded, $M_X(s)$ is finite for all $s \in \mathbb{R}$. From standard results on moment-generating functions (e.g., [Bil12, Section 21]; see also [Dur10, Theorem A.5.1]), for any $k \in \mathbb{Z}$,

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

Hence

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

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

$$\Psi''_X(s) \le \frac{1}{4}(b-a)^2. \tag{2.40}$$

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 (i.e., a Radon-Nikodym derivative) on $[a, b]$ with respect to the law of X. The variance under this density—the last line above—must be less than $\frac{1}{4}(b-a)^2$. Indeed for any random variable Z taking values in $[a, b]$ we have

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

and

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

This establishes (2.40) and concludes the proof. ■

Remark 2.52. *The change of measure above is known as* tilting *and is a standard trick in large deviation theory. See, e.g., [Dur10, Section 2.6].*

■

Bennett's inequality To be written. See [BLM13, Section 2.7] or [Lug, Section 3.2].

2.3.6 ▷ *Knapsack: probabilistic analysis*

To be written. See [FR98, Section 5.3].

2.4 Matrix tail bounds

The Chernoff-Cramér method has several important extensions. We will discuss a martingale version in Section 3.2. Here we consider the case of sums of independent random matrices. That such an extension is possible is surprising because a key step in the Chernoff-Cramér bound, that exponentials of sums are products of exponentials, fails to hold for matrices. That is, in general

$$e^{A+B} \neq e^A e^B,$$

unless A and B commute.

To be written. See [Har, Lectures 12, 13, 15].

2.4.1 Ahlswede-Winter inequality

2.4.2 ▷ *Randomized algorithms: low-rank approximations*

Exercises

Exercise 2.1 (Bonferroni inequalities)**.** Let $A_1, \dots, A_n$ 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

$$X_n := \sum_{i=1}^{n} \mathbb{1}_{A_i}.$$

a) 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.

b) Show that, for all r,

$$\sum_{1 \le i_1 < \cdots < i_r \le n} \mathbb{1}_{A_{i_1}} \mathbb{1}_{A_{i_2}} \cdots \mathbb{1}_{A_{i_r}} = \binom{X_n}{r}.$$

c) Use a) and b) to show that when $\ell \in [n]$ is odd

$$\mathbb{P}[B_n] \le \sum_{r=1}^{\ell} (-1)^{r-1} S^{(r)},$$

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 2.2 (Percolation on $\mathbb{Z}^2$: better bound [Ste])**.** Let E_1 be the event that all edges are open in $[-N, N] \times [-N, N]$ and E_2 be the event that there is no closed self-avoiding dual cycle surrounding $[-N, N]^2$. By looking at $E_1 \cap E_2$, show that $\theta(p) > 0$ for $p > 2/3$.

Exercise 2.3 (Percolation on $\mathbb{Z}^d$: phase transition)**.** Consider bond percolation on $\mathbb{L}^d$.

a) Show that $p_c(\mathbb{L}^d) > 0$. [Hint: Count self-avoiding paths.]

b) Show that $p_c(\mathbb{L}^d) < 1$. [Hint: Use the result for $\mathbb{L}^2$.]

Exercise 2.4 (Sums of uncorrelated variables)**.** Centered random variables $X_1, \dots, X_n$ are *pairwise uncorrelated* if

pairwise uncorrelated variables

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

Assume further that $\mathrm{Var}[X_r] \leq C < +\infty$ for all r. Show that

$$\mathbb{P}\left[\frac{1}{n}\sum_{r\leq n} X_r \geq \beta\right] \leq \frac{C^2}{\beta^2 n}.$$

Exercise 2.5 (Pairwise independence: lack of concentration [LW06])**.** Let $\boldsymbol{U} = (U_1, \dots, U_\ell)$ be uniformly distributed over $\{0,1\}^\ell$. Let $n = 2^\ell - 1$. For all $\boldsymbol{v} \in \{0,1\}^\ell \backslash \mathbf{0}$, define

$$X_{\boldsymbol{v}} = (\boldsymbol{U} \cdot \boldsymbol{v}) \mod 2.$$

a) 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.

b) 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 2.4 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 2.6 (Chernoff bound for Poisson trials)**.** Using the Chernoff-Cramér method, prove part (a) of Theorem 2.32. Show that part (b) follows from part (a).

Exercise 2.7 (A proof of Pólya's theorem)**.** Let (X_t) be simple random walk on $\mathbb{L}^d$ started at the origin 0.

a) For $d = 1$, use Stirling's formula to show that $\mathbb{P}_0[X_{2n} = 0] = \Theta(n^{-1/2})$.

b) For $j = 1, \dots, d$, let $N_t^{(j)}$ be the number of steps in the j-th 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 $\kappa_d > 0$.

c) Use a) and b) to show that, for any $d \geq 1$, $\mathbb{P}_0[X_{2n} = 0] = \Theta(n^{-d/2})$.

Exercise 2.8 (RIP v. orthogonality)**.** Show that a $(k, 0)$-RIP matrix with $k \geq 2$ is orthogonal, i.e., its columns are orthonormal.

Exercise 2.9 (Compressed sensing: linear programming formulation)**.** Formulate (2.29) as a linear program, i.e., the minimization of a linear objective subject to linear inequalities.

Exercise 2.10 (Compressed sensing: almost sparse case)**.** Prove the almost sparse case of Claim 2.37 by adapting the proof of the sparse case.

Exercise 2.11 (Poisson convergence: method of moments)**.** Let $A_1, \ldots, A_n$ be events and $A := \cup_i A_i$. Define

$$S^{(r)} := \sum_{1 \leq i_1 < \cdots < i_r \leq n} \mathbb{P}[A_{i_1} \cap \cdots \cap A_{i_r}],$$

and

$$X_n := \sum_{i=1}^{n} A_i.$$

Assume that there is $\mu > 0$ such that, for all r,

$$S^{(r)} \to \frac{\mu^r}{r!}.$$

Use Exercise 2.1 and a Taylor expansion of $e^{-\mu}$ to show that

$$\mathbb{P}[X_n = 0] \to e^{-\mu}.$$

In fact, $X_n \xrightarrow{\mathrm{d}} \mathrm{Poi}(\mu)$ (no need to prove this). This is a special case of the *method of moments*. See e.g. [Dur10, Section 3.3.5] and [JLR11, Section 6.1].

Exercise 2.12 (Connectivity: critical window)**.** Using Exercise 2.11 show that, when $p_n = \frac{\log n + s}{n}$, the probability that an Erdös-Rényi graph $G_n \sim \mathbb{G}_{n,p_n}$ contains no isolated vertex converges to $e^{-e^{-s}}$.

Bibliographic remarks

Section 2.1 The examples in Section 2.1.1 are taken from [AS11, Sections 2.4, 3.2]. A fascinating account of the longest increasing subsequence problem is given in [Rom14], from which the material in Section 2.1.3 is taken. The contour lemma, Lemma 2.12, 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 2.2 The presentation in Section 2.2.2 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 Erdös and Rényi [ER60]. For variants of the small subgraph containment problem involving copies that are induced, disjoint, isolated etc., see e.g. [JLR11, Chapter 3]. For corresponding results for larger graphs, such as cycles or matchings, see e.g. [Bol01]. The result in that section is due to Erdös and Rényi [ER60]. The connectivity threshold in Section 2.2.3 is also due to the same authors [ER59]. The presentation here follows [vdH14, Section 5.2]. Theorem 2.29 is due to R. Lyons [Lyo90].

Section 2.3 The use of the moment-generating function to derive tail bounds for sums of independent random variables was pioneered by Cramér [Cra38], Bernstein [Ber46], and Chernoff [Che52]. For much more on concentration inequalities, see e.g. [BLM13]. The basics of large deviation theory are covered in [Dur10, Section 2.6]. See also [RAS] and [DZ10]. The presentation in Section 2.3.2 is based on [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 e.g. [Har, Lecture 7]. Compressed sensing emerged in the works of Donoho [Don06] and Candès, Romberg and Tao [CRT06a, CRT06b]. The restricted isometry property was introduced by Candès and Tao [CT05]. Claim 2.37 is due to Candés, Romberg and Tao [CRT06b]. The proof of Claim 2.38 presented here is due to Baraniuk et al. [BDDW08]. A survey of compressed sensing can be found in [CW08]. The presentation in Section 2.3.4 follows [KP, Section 3] and [LP, Section 13.3]. The Varopoulos-Carne bound is due to Carne [Car85] and Varopoulos [Var85]. For a probabilistic approach to the Varopoulos-Carne bound see Peyre's proof [Pey08]. The application to mixing times is from [LP]. The material in Section 2.3.5 can be found in [BLM13, Chapter 2]. Hoeffding's lemma and inequality are due to Hoeffding [Hoe63]. Bennett's inequality is due to Bennett [Ben62].