CSE 525: Randomized algorithms & probabilistic analysis
Lecture notes

Spring 2019

James R. Lee
Paul G. Allen School of Computer Science & Engineering
University of Washington

Contents

1 First moments

1.1 The probabilistic method

An old math puzzle goes: Suppose there are six people in a room; some of them shake hands. Prove that there are at least three people who all shook each others' hands or three people such that no pair of them shook hands.

Generalized a bit, this is the classic Ramsey problem. The *diagonal Ramsey numbers* $R(k)$ are defined as follows. $R(k)$ is the smallest integer n such that in every two-coloring of the edges of the complete graph K_n by red and blue, there is a monochromatic copy of K_k, i.e. there are k nodes such that all of the $\binom{k}{2}$ edges between them are red or all of the edges are blue. A solution to the puzzle above asserts that $R(3) \leqslant 6$ (and it is easy to check that, in fact, $R(3) = 6$).

In 1929, Ramsey proved that $R(k)$ is finite for every k. We want to show that $R(k)$ must grow pretty fast; in fact, we'll prove that for $k \geqslant 3$, we have $R(k) > \lfloor 2^{k/2} \rfloor$. This requires finding a coloring of K_n that doesn't contain any monochromatic K_k. To do this, we'll use the *probabilistic method:* We'll give a random coloring of K_n and show that it satisfies our desired property with positive probability. This proof appeared in a paper of Erdös from 1947, and this is the example that starts Alon and Spencer's famous book devoted to the probabilistic method.

Lemma 1.1. *If* $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$, *then* $R(k) > n$. *In particular,* $R(k) > \lfloor 2^{k/2} \rfloor$ *for* $k \geqslant 3$.

Proof. Consider a uniformly random 2-coloring of the edges of K_n. Every edge is colored red or blue independently with probability half each. For any fixed set of k vertices H, let $\mathcal{E}_H$ denote the event that the induced subgraph on H is monochromatic. An easy calculation yields

$$\mathbb{P}(\mathcal{E}_H) = 2 \cdot 2^{-\binom{k}{2}}.$$

Since there are $\binom{n}{k}$ possible choices for H, we can use the union bound:

$$\mathbb{P}(\text{exists } R \text{ such that } \mathcal{E}_H) \leqslant 2 \cdot 2^{-\binom{k}{2}} \cdot \binom{n}{k}.$$

Thus if $2^{1-\binom{k}{2}}\binom{n}{k} < 1$, then with positive probability, no event $\mathcal{E}_H$ occurs. Thus there must exist at least one coloring with no monochromatic K_k. One can check that if $k \geqslant 3$ and $n = \lfloor 2^{k/2} \rfloor$, then this is satisfied. □

We have employed the following basic tool.

Tool 1.2 (Union bound)**.** If $A_1, A_2, \ldots, A_m$ are arbitrary events, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_m) \leqslant \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_m)$$

1.2 Linearity of expectation

Let's look at a couple more examples of the probabilistic method in action. We'll use a basic fact in probability: Linearity of expectation.

Tool 1.3 (Linearity of expectation)**.** If $X_1, X_2, \ldots, X_n$ are discrete real-valued random variables, then

$$\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$$

The great fact about this inequality is that we don't need to know anything about the relationships between the random variables; linearity of expectation holds no matter what the dependence structure.

MAX-3SAT. Let's consider a 3-CNF formula over the variables $x_1, x_2, \ldots, x_n$. Such a formula has the form $\varphi = C_1 \wedge C_2 \wedge \cdots \wedge C_m$ where each clause is an OR of three literals involving distinct variables: $C_i = z_{i_1} \vee z_{i_2} \vee z_{i_3}$. A literal is a variable or its negation. For instance, $(x_2 \vee \bar{x}_3 \vee \bar{x}_4) \wedge (x_3 \vee \bar{x}_5 \vee \bar{x}_1) \wedge (x_1 \vee x_5 \vee x_4)$ is a 3-CNF formula.

Claim 1.4. *If φ is a 3-CNF formula with m clauses, then there exists an assignment that makes at least $\frac{7}{8}m$ clauses evaluate to true.*

Proof. We will prove this using the probabilistic method. For every variable independently, we choose a uniformly random truth assignment: true or false each with probability 1/2. Let A_i equal 1 if clause C_i is satisfied by our random assignment, and equal 0 otherwise. Then $\mathbb{P}(A_i = 1) = 7/8$ because there are 7 ways to satisfy a clause out of the 8 possible truth values for its literals.

Let $A = A_1 + \cdots + A_m$ denote the total number of satisfied clauses. By linearity of expectation, we have

$$\mathbb{E}[A] = \sum_{i=1}^{m} \mathbb{E}[A_i] = \frac{7}{8}m\,. \tag{1.1}$$

Since a random assignment satisfies $\frac{7}{8}m$ clauses in expectation, there must exist *at least one* assignment that satisfies this many clauses. □

MAX-CUT. Consider an undirected graph $G = (V, E)$. A *cut* is a subset $S \subseteq V$, and we use $E(S, \bar{S})$ to denote the set of edges *crossing the cut* S. This is the set of edges with one endpoint in S and one not in S.

Claim 1.5. *In any graph $G = (V, E)$, there exists a cut $S \subseteq V$ that cuts at least half the edges, i.e., $|E(S, \bar{S})| \geqslant \frac{|E|}{2}$.*

Proof. We construct a random set $S \subseteq V$ by including every vertex in S independently with probability 1/2. For an edge $e \in E$, let $A_e = 1$ if e crosses the cut S, and 0 otherwise. First, it should be apparent that $\mathbb{P}(A_e = 1) = 1/2$. Therefore by linearity of expectation,

$$\mathbb{E}\left[|E(S, \bar{S})|\right] = \sum_{e \in E} \mathbb{E}[A_e] = \frac{|E|}{2}\,.$$

Thus there must exist at least one cut S that has at least half the edges crossing it. □

1.3 The method of conditional expectation

Claim 1.4 asserts that there *exists* an assignment satisfying at least $\frac{7}{8}m$ clauses, but what if we wish to actually find one? One way is to randomly sample from the underlying distribution and then check the resulting assignment. Analyzing the probability of success will require our first *tail bound*; we'll get there in the next section.

Let's examine another way that actually results in a deterministic algorithm. Let $S(x_1, x_2, \ldots, x_n)$ denote the expected number of satisfied clauses given a partial truth assignment to the input variables, where we choose the unassigned variables uniformly at random. We will use T to denote true, F to denote false, and $\star$ to denote that no assignment has been chosen for that variable.

For instance, $S(\star, \star, \ldots, \star)$ denotes the expected number of satisfied clauses in a random assignment, and we have already seen (cf. (1.1)) that

$$S(\star, \star, \ldots, \star) = \frac{7}{8}m.$$

Note that a simple linear-time algorithm can estimate $S(x_1, x_2, \ldots, x_n)$ for any partial assignment $x_1, \ldots, x_n \in \{\mathsf{T}, \mathsf{F}, \star\}$ by simply going through the clauses one by one.

As an example, consider the clause $x_1 \vee \bar{x}_2 \vee \bar{x}_4$. The probability that a random assignment satisfies this is 7/8. If we assign $x_1 = \mathsf{F}$, then the probability becomes 3/4, and if we set $x_1 = \mathsf{T}$, then the probability becomes 1.

Observe that

$$S(\star, \star, \ldots, \star) = \frac{1}{2} S(\mathsf{F}, \star, \ldots, \star) + \frac{1}{2} S(\mathsf{T}, \star, \ldots, \star) .$$

Since $S(\star, \star, \ldots, \star) \geqslant \frac{7}{8}m$, it must hold that $S(\mathsf{F}, \star, \ldots, \star) \geqslant \frac{7}{8}m$ or $S(\mathsf{T}, \star, \ldots, \star) \geqslant \frac{7}{8}m$. As we have just argued, it's possible to compute both these quantities and figure out which is larger. We can then set x_1 to the corresponding value and keep assigning truth values recursively. Eventually, this process ends at a full assignment to the variables that satisfies at least $\frac{7}{8}m$ clauses. The key property we employed here is the ability to efficiently compute the conditional expectation of the underlying random variable under a partial assignment.

1.4 Markov's inequality

The probabilistic method shows the *existence* of an object, but it doesn't necessarily give us a randomized algorithm to construct it. If we just know that the probability of an event is non-zero, it could still be very tiny; we might need to do an arbitrarily large number of random experiments before we get a positive outcome. Sometimes we can say more.

Tool 1.6 (Markov's inequality)**.** Let X be a non-negative random variable. Then for any $\alpha > 0$, we have

$$\mathbb{P}[X \geqslant \alpha] \leqslant \frac{\mathbb{E}\, X}{\alpha} .$$

The proof of this lemma is easy; we leave it as an exercise.

Consider now our **MAX-3SAT** example above. Let X denote the number of *unsatisfied* clauses in a random truth assignment. We know from the preceding analysis that $\mathbb{E}[X] \leqslant \frac{1}{8}m$. Markov's inequality tells us that for any $\varepsilon > 0$,

$$\mathbb{P}\left[X > \left(\frac{1}{8} + \varepsilon\right) m\right] \leqslant \frac{m/8}{(1/8 + \varepsilon)m} = \frac{1}{1 + 8\varepsilon} \leqslant 1 - \varepsilon .$$

The last inequality is only true if we assume $\varepsilon \leqslant 7/8$, but for any value $\varepsilon > 7/8$, the probability is clearly zero.

This means that, with probability at least ε, we will get an assignment that satisfies at least $(7/8 - \varepsilon)$-fraction of clauses. So in expectation, after $1/\varepsilon$ samples, we will get an assignment that is very close to the one guaranteed to exist. The same kind of reasoning applies to our **MAX-CUT** analysis.

1.5 Crossing number inequalities

Let's look at one more application of the linearity of expectation. It is almost as elementary as the examples above, but has some powerful consequences in incidence geometry and sum-product estimates.

If $G = (V, E)$ is an undirected graph, we use the notation $\mathrm{cr}(G)$ to denote the *crossing number of* G. This is the minimum number of edge crossings required to draw G in the plane. A drawing of

the graph means that the vertices are mapped to distinct points, and each edge is drawn as a closed, continuous curve of bounded length. The following result is due independently to Leighton and Atjai-Chvatal-Newborn-Szemeredi.

Theorem 1.7. *If G is a graph with n vertices and m edges, and $m \geqslant 4n$, then*

$$\mathrm{cr}(G) \geqslant \frac{m^3}{64n^2}\,.$$

Note that for dense graphs, i.e. those with $m = \Omega(n^2)$, we get $\Omega(n^4)$ crossings (the most possible up to a constant factor). We start with a basic fact: Euler's formula implies that, in every planar graph (a planar graph G is one for which $\mathrm{cr}(G) = 0$), we have $m \leqslant 3n - 6$.

Thus if $m > 3n$, we must have $\mathrm{cr}(G) \geqslant 1$. Since we can always remove one crossing from a drawing by removing one edge from the underlying graph, this gives us

$$\mathrm{cr}(G) \geqslant m - 3n\,. \tag{1.2}$$

This is still pretty weak. But now we will use random sampling to do seriously heavy amplification.

Proof of Theorem 1.7. Suppose we have a drawing of G in the plane. We will make some assumptions about this drawing (which are without loss of generality). We may assume that every edge crossing involves *four* distinct vertices. If an edge crosses itself, that can be fixed by short-circuiting the loops. If two edges emanating from the same vertex cross each other, they can be uncrossed without affecting the rest of the drawing (draw a picture to convince yourself). So we may assume that the only crossings are between edges $\{x, y\}$ and $\{u, v\}$ where x, y, u, v are all distinct vertices.

Now we will construct a (random) graph G_p by keeping every vertex of G independently with probability p. The value of p will be chosen soon. Let n_p and m_p denote the number of edges and vertices remaining in G_p, and let c_p denote the number of crossings remaining in our drawing (after the edges and vertices not remaining in G_p are removed).

Every vertex remains with probability p. By independence, an edge remains with probability p^2. Finally, a crossing remains with probability p^4 since we said that every crossing has to involve four distinct vertices. In order for a crossing to remain, all of those four vertices must be in G_p. Thus linearity of expectation gives us:

$$\mathbb{E}[n_p] = pn \tag{1.3}$$
$$\mathbb{E}[m_p] = p^2 m \tag{1.4}$$
$$\mathbb{E}[c_p] = p^4 \mathrm{cr}(G)\,. \tag{1.5}$$

But from (1.2), we know that $c_p \geqslant m_p - 3n_p$, and thus $\mathbb{E}[c_p] \geqslant \mathbb{E}[m_p] - 3\,\mathbb{E}[n_p]$. Plugging in our values above yields

$$p^4 \mathrm{cr}(G) \geqslant p^2 m - 3pn\,,$$

or equivalently

$$\mathrm{cr}(G) \geqslant \frac{m}{p^2} - \frac{3n}{p^3}\,.$$

Finally, we set $p = \frac{4n}{m}$ ($p \leqslant 1$ since we have assumed $m \geqslant 4n$). This yields

$$\mathrm{cr}(G) \geqslant \frac{m^3}{16n^2} - \frac{3m^3}{64n^2} = \frac{m^3}{64n^2}\,,$$

completing our proof. □

2 Second moments

2.1 Threshold phenomena in random graphs

Consider a positive integer n and value $p \in [0,1]$. Perhaps the simplest model of random (undirected) graphs is $\mathcal{G}_{n,p}$. To sample a graph from $\mathcal{G}_{n,p}$, we add every edge $\{i,j\}$ (for $i \neq j$ and $i,j \in \{1,\dots,n\}$) independently with probability p. For example, if X denotes the number of edges in a $\mathcal{G}_{n,p}$ graph, then $\mathbb{E}\,X = p\binom{n}{2}$.

A 4-*clique* in a graph is a set of four nodes such that all $\binom{4}{2} = 6$ possible edges between the nodes are present. Let G a random graph sampled according to $\mathcal{G}_{n,p}$, and let C_4 denote the event that G contains a 4-clique. It will turn out that if $p \gg n^{-2/3}$, then G contains a 4-clique with probability close to 1, while if $p \ll n^{-2/3}$, then $\mathbb{P}[C_4]$ will be close to 0. Thus $p = n^{-2/3}$ is a "threshold" for the appearance of a 4-clique.

Remark 2.1. Here we use the asymptotic notation $f(n) \gg g(n)$ to denote that $\lim_{n\to\infty} f(n)/g(n) = \infty$. Similarly, $f(n) \ll g(n)$ means that $\lim_{n\to\infty} f(n)/g(n) = 0$.

We can use a simple first moment calculation for one side of our desired threshold behavior.

Lemma 2.2. *If $p \ll n^{-2/3}$ then $\mathbb{P}[C_4] \to 0$ as $n \to \infty$.*

Proof. Let X denote the number of 4-cliques in $G \sim \mathcal{G}_{n,p}$. We can write $X = \sum_S X_S$ where the set S runs over all $\binom{n}{4}$ subsets of four vertices in G, and $X_S = 1$ if there is a 4-clique on S, and $X_S = 0$ otherwise.

We have $\mathbb{P}[X_S = 1] = p^6$ since all 6 edges must be present, thus by linearity of expectation $\mathbb{E}\,X = p^6\binom{n}{4}$. So if $p \ll n^{-2/3}$, then $\mathbb{E}\,X \to 0$ as $n \to \infty$. But now Markov's inequality implies that

$$\mathbb{P}[C_4] = \mathbb{P}[X \geqslant 1] \leqslant \mathbb{E}\,X \to 0\,. \qquad \square$$

On the other hand, proving that $p \gg n^{-2/3} \implies \mathbb{P}[C_4] \to 1$ is more delicate. Even though a first moment calculation implies that, in this case, $\mathbb{E}\,X \to \infty$, this is not enough to conclude that $\mathbb{P}[C_4] \to 1$. For instance, it could be the case that with probability $1 - \frac{1}{n^2}$, we have no 4-cliques, but with probability $\frac{1}{n^2}$ we see all possible $\binom{n}{4}$ 4-cliques. In that case, $\mathbb{E}\,X \asymp n^2$, but still the probability of seeing a 4-clique would be $\frac{1}{n^2}$.

We need to exploit the fact that the appearances of distinct 4-cliques are mostly independent events. Certainly if S and S' are two disjoint sets of vertices, then the corresponding random variables X_S and $X_{S'}$ are independent.

2.2 Chebyshev's inequality and second moments

First, we recall the notion of the *variance* of a real-valued random variable X: $\mathrm{Var}(X) = \mathbb{E}\,(X - \mathbb{E}[X])^2$. If we know something about the variance, we can improve upon Markov's inequality.

Tool 2.3 (Chebyshev inequality)**.** If X is a real-valued random variable with $\mu = \mathbb{E}\,X$, then for every $\alpha > 0$,

$$\mathbb{P}\left[|X - \mu| \geqslant \alpha\right] \leqslant \frac{\mathrm{Var}(X)}{\alpha^2}\,.$$

Proof. Simply apply Markov's inequality to the nonnegative random variable $(X - \mu)^2$. $\square$

One consequence of the Chebyshev inequality is the following simple fact.

Corollary 2.4. *If X is a real-valued random variable, then*

$$\mathbb{P}(X = 0) \leqslant \frac{\mathrm{Var}(X)}{(\mathbb{E}\, X)^2}\,.$$

Proof. Using the Chebyshev inequality, we have:

$$\mathbb{P}(X = 0) \leqslant \mathbb{P}\left(|X - \mathbb{E}\, X| \geqslant \mathbb{E}\, X\right) \leqslant \frac{\mathrm{Var}(X)}{(\mathbb{E}\, X)^2}\,. \qquad \square$$

Now we are in position to analyze the other side of the threshold. It will help to have the following definition: For two random variables X and Y, we define their *covariance* by $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$. Note that if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Lemma 2.5. *If $X = X_1 + \cdots + X_n$, then*

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{Cov}(X_i, X_j)\,.$$

Proof. One observe first that for any random variable X, we have $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}\, X)^2$, hence

$$\mathrm{Var}(X) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^2 - \left(\mathbb{E} \sum_{i=1}^{n} X_i\right)^2\,.$$

Expanding the squares and collecting terms yields the desired result. $\square$

Lemma 2.6. *If $p \gg n^{-2/3}$, then $\mathbb{P}[C_4] \to 1$ as $n \to \infty$*

Proof. As before, let $X = \sum_S X_S$ denote the number of 4-cliques in $G \sim \mathcal{G}_{n,p}$. Our goal is to show that $\mathbb{P}[X = 0] \to 0$ as $n \to \infty$. Using Corollary 2.4, it suffices to show that $\mathrm{Var}(X) \ll (\mathbb{E}\, X)^2$. The main task will be evaluating $\mathrm{Var}(X)$.

Write

$$\mathrm{Var}(X) = \sum_S \mathrm{Var}(X_S) + \sum_S \sum_{T \neq S} \mathrm{Cov}(X_S, X_T)\,.$$

First, observe that since X_S is a $\{0, 1\}$ random variable, we have $\mathrm{Var}(X_S) \leqslant \mathbb{E}[X_S^2] = \mathbb{E}[X_S]$, yielding

$$\mathrm{Var}(X) \leqslant \mathbb{E}[X] + \sum_S \sum_{T \neq S} \mathrm{Cov}(X_S, X_T)\,. \tag{2.1}$$

Now we evaluate the second sum. The value $\mathrm{Cov}(X_S, X_T)$ depends on $|S \cap T|$. In particular, if $|S \cap T| \leqslant 1$, then $\mathrm{Cov}(X_S, X_T) = 0$ since S and T share no possible edges, hence the events X_S and X_T are independent.

Next, note that $\mathrm{Cov}(X_S, X_T) \leqslant \mathbb{E}[X_S X_T]$, and the latter event is that we have a 4-clique on *both* S and T. Thus if $|S \cap T| = 2$, then $\mathrm{Cov}(X_S, X_T) \leqslant p^{11}$ (since for $X_S = X_T = 1$ to happen, we need 11 edges to be present). Similarly, if $|S \cap T| = 3$, then $\mathrm{Cov}(X_S, X_T) \leqslant p^9$. So now we are simply left to count the possibilities:

$$\sum_S \sum_{T \neq S} \mathrm{Cov}(X_S, X_T) \leqslant O(n^6)p^{11} + O(n^5)p^9\,.$$

(Make sure you understand why this line is true!) From (2.1), we conclude that

$$\mathrm{Var}(X) \leqslant O(n^4)p^6 + O(n^6)p^{11} + O(n^5)p^9\,.$$

Also, recall that $\mathbb{E}\, X = \binom{n}{4}p^6 = \Theta(n^4 p^6)$, so $(\mathbb{E}\, X)^2 = \Theta(n^8 p^{12})$. In particular, if $p \ll n^{-2/3}$, then $\mathrm{Var}(X) \ll (\mathbb{E}\, X)^2$. Combined with Corollary 2.4, this yields the desired result. $\square$

2.3 Unbiased estimators

Suppose we want to estimate the area of a unit disk in the plane. One way to accomplish this is via a *Monte Carlo* algorithm: We sample a uniformly random point in $x \in [0,1]^2$ in the unit square, and then check whether $x_1^2 + x_2^2 \leqslant 1$. Let X be the indicator of the event that x is in the unit circle. Then

$$\mathbb{E}[X] = \frac{\text{area(disk)}}{\text{area}([0,1]^2)} = \frac{\pi}{4}\,.$$

So the random variable X is an unbiased estimator for $\pi/4$.

There are many situations in which one wants to estimate the probability of some event, but doing so directly is difficult. In those cases, having an unbiased estimator is quite helpful. Crucially, the usefulness of such an estimator depends on its variance. First, let's recall the following simple fact (you should be able to prove it easily using Lemma 2.5).

Fact 2.7. *If $X_1, X_2, \ldots, X_n$ are independent, then*

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\,.$$

We recall that a family of random variables is said to be "i.i.d." if they are "identical and independently distributed," i.e. they are independent samples from the same distribution.

Theorem 2.8. *For every $\varepsilon > 0$, the following holds. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $X = \frac{X_1+X_2+\cdots+X_n}{n}$ be their* empirical mean. *If $n \geqslant \frac{4}{\varepsilon^2}\frac{\sigma^2}{\mu^2}$, then*

$$\mathbb{P}\left[|X - \mu| \geqslant \varepsilon\mu\right] \leqslant \frac{1}{4}\,.$$

Proof. By linearity, $\mathbb{E}[X] = \mu$ and by independence and Fact 2.7, $\mathrm{Var}(X) = \sigma^2/n$. So Chebyshev's inequality implies that

$$\mathbb{P}\left[|X - \mu| \geqslant \varepsilon\mu\right] \leqslant \frac{\sigma^2}{n\mu^2\varepsilon^2} \leqslant \frac{1}{4}\,,$$

where the final bound uses our assumption on n. □

Observe that if each X_i is a $\{0,1\}$ random variable, then $\mathrm{Var}(X_i) = \mu - \mu^2$, so $\sigma^2 \leqslant \mu$. In this case, the required number of samples simplifies to $n \geqslant \frac{4}{\varepsilon^2} \cdot \frac{1}{\mu}$. This represents the intuitive fact that if we are trying to estimate probability of a very rare event (so that μ is very small), then we will need many samples to get a decent estimate.

The median trick. Although we stopped at $\frac{1}{4}$, we can improve the probability of a poor estimate with a small additional expense. Suppose that we do N trials of the above experiment (requiring $N \cdot n$ samples overall). Let X' be the median value of these N trials. Then as long as $N > 2\log_{4/3}\frac{1}{\delta}$, we will have $\mathbb{P}[|X' - \mu| \geqslant \varepsilon\mu] \leqslant \delta$.

To see this, let $Y_1, \ldots, Y_N$ be indicators that are equal to 1 if the ith trial gives an empirical mean in the range $[\mu(1-\varepsilon), \mu(1+\varepsilon)]$. We know from the preceding lemma that $\mathbb{P}[Y_i = 1] \geqslant 3/4$, and the variables $\{Y_i\}$ are independent. The only way that $X' \notin [\mu(1-\varepsilon), \mu(1+\varepsilon)]$ is if at least half the trials end negatively, i.e. $Y_1 + \cdots + Y_N \leqslant N/2$.

Claim 2.9. *If we perform $2k+1$ coin flips and $\mathbb{P}[\text{heads}] \geqslant 3/4$, then $\mathbb{P}[\leqslant k$ flips are heads$] \leqslant (3/4)^k$.*

One can prove this directly by counting and estimating some binomial coefficients. In Lecture 5, we will see how to prove this and related results in a more general way, and that will start our foray into the *concentration of measure* phenomenon.

2.4 Percolation on a tree

In the last lecture, we used the fact that for a real-valued random variable X, we have

$$\mathbb{P}(X = 0) \leqslant \frac{\mathrm{Var}(X)}{(\mathbb{E}\, X)^2}\,.$$

Let us see that second moments can also control the probability that a non-negative random variable is non-zero. (A generalization of this inequality often goes by the name "Payley-Zygmund inequality.")

Lemma 2.10. *If X is a non-negative random variable, then*

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

Proof. The proof relies on the Cauchy-Schwarz inequality: For any two real-valued random variables Y and Z,

$$\mathbb{E}\,|YZ| \leqslant \sqrt{\mathbb{E}[Y^2]}\sqrt{\mathbb{E}[Z^2]}\,.$$

We will use $Y = X$ and $Z = \mathbf{1}_{\{X>0\}}$, i.e. Z is the indicator of the event that $X > 0$:

$$\mathbb{E}\, X = \mathbb{E}\left[X\mathbf{1}_{\{X>0\}}\right] \leqslant \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}\,\mathbf{1}_{\{X>0\}}} = \sqrt{\mathbb{E}[X^2]}\,\sqrt{\mathbb{P}(X > 0)}\,.$$

Rearranging and squaring both sides completes the proof. □

Let's now consider the following model of *percolation* on the complete (rooted) binary tree of depth n. Fix a parameter $p \in [0, 1]$ and also an orientation of the tree (so that we can refer to "left" and "right" children). Suppose that every left edge is independently deleted with probability $1 - p$ and every right edge is independently deleted with probability p. Let $X = \sum_\ell Z_\ell$ denote the number of leaves which can reach the root of the tree. Here, the sum is over all 2^n leaves ℓ and Z_ℓ is the indicator random variable that is 1 precisely when the path from the root to the leaf ℓ remains intact. Here we are interested in the probability that there exists at least one reachable leaf, i.e. $\mathbb{P}[X > 0]$.

The first moment. It is relatively easy to compute the expected value of X:

$$\mathbb{E}\, X = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = (p + 1 - p)^n = 1\,.$$

The summation variable i indexes the number of left turns that a root-leaf path makes. If a path makes i left turns, then the probability it remains intact is $p^i(1-p)^{n-i}$. Furthermore, we can specify a root-leaf path by a sequence of "left" and "right" turns; this implies that the number of leaves whose root-leaf path contains exactly i left turns is $\binom{n}{i}$.

This calculation doesn't tell us that there is a root-leaf with decent probability. Certainly if $p = 0$ or $p = 1$, then there is a leaf with probability one (the left-most and right-most paths, respectively). But what about intermediate values of p? What if $p = 1/2$?

The second moment. Let's now compute the second moment, but we'll be a bit more clever. Let X_n denote the number of reachable leaves in a tree of depth n. If X_L and X_R denote the number of reachable leaves under the left and right subtrees, then $X_n^2 = (X_L + X_R)^2 = X_L^2 + X_R^2 + 2X_L X_R$.

Observe that $\mathbb{E}[X_L^2] = p\,\mathbb{E}[X_{n-1}^2]$, $\mathbb{E}[X_R^2] = (1-p)\,\mathbb{E}[X_{n-1}^2]$, $\mathbb{E}[X_L] = p\,\mathbb{E}[X_{n-1}] = p$ and $\mathbb{E}[X_R] = (1-p)\,\mathbb{E}[X_{n-1}]$. Since X_L and X_R are independent, we have

$$\mathbb{E}[X_n^2] = \mathbb{E}[X_{n-1}^2] + 2p(1-p)\,.$$

Since $X_0 = 1$, we get $\mathbb{E}[X_n^2] = 1 + 2np(1-p)$.

Now applying Lemma 2.10 gives

$$\mathbb{P}(X > 0) \geqslant \frac{1}{1 + 2np(1-p)}\,.$$

Thus if $p = \Theta(1/n)$, there exists a reachable leaf with constant probability. If $p = 1/2$, we get a leaf with probability at least $2/(n+2)$.

Notice that this still leaves a number of interesting questions to be answered. We know that $\mathbb{P}(X > 0) \geqslant 2/(n+2)$, but $\mathbb{E}[X_n^2] = \Theta(n)$. Is it the case that we see $\Theta(\sqrt{n})$ reachable leaves with probability $\Theta(1/n)$?

[See video of the discrete torus being covered by random walk—the asymptotics of the cover time are determined by a related percolation process on a complete tree.]

2.5 Using unbiased estimators to count

Consider a formula in disjunctive normal form (a DNF formula), e.g.

$$\varphi = (X_1 \wedge X_2 \wedge \bar{X}_3) \vee (\bar{X}_2 \wedge X_5) \vee \cdots$$

We can easily determine whether such a formula is satisfiable (we just check whether each term separately is satisfiable). On the other hand, *counting* the number of satisfying assignments to such a formula is #P-hard. (That means the problem is at least as hard as an NP-complete problem; this class of decision problems is thought to be extremely difficult to solve.)

The reason for this hardness is easy to see. Suppose that φ is a CNF formula. Then de Morgan's laws can be used to write $\neg\varphi$ as a DNF formula in polynomial time. But if φ is a formula on n variables, then

$$\#\text{ satisfying assignments to } \varphi = 2^n - \big(\#\text{ satisfying assignments to } \neg\varphi\big)\,.$$

Nevertheless, we will show that one can obtain an efficient algorithm to approximate the number of satisfying assignments.

Theorem 2.11 (Karp and Luby, 1983)**.** *Given a DNF formula φ with n variables and m terms, and a number $\varepsilon > 0$, there is an algorithm running in time $O(mn/\varepsilon^2)$ that outputs a value Z such that*

$$\mathbb{P}\left[(1-\varepsilon)Z \leqslant N(\varphi) \leqslant (1+\varepsilon)Z\right] \geqslant \frac{3}{4}\,,$$

where $N(\varphi)$ is the number of satisfying assignments to φ.

As we saw in the last lecture, one can use the median trick to amplify the probability of correctness to be very close to 1. One should note that a naïve Monte Carlo algorithm will not work so well: If we choose one of the 2^n assignments uniformly at random and check whether it satisfies φ, then the number we are trying to estimate is $\mu = \frac{N(\varphi)}{2^n}$, but this could be exponentially small. The unbiased estimator theorem would dictate that we use exponentially many samples, making our algorithm quite inefficient.

Proof of Theorem 2.11. The main idea will be to apply the Monte Carlo method to a more suitable space. Let S_i be the set of assignments which satisfy the ith term in φ. Our goal is to compute the size of the union $\bigcup_{i=1}^m S_i$.

Let $\mathcal{U} = \{(a, i) : a \in S_i\}$. Say that the pair $(a, i) \in \mathcal{U}$ is *special* if i is the first term that the assignment a satisfies. Since every satisfying assignment has exactly one first term that it satisfies, we have the following.

Claim 2.12. *The number of special pairs is precisely* $\left|\bigcup_{i=1}^m S_i\right|$

Now we apply the Monte Carlo algorithm to estimate $|\mathcal{S}|/|\mathcal{U}|$ where $\mathcal{S}$ is the set of special pairs. Note that $|S_i| = 2^{n-q_i}$ where q_i is the number distinct variables in term i. Thus we can easily compute $|\mathcal{U}| = \sum_{i=1}^m |S_i|$. In particular, if we obtain a multiplicative approximation to $\mu = |\mathcal{S}|/|\mathcal{U}|$, then we can obtain a multiplicative approximation to $|\mathcal{S}| = N(\varphi)$.

To generate a random sample from $\mathcal{U}$, we choose i with probability $|S_i|/|\mathcal{U}|$, and then pick a random satisfying assignment for term i. Finally, this is extended to a uniformly random assignment on the remaining $n - q_i$ variables.

Thus we are left to estimate $\mu = |\mathcal{S}|/|\mathcal{U}|$. But it's easy to see that $|\mathcal{U}| \leqslant m|\mathcal{S}|$ since each satisfying assignment can satisfy at most m terms. Therefore $\mu \geqslant \frac{1}{m}$. The unbiased estimator theorem now states that we need at most $\frac{4}{\varepsilon^2 \mu} \leqslant \frac{4m}{\varepsilon^2}$ samples in order to achieve our goal.

A naïve implementation of the algorithm requires $O(nm^2/\varepsilon^2)$ time, but further improvements [Karp-Luby-Madras 1989] show how to implement the algorithm in $O(mn/\varepsilon^2)$ time. □

Remark 2.13. The algorithm can be used to approximate the size of the union of any collection of finite sets as long as we can compute their individual sizes and sample random elements from each of them.

A straightforward generalization of the algorithm allows us to compute the probability of a random assignment in a probabilistic DNF model where each variable is true independently with some probability p_i.

3 Chernoff bounds

We have seen how knowledge of the variance of a random variable X can be used to control deviation of X from its mean. This is the heart of the second moment method. But often we can control even higher moments, and this allows us to obtain much stronger concentration properties.

A prototypical example is when $X_1, X_2, \ldots, X_n$ is a family of independent (but not necessarily identically distributed) $\{0, 1\}$ random variables and $X = X_1 + X_2 + \cdots + X_n$. Let $p_i = \mathbb{E}[X_i]$ and define $\mu = \mathbb{E}[X] = p_1 + p_2 + \cdots + p_n$. In that case, we have the following multiplicative form of the "Chernoff bound."

Theorem 3.1 (Chernoff bound, mutiplicative error)**.** *For every $\beta \geqslant 1$, it holds that*

$$\mathbb{P}\left(X \geqslant \beta\mu\right) \leqslant \left(\frac{e^{\beta-1}}{\beta^{\beta}}\right)^{\mu}, \tag{3.1}$$

and

$$\mathbb{P}\left(X \leqslant \frac{\mu}{\beta}\right) \leqslant \left(\frac{e^{1/\beta-1}}{\beta^{\beta}}\right)^{\mu}. \tag{3.2}$$

It's easy to use these formulae, but it sometimes helps to employ the slightly weaker bounds: If we put $\beta = 1 + \delta$ for $0 < \delta < 1$,

$$\mathbb{P}\left(X \geqslant (1+\delta)\mu\right) \leqslant e^{-\frac{\delta^2\mu}{3}},$$
$$\mathbb{P}\left(X \leqslant (1-\delta)\mu\right) \leqslant e^{-\frac{\delta^2\mu}{2}}.$$

The main point is that these tail bounds go down *exponentially* in the mean μ and the multiplicative deviation β, as opposed to the previous tail bounds we've seen (Markov and Chebyshev) that only go down polynomially.

Proof of Theorem 3.5. Much as we proved Chebyshev's inequality by applying Markov's inequality to the random variable $|X - \mathbb{E}\,X|^2$, the Chernoff bound is proved by applying a function to the underlying random variable X and then applying Markov's inequality.

Let $t \geqslant 0$ be a parameter we will choose later and write

$$\mathbb{P}[X \geqslant \beta\mu] = \mathbb{P}\left[e^{tX} \geqslant e^{t\beta\mu}\right] \leqslant \frac{\mathbb{E}[e^{tX}]}{e^{t\beta\mu}}. \tag{3.3}$$

The point of applying the function $X \mapsto e^{tX}$ is that we can exploit independence:

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{t(X_1+\cdots+X_n)}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{tX_i}\right]. \tag{3.4}$$

Now write:

$$\mathbb{E}[e^{tX_i}] = (1-p_i) + p_i e^t = 1 + p_i(e^t - 1) \leqslant e^{p_i(e^t-1)},$$

where the last inequality uses $1 + x \leqslant e^x$ which is valid for all $x \in \mathbb{R}$.

Plugging this into (3.4) yields

$$\mathbb{E}\left[e^{tX}\right] \leqslant \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{\mu(e^t-1)}$$

Now recalling (3.3), we have

$$\mathbb{P}[X \geqslant \beta\mu] \leqslant e^{\mu(e^t-1-\beta t)}.$$

Choosing $t = \ln\beta$ yields (3.7). One can prove (3.8) similarly. □

3.1 Randomized rounding

A classical technique in the field of approximation algorithms is to write down a linear programming relaxation of a combinatorial problem. The linear program (LP) is then solved in polynomial time, and one *rounds* the fractional solution to an integral solution that is, hopefully, not too much worse than the optimal solution. A classical example goes back to Raghavan and Thompson.

Let $D = (V, A)$ be a directed network, and suppose that we are given a sequence of *terminal pairs* $(s_1, t_1), (s_2, t_2), \ldots, (s_k, t_k)$ where $\{s_i\}, \{t_i\} \subseteq V$. We use $\mathfrak{I} = (D, \{(s_i, t_i)\})$ denote this instance of the *min-congestion disjoint paths problem*. The goal is to choose, for every i, a directed s_i-t_i path γ_i in D so as to minimize the maximum *congestion* of an arc $e \in A$:

$$\mathrm{opt}(\mathfrak{I}) = \text{minimize} \left\{ \max_{e \in A} \#\{i : e \in \gamma_i\} \right\} .$$

This problem is NP-hard. Our goal will be an *approximation algorithm* that outputs a solution $\{\gamma_i\}$ so that the congestion of every edge is at most $\alpha \cdot \mathrm{opt}$. The number α is called the *approximation factor* of our algorithm.

A fractional relaxation. Our approach will be to compute first a fractional solution that sends 1 unit of flow from s_i to t_i for every i. A flow can be thought of in the following way. Let $\mathcal{P}_i$ denote the set of simple, directed s_i-t_i paths in D, and let $\mathcal{P}$ denote the set of all simple directed paths in D. Here, simple means that no arc is repeated.

A *multi-flow* F is a mapping $F : \mathcal{P} \to \mathbb{R}_+$ of paths to nonnegative real numbers. The multi-flow F *routes the demands* $\{(s_i, t_i)\}$ if, for every $i = 1, \ldots, k$, we have $\sum_{\gamma \in \mathcal{P}_i} F(\gamma) = 1$, i.e. we send at least one unit of flow from s_i to t_i for every i. Finally, the *congestion* of an arc $e \in A$ under the flow F is the value $\mathrm{con}_F(e) = \sum_{\gamma \in \mathcal{P} : e \in \gamma} F(\gamma)$, i.e. the amount of flow passing through the arc e.

We make the definition:

$$\mathrm{LP}(\mathfrak{I}) = \text{minimize}_F \left\{ \max_{e \in A} \mathrm{con}_F(e) \right\} ,$$

where the minimum is over all multi-flows F that route the demands $\{(s_i, t_i)\}$. It should be clear that $\mathrm{LP}(\mathfrak{I}) \leqslant \mathrm{opt}(\mathfrak{I})$. The reason we write $\mathrm{LP}(\mathfrak{I})$ is that this value can be computed by a linear program of polynomial-size. This is not precisely clear from our formulation because there are possibly exponentially many paths in $\mathcal{P}$, but there is a compact formulation of the LP using standard techniques (see the remark at the end of this section).

Given a multi-flow F, we will round it to an *integral* multi-flow F', where an integral flow is one such that, for every $i = 1, \ldots, k$, we have $F'(\gamma) = 1$ for *exactly* one $\gamma \in \mathcal{P}_i$. Note that an integral flow represents a solution to the initial disjoint paths problem. Furthermore, we will now show that for some $\alpha \geqslant 1$, we have

$$\max_{e \in A} \mathrm{con}_{F'}(e) \leqslant \alpha \cdot (1 + \max_{e \in A} \mathrm{con}_F(e)) .$$

In particular, if we apply this to the optimal fractional flow F^*, we arrive at

$$\max_{e \in A} \mathrm{con}_F(e) \leqslant \alpha \cdot (1 + \max_{e \in A} \mathrm{con}_{F^*}(e)) = \alpha(1 + \mathrm{LP}(\mathfrak{I})) \leqslant \alpha(1 + \mathrm{opt}(\mathfrak{I})) \leqslant 2\alpha \cdot \mathrm{opt}(\mathfrak{I}) ,$$

implying that we have achieved an 2α-approximation to the optimal solution. (Note that we have used the trivial bound $\mathrm{opt} \geqslant 1$.)

Theorem 3.2. *Let $n = |V|$ and suppose that $n \geqslant 4$. If there is a multi-flow F that routes the demands $(s_1, t_1), \ldots, (s_k, t_k)$, then there exists an integral multi-flow F' that routes the demands, and furthermore*

$$\max_{e \in A} F'(e) \leqslant C \frac{\log n}{\log \log n} \left(1 + \max_{e \in A} F(e)\right), \tag{3.5}$$

where $C > 0$ is a universal constant.

Proof. We will produce a *random* integral multi-flow F' that routes the demands $\{(s_i, t_i)\}$ and argue that it satisfies the conditions of the theorem with high probability.

For every $i = 1, \ldots, k$, we do the following independently. We know that $\sum_{\gamma \in \mathcal{P}_i} F(\gamma) = 1$. Thus we can think of F as providing a probability distribution over s_i-t_i paths. We let γ_i denote a random s_i-t_i path chosen with probability $F(\gamma)$ for $\gamma \in \mathcal{P}_i$.

The set of paths $\{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ gives us an integral multi-flow F'. We are left to bound the maximum congestion of an edge. To this end, fix an edge $e \in A$. For every $\gamma \in \mathcal{P}$ such that $e \in \gamma$, let X_γ be the indicator random variable that is 1 when the path γ is chosen in the rounding. Then the number of edges going through the edge e after rounding is given by the random variable

$$\mathrm{con}_{F'}(e) = \sum_{\gamma : e \in \gamma} X_\gamma \,. \tag{3.6}$$

We may assume that $\mathrm{con}_F(e) \geqslant 1$ because we are comparing $\mathrm{con}_{F'}(e)$ to $1 + \mathrm{con}_F(e)$ in (3.5).

First, we have

$$\mathbb{E}[\mathrm{con}_{F'}(e)] = \sum_{\gamma : e \in \gamma} \mathbb{E}[X_\gamma] = \sum_{\gamma : e \in \gamma} F(\gamma) = \mathrm{con}_F(e) \,.$$

So at least in expectation, the congestion does not increase. If the $\{X_\gamma\}$ random variables were independent, then we could apply the Chernoff bound. Unfortunately, this not necessarily the case. For instance, if $\gamma, \gamma' \in \mathcal{P}_i$ both contain the edge e, then X_γ and $X_{\gamma'}$ are not independent; in fact, at most one of them can be equal to 1. Thus we will first rewrite $\mathrm{con}_{F'}(e)$ as a sum of independent $\{0, 1\}$ random variables.

Let Y_i be the indicator variable that equals 1 if the unique s_i-t_i path in F' uses the edge e, i.e. if $e \in \gamma_i$. Then the $\{Y_i\}$ are independent (since we round each s_i-t_i pair independently). Moreover, we have $Y_i = \sum_{\gamma \in \mathcal{P}_i : e \in \gamma} X_\gamma$, so $\mathrm{con}_{F'}(e) = \sum_{i=1}^k Y_i$.

Since $\mathrm{con}_{F'}(e)$ is a sum of independent $\{0, 1\}$ random variables, we can apply the Chernoff bound (Theorem 3.5) to conclude that

$$\mathbb{P}\left[\mathrm{con}_{F'}(e) \geqslant \beta \cdot \mathrm{con}_F(e)\right] \leqslant \left(\frac{e^{\beta - 1}}{\beta^\beta}\right)^{\mathrm{con}_F(e)} \leqslant \frac{e^{\beta - 1}}{\beta^\beta} \,,$$

where in the last inequality we have used our assumption that $\mathrm{con}_F(e) \geqslant 1$. We would like to choose the latter bound to be at most n^{-3}. To do this, we need to choose $\beta = C \frac{\log n}{\log \log n}$ for some constant C. (You should check that this is the right choice of β.)

Setting β like this, we have

$$\mathbb{P}\left[\mathrm{con}_{F'}(e) \geqslant \beta \cdot \mathrm{con}_F(e)\right] \leqslant \frac{1}{n^3} \,,$$

and thus by a union bound over the n^2 possible edges,

$$\mathbb{P}\left[\exists e \in A \text{ such that } \mathrm{con}_{F'}(e) \geqslant \beta \cdot \mathrm{con}_F(e)\right] \leqslant n^2 \cdot \frac{1}{n^3} \leqslant \frac{1}{n} \,.$$

Thus with probability at least $1 - \frac{1}{n}$, our integral flow F' satisfies the claim of the theorem. □

Remark 3.3. Note that if we knew $\mathrm{con}_F(e) \geqslant C' \log n$ for some constant C' and every $e \in A$, then we could actually choose $\beta = O(1)$ and still achieve a bound of n^{-3} on the probability of an over congested edge. This means that if all the fractional congestions are $\Omega(\log n)$, we can get an $O(1)$ approximation.

Remark 3.4. To compute the optimal fractional multi-flow, we write a linear program with variables $\{F_e : e \in A\}$. Our program should minimize the value λ such that $F_e \leqslant \lambda$ for every $e \in A$. Moreover, to make sure that the variables $\{F_e : e \in A\}$ correspond to an optimal flow, we should add the flow constraints at every non-terminal vertex: The flow in should be equal to the flow out of the vertex. At terminals, we have to allow there to be a surplus or deficit based on whether we are at a source or a sink. This program has $O(m)$ variables and $O(m + n)$ linear constraints, where $m = |A|$ and $n = |V|$.

3.2 Some more applications

In the preceding lecture, we saw our first large-deviation inequality. Let $X_1, X_2, \ldots, X_n$ be a family of independent (but not necessarily identically distributed) $\{0, 1\}$ random variables and $X = X_1 + X_2 + \cdots + X_n$. Let $p_i = \mathbb{E}[X_i]$ and define $\mu = \mathbb{E}[X] = p_1 + p_2 + \cdots + p_n$. We recall the following multiplicative form of the "Chernoff bound."

Theorem 3.5 (Chernoff bound, mutiplicative error)**.** *For every $\beta \geqslant 1$, it holds that*

$$\mathbb{P}\left(X \geqslant \beta\mu\right) \leqslant \left(\frac{e^{\beta-1}}{\beta^\beta}\right)^\mu, \tag{3.7}$$

and

$$\mathbb{P}\left(X \leqslant \frac{\mu}{\beta}\right) \leqslant \left(\frac{e^{1/\beta-1}}{\beta^\beta}\right)^\mu. \tag{3.8}$$

3.2.1 Balls in bins

Suppose we throw m balls uniformly at random into n bins. For $i = 1, \ldots, n$, let $X^{(i)}$ denote the number of balls that land in bin i. Let $Z := \max(X^{(1)}, \ldots, X^{(n)})$ denote the maximum load. Even to bound $\mathbb{E}[Z]$ seems tricky, and it is often the case that evaluating the expected maximum of a family of random variables requires understanding their tail behavior.

Let $X_j^{(i)}$ be the indicator random variable that is 1 if ball j lands in bin i. Then $\mathbb{E}[X_j^{(i)}] = 1/n$ and hence by linearity of expectation, $\mathbb{E}[X^{(i)}] = m/n$. Applying Theorem 3.5 yields, for any $i = 1, \ldots, n$,

$$\mathbb{P}\left[X^{(i)} \geqslant \beta\frac{m}{n}\right] \leqslant \left(\frac{e}{\beta}\right)^{\beta m/n}. \tag{3.9}$$

Let's consider two representative regimes.

Regime I: $m = n$**.** By choosing $\beta = \frac{c \log n}{\log \log n}$ for $c > 1$ large enough (as in the preceding lecture), (3.9) gives

$$\mathbb{P}\left[X^{(i)} \geqslant \beta\right] \leqslant n^{-2}$$

Now the deviation probability is small enough to apply a union bound:

$$\mathbb{P}[Z \geqslant \beta] \leqslant \sum_{i=1}^{n} \mathbb{P}[X^{(i)} \geqslant \beta] \leqslant n \cdot n^{-3} = \frac{1}{n}.$$

Thus with high probability, the maximum load is $O(\frac{\log n}{\log \log n})$.

Regime II: $m \geqslant cn \log n$**,** $c > 0$**.** In this case, we have $m/n \geqslant c \log n$, so applying (3.9) and a union bound gives

$$\mathbb{P}[Z \geqslant \beta] \leqslant n \cdot \beta^{-c\beta \log n}.$$

Now choosing $\beta \asymp 1/c$ gives $\mathbb{P}[Z \geqslant \beta] \leqslant \frac{1}{n}$, implying that the maximum load is only an $O(1)$ factor more than its expectation.

In comparing the two regimes, note that as the expected number of balls per bin rises, we actually get more concentration around the mean.

3.2.2 Randomized Quicksort

I don't particularly like this application since it requires some unnatural machinations to get independent random variables. In the next lecture, we will see that large-deviation inequalities hold for martingales, and this argument becomes more natural.

But a quick recap: Consider the numbers $\{1, 2, \ldots, n\}$. We construct a random rooted tree $\boldsymbol{T}$ where each node $v \in V(\boldsymbol{T})$ has an associated subset $S_v \subseteq \{1, \ldots, n\}$ defined as follows: For the root $r \in V(\boldsymbol{T})$, we have $S_r = \{1, \ldots, n\}$. Then inductively, for a node v with $|S_v| > 1$, we partition S_v uniformly at random into two sets S_v^L, S_v^R with $|S_v^L|, |S_v^R| \geqslant 1$, and we give v two children labeled by these sets. Thus $\boldsymbol{T}$ has precisely n leaves labeled by the singleton sets $\{1\}, \ldots, \{n\}$.

Let D_i denote the depth of the leaf labeled by $\{i\}$. The following claim is straightforward to verify inductively: The number of comparisons made by Quicksort is precisely $D_1 + D_2 + \cdots + D_n$. (Strictly speaking, this is only true because we are including the pivot in one of the two child lists, but a more clever implementation would only do better.)

Claim 3.6. *There is a constant $C \geqslant 1$ such that for any $i \in \{1, 2, \ldots, n\}$, it holds that*

$$\mathbb{P}[D_i \geqslant C \log n] \leqslant n^{-2}.$$

Taking a union bound gives

$$\mathbb{P}[\#\text{ comparisons} > Cn \log n] \leqslant \frac{1}{n},$$

i.e. Quicksort runs in $O(n \log n)$ time with high probability.

Fix an element $i \in \{1, \ldots, n\}$. Let $S_0, S_1, \ldots, S_{D_i} = \{i\}$ be the labels of the nodes occuring from the root down to the leaf labeled $\{i\}$ in $\boldsymbol{T}$. Define $S_j := \{i\}$ for $j > D_i$. Then an elementary calculation gives

$$\mathbb{P}\left[|S_{j+1}| \leqslant \frac{1}{2}|S_j|\right] \geqslant \frac{1}{2} \qquad \forall j \geqslant 0.$$

Define:

$$Y_j = \begin{cases} 1 & \text{if } |S_{j+1}| \leqslant \frac{1}{2}|S_j| \text{ or } S_{j+1} = S_j \\ 0 & \text{otherwise.} \end{cases}$$

The next claim is straightforward to verify.

Claim 3.7. *For every $j \geqslant 0$, we have $\mathbb{P}[Y_j = 1] \geqslant \frac{1}{2}$. Moreover, if $\sum_{j=0}^{M-1} Y_j \geqslant \log_2 n$, then $D_i \leqslant M$.*

If $\{Y_j\}$ were independent random variables, then we could apply (3.8) to conclude that

$$\mathbb{P}\left[Y_0 + Y_1 + \cdots + Y_{M-1} \leqslant \frac{M}{2\beta}\right] \leqslant \left(\frac{e^{1/\beta}}{\beta}\right)^{\beta M/2}$$

Choosing $M = \Theta(\log n)$ and $\beta = \Theta(1)$ would then yield Claim 3.6.

Unfortunately, these random variables are not independent, as clearly $Y_j = 1$ for $j \geqslant D_i$, for instance. One solution is to use a hack: We can couple the $\{Y_j\}$ random variables to *independent* random variables $\{\tilde{Y}_j\}$ such that $\tilde{Y}_j = 1 \implies Y_j = 1$ and $\mathbb{P}[\tilde{Y}_j = 1] = 1/2$. Then we can legitimately apply the Chernoff bound to the family $\{\tilde{Y}_j\}$ and reach the same conclusion.

This is easy to do by defining

$$\tilde{Y}_j = Z_j Y_j \,,$$

where $\{Z_j\}$ is a collection of $\{0,1\}$ random variables such that $Z_j = 1$ with probability $1/(2\,\mathbb{P}[Y_j = 1])$ (so that Z_j is independent of Y_j conditioned on S_j). Note that this definition makes sense since $\mathbb{P}[Y_j = 1] \geqslant 1/2$.

Now we have $\mathbb{P}[\tilde{Y}_j = 1] = 1/2$ for all $j \geqslant 0$, and moreover, the random variables $\{\tilde{Y}_j\}$ are independent:

$$\mathbb{P}\left[\tilde{Y}_j = 1 \mid \tilde{Y}_0, \ldots, \tilde{Y}_{j-1}, \tilde{Y}_{j+1}, \ldots\right] = \frac{1}{2} \qquad \forall j \geqslant 0 \,.$$

Even this preceding fact is slightly tricky to verify and the whole argument doesn't reflect one's natural intuition: Independence shouldn't matter as long as we have probability at least $1/2$ to reduce the size of S_{j+1} conditioned on S_j with $|S_j| > 1$. That is the purview of *martingale theory*, and we will cover large-deviation inequalities for martingales in the next lecture.

3.3 Negative correlation

Say that a collection $\{X_1, \ldots, X_n\}$ of random variables are *negatively correlated* if it holds that for any subset $S \subseteq [n]$:

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] \leqslant \prod_{i=1}^{n} \mathbb{E}[X_i] \,. \tag{3.10}$$

Note that if $\{X_1, \ldots, X_n\}$ are independent, then this holds with equality.

Examples. We will state some examples of negatively correlated families (without proof).

1. **Loads of the bins.** The variables $\{X^{(i)} : i = 1, \ldots, n\}$ from Section 3.2.1 are negatively correlated.

   This stands to reason: Telling you that some bins have unusually large (resp., small) load makes the expected load of the remaining bins smaller (resp., larger).

2. **Random permutations.** If $\{X_1, \ldots, X_n\} = \{1, 2, \ldots, n\}$, then the family $\{X_i\}$ is negatively correlated.

   The intuition here is the same as the previous example.

3. **Random spanning trees.** Suppose $G = (V, E)$ is an undirected graph and $\boldsymbol{T}$ is a uniformly random spanning tree of G. For $e \in E$, let X_e denote the indicator random variable that is 1 precisely when e is an edge of $\boldsymbol{T}$. Then the family $\{X_e : e \in E\}$ is negatively correlated.

   Suppose I tell you that $X_{e_1} = \cdots = X_{e_k} = 1$ for some edges $e_1, \ldots, e_k \in E$. Intuitively, one can contract the connected components in the graph spanned by $\{e_1, \ldots, e_k\}$ and consider a uniformly random spanning tree on the rest. Now the problem of connecting everything together has become easier, and thus $\mathbb{P}[X_e]$ decreases for $e \notin \{e_1, \ldots, e_k\}$. Certainly if e

connects two vertices that are already connected in the graph with edges $\{e_1, \ldots, e_k\}$, then $\mathbb{P}[X_e = 1 \mid X_{e_1} = \cdots = X_{e_k} = 1] = 0$.

Actually proving negative correlation for this family is non-trivial.

It turns out that the Chernoff bounds Theorem 3.5 hold if we consider negatively correlated random variables. It is still an area of active research to determine good notions for "negative dependence" in general settings. In particular, the notion of of negative correlation above is unsuitable for many settings, especially because it can be hard to verify and does not satisfy natural closure properties (making it difficult to derive new negatively correlated families from old ones).

Theorem 3.8 (Chernoff for negatively correlated random variables)**.** *If, in the statement of Theorem 3.5, we only assume that $X_1, \ldots, X_n$ are negatively correlated $\{0,1\}$ random variables (instead of independent), then the conclusion still holds.*

Proof. To see this, note that the one place we used independence in the proof of Theorem 3.5 is in the calculation: When $X = X_1 + \cdots + X_n$,

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{tX_i}\right].$$

Let us see that the inequality $\mathbb{E}\left[e^{tX}\right] \leqslant \prod_{i=1}^{n} \mathbb{E}\left[e^{tX_i}\right]$ still holds when $\{X_1, \ldots, X_n\}$ are only assumed to be negatively correlated.

To this end, let $\{\tilde{X}_1, \ldots, \tilde{X}_n\}$ be *independent* $\{0,1\}$ random variables with $\mathbb{E}[\tilde{X}_i] = \mathbb{E}[X_i]$ for each $i = 1, \ldots, n$, and define $\tilde{X} := \tilde{X}_1 + \cdots + \tilde{X}_n$. For any nonnegative integer k,

$$\mathbb{E}[X^k] = \sum_{\alpha} \mathbb{E}[X_1^{\alpha_1} X_2^{\alpha_2} \cdots X_n^{\alpha_n}],$$

where the sum is over all $\alpha \in \mathbb{R}^n$ with $\alpha_i \geqslant 0$ and $\sum_i \alpha_i = k$. Using the negative correlation property, this gives

$$\mathbb{E}[X^k] \leqslant \sum_{\alpha} \mathbb{E}[X_1^{\alpha_1}]\,\mathbb{E}[X_2^{\alpha_2}] \cdots \mathbb{E}[X_n^{\alpha_n}] = \sum_{\alpha} \mathbb{E}[\tilde{X}_1^{\alpha_1}]\,\mathbb{E}[\tilde{X}_2^{\alpha_2}] \cdots \mathbb{E}[\tilde{X}_n^{\alpha_n}],$$

where the last line follows because X_i and $\tilde{X}_i$ have the same distribution for every i. Finally, note that by independence,

$$\sum_{\alpha} \mathbb{E}[\tilde{X}_1^{\alpha_1}]\,\mathbb{E}[\tilde{X}_2^{\alpha_2}] \cdots \mathbb{E}[\tilde{X}_n^{\alpha_n}] = \sum_{\alpha} \mathbb{E}[\tilde{X}_1^{\alpha_1} \tilde{X}_2^{\alpha_2} \cdots \tilde{X}_n^{\alpha_n}] = \mathbb{E}[\tilde{X}^k].$$

We conclude that for every integer $k \geqslant 0$,

$$\mathbb{E}[X^k] \leqslant \mathbb{E}[\tilde{X}^k]. \tag{3.11}$$

Using the Taylor expansion

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2} + \frac{t^3X^3}{6} + \cdots,$$

and applying (3.11) to each term gives

$$\mathbb{E}[e^{tX}] \leqslant \mathbb{E}[e^{t\tilde{X}}] = \prod_{i=1}^{n} \mathbb{E}[e^{t\tilde{X}_i}] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}],$$

yielding our desired inequality. Now the proof of the Chernoff bound can proceed exactly as in the preceding lecture. □

4 Martingales

We have seen that if $X = X_1 + \cdots + X_n$ is a sum of independent $\{0, 1\}$ random variables, then X is tightly concentrated around its expected value $\mathbb{E}[X]$. The fact that the random variables were $\{0, 1\}$-valued was not essential; similar concentration results hold if we simply assume that they are in some bounded range $[-L, L]$. One can also relax the independence assumption, as we will see next.

Consider a sequence of random variables $X_0, X_1, X_2, \ldots$. The sequence $\{X_i\}$ is called a *discrete-time martingale* if it holds that

$$\mathbb{E}[X_{i+1} \mid X_0, X_1, \ldots, X_i] = X_i$$

for every $i = 0, 1, 2, \ldots$. More generally, the sequence is $\{X_i\}$ is a martingale with respect to another sequence of random variables $\{Y_i\}$ if it for every i, it holds that

$$\mathbb{E}[X_{i+1} \mid Y_0, Y_1, \ldots, Y_i] = X_i \,.$$

Note that this is equivalent to $\mathbb{E}[X_{i+1} - X_i \mid Y_0, Y_1, \ldots, Y_i] = 0$. If one thinks of $\{Y_0, Y_1, \ldots, Y_i\}$ as all the "information" up to time i, then this says that the difference $X_{i+1} - X_i$ is unbiased conditioned on the past up to time i. Observe that for any i, we have

$$\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i \mid X_0, \ldots, X_{i-1}]] = \mathbb{E}[X_{i-1}] = \cdots = \mathbb{E}[X_0] \,.$$

Martingales form an extremely useful class of random processes that appear in a vast array of settings (e.g., finance, machine learning, information theory, statistical physics, etc.). The classic example is that of a Gambler whose bank roll is X_0. At each time, she chooses to play some game in the casino at some stakes. If we assume that every game is fair (that is, the expected utility from playing the game is 0), then the sequence $\{X_0, X_1, \ldots\}$ forms a martingale, where X_i is the amount of money she has at time i.

Remark 4.1. The correct level of generality at which to define martingales involves a filtration. Formally, this is an increasing sequence of σ-algebras on our measure space $(\Omega, \mu, \mathcal{F})$: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$. A sequence of random variables $\{X_i\}$ is a martingale with respect to the filtration $\{\mathcal{F}_i\}$ if $\mathbb{E}[X_{i+1} \mid \mathcal{F}_i] = X_i$ for every $i \geqslant 0$.

4.1 Doob martingales

One reason martingales are so powerful is that they model a situation where one gains progressively more information over time. Suppose that $\mathcal{U}$ is a set of objects, and $f : \mathcal{U} \to \mathbb{R}$. Let X be a random variable taking values in $\mathcal{U}$, and let $\{Y_i\}$ be another sequence of random variables. The associated *Doob martingale* is given by

$$X_i = \mathbb{E}[f(X) \mid Y_0, Y_1, \ldots, Y_i] \,.$$

In words, this is our "estimate" for the value of $f(X)$ given the information contained in $\{Y_0, \ldots, Y_i\}$. To see that this is always a martingale with respect to $\{Y_i\}$, observe that

$$\mathbb{E}[X_{i+1} \mid Y_0, \ldots, Y_i] = \mathbb{E}[\mathbb{E}[f(X) \mid Y_0, \ldots, Y_{i+1}] \mid Y_0, \ldots, Y_i] = \mathbb{E}[f(X) \mid Y_0, \ldots, Y_i] = X_i \,,$$

where we have used the tower rule of conditional expectations.

Balls in bins. Suppose we throw m balls into n bins one at a time. At step i, we place ball i in a uniformly random bin. Let $C_1, C_2, \ldots, C_m$ be the sequence of (random) choices, and let C denote the final configuration of the system, i.e. exactly which balls end up in which bins.

Now we can consider a functional like $f(C) = \#$ of empty bins. If $X_i = \mathbb{E}[f(C) \mid C_1, \ldots, C_i]$, then $\{X_i\}$ is a (Doob) martingale. It is straightforward to calculate that

$$\mathbb{E}[X_m] = \mathbb{E}[X_0] = \mathbb{E}[f(C)] = n \cdot \left(1 - \frac{1}{n}\right)^m .$$

Suppose we are interested the concentration of $X_m = f(C)$ around its mean value. Of course, we can write $X_m = Z_1 + \cdots Z_m$ where Z_i is the indicator of whether the ith been is empty after all the balls have been thrown. But note that the $\{Z_i\}$ variables are not independent—in particular, if I tell you that $Z_1 = 1$ (bin 1 is empty), it decreases slightly the likelihood that other bins are empty.

The vertex exposure filtration. Recall that $\mathcal{G}_{n,p}$ denotes the random graph model where an undirected graph on n vertices is chosen by including every edge independently with probability p. Suppose the vertices are numbered $\{1, 2, \ldots, n\}$ Let $G \sim \mathcal{G}_{n,p}$ and denote by G_i the induced subgraph on the vertices $\{1, \ldots, i\}$. G_0 denotes the empty graph.

Let $\chi(G)$ denote the chromatic number of G, and consider the Doob martingale

$$X_i = \mathbb{E}[\chi(G) \mid G_0, \ldots, G_i] .$$

If we wanted to understand concentration properties of $X_n = \chi(G)$, this seems even more daunting. The chromatic number is a very complicated parameter of a graph! Nevertheless, we will now see that martingale concentration inequalities allow us to achieve tight concentration using very limited information about a sequence of random variables.

4.2 The Hoeffding-Azuma inequality

Say that a martingale $\{X_i\}$ has L-bounded increments if

$$|X_{i+1} - X_i| \leqslant L$$

for all $i \geqslant 0$. (The preceding inequality is meant to hold with probability 1.)

Theorem 4.2. *For every $L > 0$, if $\{X_i\}$ is a martingale with L-bounded increments, then for every $\lambda > 0$ and $n \geqslant 0$, we have*

$$\mathbb{P}[X_n \geqslant X_0 + \lambda] \leqslant e^{-\frac{\lambda^2}{2L^2 n}}$$
$$\mathbb{P}[X_n \leqslant X_0 - \lambda] \leqslant e^{-\frac{\lambda^2}{2L^2 n}}$$

It's useful to note the following special case of the theorem.

Corollary 4.3. *Suppose that $Z_1, Z_2, \ldots, Z_n$ are independent random variables taking values in the interval $[-L, L]$. Put $Z = Z_1 + \cdots + Z_n$ and $\mu = \mathbb{E}[Z]$. Then for every $\lambda > 0$, we have*

$$\mathbb{P}\left[Z \geqslant \mu + \lambda\right] \leqslant e^{-\lambda^2/(2L^2 n)}$$
$$\mathbb{P}\left[Z \leqslant \mu - \lambda\right] \leqslant e^{-\lambda^2/(2L^2 n)}$$

The Lispchitz condition. Recall the setting of Doob martingales, where $\mathcal{U}$ is a set. Suppose that we can describe every element $u \in \mathcal{U}$ by a sequence of values $u = (u_1, u_2, \ldots, u_n)$. (For instance, every configuration of m balls in n bins can be described by the sequence of which balls go into which bins.)

Say that f is *L-Lipschitz* if it holds that for every $i = 1, \ldots, n$ and for every two elements $u = (u_1, u_2, \ldots, u_i, \ldots, u_n) \in \mathcal{U}$ and $u' = (u_1, u_2, \ldots, u'_i, \ldots, u_n) \in \mathcal{U}$ that differ only in the ith coordinate, we have

$$|f(u) - f(u')| \leqslant L\,.$$

Let $Z = (Z_1, \ldots, Z_n)$ be a $\mathcal{U}$-valued random variable such that the random variables $\{Z_i\}$ are independent. We now confirm that the Doob martingale $X_i = \mathbb{E}[f(Z) \mid Z_1, \ldots, Z_i]$ has L-bounded increments.

Let Z'_{i+1} be an independent copy of Z_{i+1} conditioned on $Z_1, \ldots, Z_i$, and let $Z' = (Z_1, \ldots, Z_i, Z'_{i+1}, \ldots, Z_n)$. Then:

$$\begin{aligned}
|X_{i+1} - X_i| &= \left|\mathbb{E}[f(Z) \mid Z_1, \ldots, Z_{i+1}] - \mathbb{E}[f(Z) \mid Z_1, \ldots, Z_i]\right| \\
&= \left|\mathbb{E}\left[f(Z) - f(Z') \mid Z_1, \ldots, Z_{i+1}\right]\right| \\
&\leqslant \mathbb{E}\left[\left|f(Z) - f(Z')\right| \mid Z_1, \ldots, Z_{i+1}\right] \\
&\leqslant L\,,
\end{aligned}$$

where in the last step we have used the fact that the term inside the absolute value signs is always at most L by the L-Lipschitz property of f, and the fact that Z and Z' differ in at most one coordinate.

Remark 4.4. Note the power of Theorem 4.2 combined with this construction of Doob marginales. It means that if we have any random variable $Z = (Z_1, \ldots, Z_n)$ that is built out of independent pieces of information $\{Z_i\}$ and some quantity $f(Z)$ that we care about does not depend too much on changing any single piece of information, then $f(Z)$ is tightly concentrated about its mean. This is a vast generalization of the fact that sums of independent, bounded random variables are highly concentrated (cf. Corollary 4.3).

The number of empty bins. First let's apply this to balls and bins. Recall that for a sequence of choices $C_1, \ldots, C_m$ (where C_i is the bin that the ith ball is thrown into), we put $f(C_1, \ldots, C_m)$ to be the number of empty bins. Then clearly f is 1-Lipschitz: Changing the fate of ball i can only change the number of empty bins by 1. Therefore the corresponding martingale $X_i = \mathbb{E}[f(C_1, \ldots, C_m) \mid C_1, \ldots, C_i]$ has 1-bounded increments, and Azuma's inequality implies that

$$\mathbb{P}[X_n \geqslant X_0 + \lambda] \leqslant e^{-\frac{\lambda^2}{2m}}\,.$$

Recall that $X_0 = \mathbb{E}[X_n] = n(1 - \frac{1}{m})^n$. Consider the situation where $m = n$ and thus $X_0 \asymp \frac{n}{e}$. If we put $\lambda = C\sqrt{n}$, we see that with high probability we expect the number of empty bins to be in the interval $\frac{n}{e} \pm O(\sqrt{n})$.

The chromatic number. Similarly, consider the vertex exposure martingale. We have to be a little more careful here to describe a graph G by a sequence $(Z_1, \ldots, Z_n)$ of *independent* random variables. The key is to think about Z_i containing the information on edges from vertex i to the vertices $\{1, \ldots, i-1\}$ so that we have independence.

Since we can identify a graph G with the vector $(Z_1, \ldots, Z_n)$, we can think of the chromatic number as a function $\chi(Z_1, \ldots, Z_n)$. The function χ satisfies the 1-Lipschitz property because

changing the edges adjacent to some vertex i can only change the chromatic number by 1. The chromatic number cannot increase by more than one because we could always color i a new color; it cannot decrease by more than one because if we could color the graph without vertex i with c colors, then we can color the whole graph with $c+1$ colors.

So the martingale $X_i = \mathbb{E}[\chi(G) \mid Z_1, \dots, Z_i] = \mathbb{E}[\chi(G) \mid G_1, \dots, G_i]$ has 1-bounded increments and Azuma's inequality tells us that

$$\mathbb{P}[\chi(G) \geqslant \mathbb{E}[\chi(G)] + \lambda] \leqslant e^{-\frac{\lambda^2}{2n}}.$$

Even without having any idea how to compute $\mathbb{E}[\chi(G)]$, we are able to say something significant about its concentration properties.

Remark 4.5. By the way, if $G \sim \mathcal{G}_{n,1/2}$, then $\mathbb{E}[\chi(G)] = n/(2\log_2 n)$, so the concentration window here—which is $O(\sqrt{n})$—is again quite small with respect to the expectation. In the next lecture, we will see how a more clever use of Azuma's inequality can achieve even better concentration of $\chi(G)$.

4.3 Proof

We will actually prove the following generalization of Theorem 4.2.

Theorem 4.6. *Suppose that $\{X_i\}$ is a sequence of random variables satisfying the property that for every subset of distinct indices $i_1 < i_2 < \cdots < i_k$, we have*

$$\mathbb{E}\left[X_{i_1} X_{i_2} \cdots X_{i_k}\right] = 0.$$

Then for every $\lambda > 0$ and $n \geqslant 1$, it holds that

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geqslant \lambda\right] \leqslant \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^{n} \|X_i\|_\infty^2}\right).$$

Here, $\|X_i\|_\infty$ is the essential supremum of X_i, i.e. the least value L such that $|X_i| \leqslant L$ with probability one.

The reason Theorem 4.6 proves Theorem 4.2 is as follows: Suppose that $\{Z_i\}$ is a martingale with respect to the sequence of random variables $\{Y_i\}$, and let $X_i = Z_i - Z_{i-1}$. Consider distinct indices $i_1 < i_2 < \cdots < i_k$. Then:

$$\mathbb{E}[X_{i_1} \cdots X_{i_k}] = \mathbb{E}\left[X_{i_1} \cdots X_{i_{k-1}} \mathbb{E}\left[Z_{i_k} - Z_{i_k-1} \mid Y_0, \dots, Y_{i_k-1}\right]\right] = 0,$$

where the final inequality follows from defining property of a martingale.

Proof of Theorem 4.6. Note that from our assumptions, we have that for any sequences of constants $\{a_i\}$ and $\{b_i\}$, we have

$$\mathbb{E}\left[\prod_{i=1}^{n}(a_i + b_i X_i)\right] = \prod_{i=1}^{n} a_i. \tag{4.1}$$

Also, observe that for any a, the functions $f(x) = e^{ax}$ is convex. Thus for $x \in [-1, 1]$, it lies below the line connecting e^{-a} to e^{a}. In other words, for $x \in [-1, 1]$,

$$e^{ax} \leqslant \frac{e^a + e^{-a}}{2} + x\frac{e^a - e^{-a}}{2} = \cosh(a) + x\sinh(a).$$

Combining this with (4.1), we have for any t:

$$\mathbb{E}\left[e^{t\sum_{i=1}^n X_i}\right] \leqslant \mathbb{E}\left[\prod_{i=1}^n \cosh(t\|X_i\|_\infty) + \frac{X_i}{\|X_i\|_\infty}\sinh(t\|X_i\|_\infty)\right] = \prod_{i=1}^n \cosh(t\|X_i\|_\infty) \leqslant e^{t^2\|X_i\|_\infty^2/2},$$

where the final inequality follows from $\cosh(x) = \sum \frac{x^{2k}}{(2k)!} \leqslant \sum \frac{x^{2k}}{2^k k!} = e^{x^2/2}$.

Now we are in position to apply the method of Laplace transforms:

$$\mathbb{P}\left[\sum_{i=1}^n X_i > \lambda\right] \leqslant \frac{\mathbb{E}[e^{t\sum_{i=1}^n X_i}]}{e^{t\lambda}} \leqslant e^{(t^2/2)\sum_{i=1}^n \|X_i\|_\infty^2 - t\lambda}.$$

Setting $t = \frac{\lambda}{\sum_{i=1}^n \|X_i\|_\infty^2}$ finishes the proof. □

4.4 Additional applications

4.4.1 Concentration in product spaces

Define $\mathcal{U} = \{1, 2, \ldots, 6\}^n$. Define the *hamming distance* between $x, y \in \mathcal{U}$ by

$$H(x, y) := \#\{i \in [n] : x_i \neq y_i\},$$

and if $A \subseteq \mathcal{U}$, define $H(x, A) := \min\{H(x, y) : y \in A\}$. The following theorem shows that $\mathcal{U}$ exhibits "concentration of measure." Starting with any sufficiently large set $A \subseteq \mathcal{U}$, most of the points in $\mathcal{U}$ will be very close to A (the distance to A will be much smaller than the diameter of $\mathcal{U}$).

Theorem 4.7. *Consider any subset $A \subseteq \mathcal{U}$ with $|A| \geqslant 6^{n-1}$. Then for any $c > 0$,*

$$\frac{\left|\{x \in \mathcal{U} : H(x, A) \leqslant (c+2)\sqrt{n}\}\right|}{6^n} \geqslant 1 - e^{-c^2/2}. \tag{4.2}$$

Proof. Let $Z = (Z_1, \ldots, Z_n) \in \mathcal{U}$ be a uniformly random point. Define the Doob martingale $X_i = \mathbb{E}[H(Z, A) \mid Z_1, \ldots, Z_i]$. Since the map $x \mapsto H(x, A)$ is 1-Lipschitz, we know that $|X_i - X_{i-1}| \leqslant 1$ for every $i = 1, 2, \ldots, n$. Thus if $\mu = \mathbb{E}[H(Z, A)]$, Azuma's inequality yields

$$\mathbb{P}[H(Z, A) \leqslant \mu - c\sqrt{n}] \leqslant e^{-c^2/2}$$
$$\mathbb{P}[H(Z, A) \geqslant \mu + c\sqrt{n}] \leqslant e^{-c^2/2}.$$

It is not immediately obvious how to calculate μ, but we can get a good bound using concentration. If $\mu < 2\sqrt{n}$, then the first inequality gives

$$\mathbb{P}[H(Z, A) = 0] \leqslant e^{-2^2/2} = e^{-2} < 1/6,$$

but we know that $\mathbb{P}[Z = 0] = |A|/6^n \geqslant 1/6$, thus $\mu \geqslant 2\sqrt{n}$. Now apply the second inequality, yielding

$$\mathbb{P}[H(Z, A) \geqslant 2\sqrt{n} + c\sqrt{n}] \leqslant e^{-c^2/2}.$$

This is precisely our goal (4.2). □

4.4.2 Tighter concentration of the chromatic number

Previously, using the vertex exposure martingale we were able to prove reasonable concentration for $\chi(G)$ when $G \sim \mathcal{G}_{n,p}$. In what follows, we will put $p = n^{-\alpha}$ for some $\alpha > 0$. We will show that, surprisingly, if $\alpha > 5/6$, then with probability tending to one, $\chi(G)$ is concentrated on one of four values. In what follows, we will say that an event $\mathcal{E}_n$ (explicitly or implicitly indexed by n) holds "with high probability" if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$.

Lemma 4.8. *For any $c > 0$ and $\alpha > 5/6$, the following holds for $G \sim \mathcal{G}_{n,p}$: With high probability, every induced subgraph of size at most $c\sqrt{n}$ is 3-colorable.*

Proof sketch. Let S be the smallest subset of $V(G)$ that is not 3-colorable (if no such set exists, we are done). Then every $x \in S$ must have at least three neighbors in S, otherwise since $S \setminus \{x\}$ is 3-colorable, it would be the case that S is also 3-colorable. Thus the number of edges in the induced subgraph $G[S]$ is at least $3|S|/2$.

But it is unlikely that any set S with $|S| \leqslant c\sqrt{n}$ has at least $3|S|/2$ edges inside it. To see this, let $t = |S|$, and we'll compute the probability for a fixed set S: It's at most

$$p^{3t/2}\binom{\binom{t}{2}}{3t/2} \leqslant p^{3t/2}O(t)^{3t/2}\,.$$

Now we take a union bound over all sets of size at most T:

$$\sum_{t \leqslant T} p^{3t/2}O(t)^{3t/2}\binom{n}{t} \leqslant O(pT)^{3T/2}O\left(\frac{n}{T}\right)^T\,.$$

The latter inequality holds as long as $T \ll n$. Now using $p = n^{-\alpha}$ and $T \leqslant c\sqrt{n}$, this is bounded by

$$O(n)^{(1/2-\alpha)3T/2}O(n)^{T/2}\,,$$

and the latter quantity is $o(1)$ as long as $3/2(1/2 - \alpha) < 1/2$, i.e. $\alpha > 5/6$. □

Theorem 4.9. *With high probability, $\chi(G)$ takes one of four different values.*

Proof. Fix a number $\varepsilon > 0$ that we will send to 0. Let $u = u(n, p, \varepsilon)$ be the smallest integer so that $\mathbb{P}[\chi(G) \leqslant u] > \varepsilon$. Observe that, by the choice of u, we have $\mathbb{P}[\chi(G) > u - 1] \geqslant 1 - \varepsilon$.

Let $Y = Y(G)$ be the minimal size of a set of vertices S such that $\chi(G \setminus S) \leqslant u$. Consider the vertex exposure martingale for $G \sim \mathcal{G}_{n,p}$. Note that Y is 1-Lipschitz with respect to the exposure process because we could always add the modified vertex to S. Thus we can apply Azuma's inequality to the corresponding Doob martingale to conclude that

$$\mathbb{P}[Y \geqslant \mu + \lambda\sqrt{n}] \leqslant e^{-\lambda^2/2} \tag{4.3}$$

$$\mathbb{P}[Y \leqslant \mu - \lambda\sqrt{n}] \leqslant e^{-\lambda^2/2}\,, \tag{4.4}$$

where $\mu = \mathbb{E}[Y]$.

Let us choose λ so that $e^{-\lambda^2/2} = \varepsilon$. By the definition of u, we have $\mathbb{P}[Y = 0] > \varepsilon$. We conclude from (4.4) that $\mu \leqslant \lambda\sqrt{n}$. Now using (4.3), we see that $\mathbb{P}[Y \geqslant 2\lambda\sqrt{n}] \leqslant \varepsilon$.

By Lemma 4.8, we may assume that every subset of size at most $2\lambda\sqrt{n}$ is 3-colorable by throwing away an ε-fraction of graphs. Now observe that $Y < 2\lambda\sqrt{n}$ implies that G is $u + 3$ colorable since $G \setminus S$ is u-colorable and $|S| < 2\lambda\sqrt{n}$ so S can be colored with an additional 3 colors. We conclude that

$$\mathbb{P}[\chi(G) \in \{u, u + 1, u + 2, u + 3\}] \geqslant 1 - 3\varepsilon\,.$$

Sending $\varepsilon \to 0$ completes the proof. □

5 Memoryless random variables and low-diameter partitions

5.1 Random tree embeddings

Let (X, d) be a metric space on n points. Many problems in computation involve data points equipped with a natural distance. Metric spaces also arise as the solutions to linear programming relaxations of combinatorial problems. Often, it is useful to embed a metric space into a "simpler" one while changing the distances as little as possible. This is beneficial if an algorithmic problem can be solved more easily on the simple space.

A prime example of a simple metric space is a *tree metric.* Let T be a graph-theoretic with vertices $V(T)$ and edge set $E(T)$. We will assume that the tree is equipped some nonnegative length $\ell(e)$ on each edge $e \in E(T)$. There is a canonical shortest-path metric which we denote d_T.

A natural goal is to try and map our space (X, d) into a tree metric so that we preserve distances multiplicatively. In other words, we would look for a map $f : X \to V(T)$ so that

$$d(x, y) \leqslant d_T(f(x), f(y)) \leqslant D \cdot d(x, y) \qquad \forall x, y \in X\,,$$

and the *distortion* D is as small as possible. Unfortunately, this doesn't work so well. For instance, if (X, d) is the shortest-path metric on an unweighted n-cycle, one can show that any such mapping must have $D \geqslant \Omega(n)$. (Getting this argument right is actually a little tricky, but it is true.)

On the other hand, if we allow ourselves to use a *random* embedding, then we can approximate distance well in expectation. Consider again the shortest-path metric on an n-cycle C_n. Let T be the random (unweighted) tree that results from a single uniformly random edge of C_n. It's easy to see that for every $x, y \in V(C_n)$, we have

$$d_{C_n}(x, y) \leqslant \mathbb{E}\left[d_T(x, y)\right] \leqslant 2 \cdot d_{C_n}(x, y)\,.$$

The pair with the largest distortion is an edge $\{x, y\}$ of C_n; the expected length of $\{x, y\}$ in T is $(1 - \frac{1}{n})1 + \frac{1}{n}(n-1) \leqslant 2$.

Non-contracting tree embeddings. We now formalize the goal of embedding into a random tree. We say that (X, d) *admits a random tree embedding with distortion* D if there exists a random tree metric T and a random map $F : X \to V(T)$ that satisfies the following two properties:

1. **Non-contracting.**

   With probability one, for every $x, y \in X$, we have $d_T(F(x), F(y)) \geqslant d(x, y)$.

2. **Non-expanding in expectation.**

   For all $x, y \in X$,
   $$\mathbb{E}[d_T(F(x), F(y))] \leqslant D \cdot d(x, y)\,.$$

There are many scenarios in approximation algorithms and online algorithms where such embeddings can be used to reduce solving a problem in the general case to solving it on a tree by losing a factor of D in the approximation ratio (see Homework #3 for an example). It's also the case that such mappings can be useful for preconditioning diagonally dominant linear systems; in this case, one usually loses a factor of D (or $D^{O(1)}$) in the running time. In a sequence of works, Bartal showed that one can achieve $D = O(\log n \log\log n)$. The optimal bound was obtained a few years later.

Theorem 5.1 (Fakcharoenphol-Rao-Talwar 2003)**.** *Every n-point metric space (X, d) admits a random tree embedding with distortion $O(\log n)$.*

In Homework #3, you will prove that the theorem holds with $D = O((\log n)^2)$. We now discuss the basic primitive one needs.

5.2 Random low-diameter partitions

Given a parameter $\Delta > 0$ (the diameter bound), our goal is to construct a *random* partition $X = C_1 \cup C_2 \cup \cdots \cup C_k$ of X into sets where $\mathrm{diam}(C_i) \leqslant \Delta$ for every $i = 1, \ldots, k$. Here, $\mathrm{diam}(S) = \max_{x,y \in S} d(x, y)$ denotes the maximum distance in a subset $S \subseteq X$.

Of course, this is easy (we could simply decompose X into sets of singletons). We will also require that for every $x, y \in X$, we have

$$\mathbb{P}[x \text{ and } y \text{ are separated in } P] \leqslant \frac{d(x, y)}{\Delta} \cdot \alpha \,. \tag{5.1}$$

We say that a partition $P = \{C_1, C_2, \ldots, C_k\}$ separates x and y if $x \in C_i$, $y \in C_j$ and $i \neq j$. Our goal will be to prove that such random partitions always exist if we take $\alpha = 8 \ln n$.

Exercise: Prove that if $X = \mathbb{R}$ and $d(x, y) = |x - y|$ then we can find such a random partition with $\alpha = 1$. This demonstrates why the scaling in (5.1) is natural.

5.3 Memoryless random variables

Geometric random variables. Write $X \sim \mathrm{Geom}(p)$ to denote the fact that X is a geometric random variable with mean $1/p$. Recall that X is the number of independent coin flips needed to get heads when the coin comes up heads with probability p. It is easy to see that for every $k \geqslant 1$, we have $\mathbb{P}[X = k] = (1 - p)^{k-1} p$. One can also check that

$$\begin{aligned} \mathbb{P}[X \geqslant k] &= (1 - p)^{k-1} \qquad \forall k \geqslant 1 \\ \mathbb{P}[X = k \mid X \geqslant j] &= (1 - p)^{k-j} p \qquad \forall 1 \leqslant j \leqslant k \,. \end{aligned} \tag{5.2}$$

The last property expresses that a geometric random variable is "memoryless" in the sense that the distribution of $X \mid \{X > j\}$ is the same as the distribution of $X + j$.

Exponential random variables. There is also a continuous memoryless distribution: The exponential distribution. If $X \sim \mathrm{Exp}(\mu)$ is exponential with mean μ, then it has density $\mu e^{-t\mu}$, and the property that the distribution of $X \mid \{X > \lambda\}$ is the same that of $X + \lambda$.

5.4 The partitioning algorithm

Consider again a metric space (X, d) and a parameter $\Delta > 0$. Also recall our goal of producing a random partition whose sets have diameter at most Δ and such that (5.1) holds for $\alpha = 8 \ln n$. We may assume that $\Delta > 4 \ln n$, else we are done.

For simplicity, let us assume that $d(x, y) \in \{0, 1, 2, \ldots, n\}$. This will make the analysis slightly easier without sacrificing any of the essential details. Recall also that for $x \in X$ and $r \geqslant 0$, the ball of radius r around x is defined by

$$B(x, r) = \{y \in X : d(x, y) \leqslant r\} \,.$$

Order the vertices $X = \{x_1, \ldots, x_n\}$ arbitrarily. We produce the following random partition. For each $i = 1, 2, \ldots, n$, we choose an independent random variable $R_i \sim \text{Geom}(\frac{4 \ln n}{\Delta})$, and set

$$C_i = B(x_i, R_i) \setminus \bigcup_{j<i} C_j \,.$$

Thus the ith set if $B(x_i, R_i)$ but we remove the points that have already been clustered. Our partition is $P = \{C_1, C_2, \ldots, C_n\}$.

Note that, as stated, the algorithm could potentially output a set of diameter bigger than Δ: There's even a chance that $R_1 > \Delta$. If $\max_i R_i > \Delta/2$, we will output the partition $P^* = \{\{x\} : x \in X\}$ into singleton clusters. That ensures that the diameter of our sets are always bounded, and we will show this eventuality happens only with very small probability. Let's use $\mathcal{E}$ to denote the event that $\max_i R_i > \Delta/2$.

Now fix $x, y \in X$. Let $\mathcal{E}_{x,y}$ be the event that x and y end up in different sets of the partition P. (We are ignoring P^* for now.) We are interested in proving an upper bound on $\mathbb{P}(\mathcal{E}_{x,y})$. In order to do this, it's helpful to think about the process as "growing" balls around $x_1, x_2, \ldots$ in order until the whole space is partitioned. The growing is because the random radii are geometrically distributed, thus we can think about each center x_i flipping coins until one comes up heads and at each step incrementing the radius by one if the coin comes up tails.

With this picture in mind, it's intuitive that we only need to start getting worried about $\mathcal{E}_{x,y}$ occurring when some ball $B(x_i, R)$ "reaches" one of x or y. Until then, there are lots of growing balls that die out before they ever see one of x or y.

Let's make this intuition precise. Let $\mathcal{Z}_i$ denote the event that $\{x, y\} \cap C_i \neq \emptyset$ and $\{x, y\} \cap C_j = \emptyset$ for $j < i$. In other words, C_i is the first set that contains one of x or y (and it possibly contains both). Then

$$\mathbb{P}(\mathcal{E}_{x,y}) = \sum_{i=1}^{n} \mathbb{P}[\mathcal{Z}_i] \cdot \mathbb{P}[\mathcal{E}_{x,y} \mid \mathcal{Z}_i] \,.$$

So if we can show that $\mathbb{P}[\mathcal{E}_{x,y} \mid \mathcal{Z}_i] \leqslant p$ it will imply that $\mathbb{P}[\mathcal{E}_{x,y}] \leqslant p$.

Fix i. Let us suppose that $d(x, x_i) \leqslant d(y, x_i)$ (otherwise interchange the roles of x and y). Next, note that $\mathcal{Z}_i \implies \{R_i \geqslant d(x_i, x)\}$. Thus conditioned on $\mathcal{Z}_i$, we have the following: If $R_i \in [d(x_i, x), d(x_i, y))$ then $\mathcal{E}_{x,y}$ occurs, and otherwise $R_i \geqslant d(x_i, y)$ and $\mathcal{E}_{x,y}$ does not occur.

The memoryless property of the geometric distribution means that $R_i - d(x_i, x) \mid \{R_i \geqslant d(x_i, x)\}$ again has law $\text{Geom}(\frac{4 \ln n}{\Delta})$. We conclude that

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_{x,y} \mid \mathcal{Z}_i] &\leqslant \mathbb{P}[R_i \in [d(x_i, x), d(y_i, x)) \mid R_i \geqslant d(x_i, x)] \\
&= \mathbb{P}[R_i < d(y_i, x) \mid R_i \geqslant d(x_i, x)] \\
&= \mathbb{P}[X < d(y_i, x) - d(x_i, x)] \,,
\end{aligned}$$

where $X \sim \text{Geom}(\frac{4 \ln n}{\Delta})$.

Using (5.2) and the fact that $d(y, x_i) - d(x, x_i) \leqslant d(x, y)$, we have

$$\mathbb{P}[X < d(y_i, x) - d(x_i, x)] \leqslant \mathbb{P}[X < d(x, y)] = 1 - \mathbb{P}[X \geqslant d(x, y)] = 1 - \left(1 - \frac{4 \ln n}{\Delta}\right)^{d(x,y)-1}$$

Computation yields

$$1 - \left(1 - \frac{4 \ln n}{\Delta}\right)^{d(x,y)-1} \leqslant 1 - \left(1 - \frac{4 \ln n}{\Delta}\right)^{d(x,y)} \leqslant 1 - \left(1 - d(x, y)\frac{4 \ln n}{\Delta}\right) = \frac{d(x, y)}{\Delta} 4 \ln n \,,$$

where we have used the fact that $(1-\varepsilon)^k \geqslant 1-\varepsilon k$ for all $\varepsilon \in [0,1]$ and $k \geqslant 1$.

Thus $\mathbb{P}[\mathbb{E}_{x,y}] \leqslant \frac{d(x,y)}{\Delta} 4 \ln n$. We are almost done, but recall that we sometimes output the partition P^* instead of P. Thus

$$\mathbb{P}[x \text{ and } y \text{ are separated}] \leqslant \mathbb{P}[\mathcal{E}] + \frac{d(x,y)}{\Delta} 4 \ln n \,.$$

Using a union bound along with (5.2), we have

$$\mathbb{P}[\mathcal{E}] \leqslant n \cdot \mathbb{P}[R_1 > \Delta/2] \leqslant n \cdot \left(1 - \frac{4 \ln n}{\Delta}\right)^{\Delta/2} \leqslant n \cdot e^{-2 \ln n} = \frac{1}{n} \,.$$

Here, we have used the fact that $\left(1 - \frac{1}{k}\right)^k \leqslant \frac{1}{e}$ for $k \geqslant 1$. We conclude that

$$\mathbb{P}[x \text{ and } y \text{ are separated}] \leqslant \frac{1}{n} + \frac{d(x,y)}{\Delta} 4 \ln n \leqslant \frac{d(x,y)}{\Delta} 8 \ln n \,,$$

using our assumption that $\frac{d(x,y)}{\Delta} \geqslant \frac{1}{n}$ if $x \neq y$ (since $d(x,y) \in \{0,1,2,\ldots,n\}$).

6 Low-distortion embeddings

Let (X,d) be a finite metric space with $n = |X|$. Recall that the distance function $d : X \times X \to \mathbb{R}_+$ satisfies the axioms of a metric: For all $x, y, z \in X$:

1. $d(x,y) = 0 \iff x = y$
2. $d(x,y) = d(y,x)$
3. $d(x,y) \leqslant d(x,z) + d(z,y)$

While properties (2) and (3) are essential for us, the implication $d(x,y) = 0 \implies x = y$ is often not particularly important. If a distance function satisfies only (2), (3), and $d(x,x) = 0$, it is commonly called a *pseudometric.*

Metric spaces arise in a variety of mathematical and scientific domains since they abstract the properties of many natural notions of "similarity" between objects. Consider, for instance, the latency between nodes in a network, the travel-distance between cities, the edit distance between genetic sequences, or various similarity measures between proteins.

Often one might first try to understand a given metric space (X,d) by trying to compare it to a well-understood space. For instance, one could think about mapping $F : X \to \mathbb{R}^k$ into a Euclidean space $\mathbb{R}^k$ equipped with the Euclidean norm $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_k^2}$. One way to measure how well this mapping preserves the geometry of X is via the *bilipschitz distortion.* This is the smallest number $D > 0$ such that

$$d(x,y) \leqslant \|F(x) - F(y)\|_2 \leqslant D \cdot d(x,y) \qquad \forall x, y \in X \,. \tag{6.1}$$

Today we will prove the following result.

Theorem 6.1 (Bourgain 1985)**.** *Every n-point metric space embeds into some Euclidean space $\mathbb{R}^k$ with bilipschitz distortion D, where $D \leqslant O(\log n)$.*

We will show that this is possible $k \leqslant O((\log n)^2)$, but in the next lecture, we will see that for general reasons, one can achieve $k \leqslant O(\log n)$.

6.1 Distances to subsets

6.1.1 Fréchet's embedding

Let us first show how we can achieve the significantly worse bounds $D \leqslant \sqrt{n}$ and $k = n$. Enumerate the points $X = \{x_1, x_2, \ldots, x_n\}$ and let $F : X \to \mathbb{R}^n$ be defined by $F(x) = (F_1(x), \ldots, F_n(x))$, where

$$F_i(x) = d(x, x_i)\,.$$

First, note that every coordinate is 1-Lipschitz: For all $x, y \in X$,

$$|F_i(x) - F_i(y)| = |d(x, x_i) - d(y, x_i)| \leqslant d(x, y)\,,$$

where we have used the triangle inequality. From this, we get

$$\|F(x) - F(y)\|_2^2 = \sum_{i=1}^{n} |F_i(x) - F_i(y)|^2 \leqslant n\, d(x, y)^2\,, \tag{6.2}$$

implying that $\|F(x) - F(y)\|_2 \leqslant \sqrt{n} \cdot d(x, y)$ for all $x, y \in X$.

On the other hand, for any $x \in X$, it holds that

$$\|F(x) - F(y)\|_2 \geqslant |F_i(x) - F_i(x_i)| = d(x, x_i)\,,$$

Therefore (6.1) is satisfied with $D = \sqrt{n}$.

6.1.2 Bourgain's embedding

To get improved distortion, we will construct our coordinates out of distances to *subsets* instead of simply to points. For a subset $S \subseteq X$ and $x \in X$, let us define

$$d(x, S) := \min_{y \in S} d(x, y)\,.$$

First, observe that such maps are also 1-Lipschitz: The triangle inequality yields

$$d(x, S) \leqslant d(y, S) + d(x, y)\,,$$

hence

$$|d(x, S) - d(y, S)| \leqslant d(x, y) \qquad \forall x, y \in X, S \subseteq X\,. \tag{6.3}$$

For some number $m \leqslant O(\log n)$ that we will choose later, let

$$\left\{\boldsymbol{S}_{t,j} : t = 1, 2, \ldots, \lfloor \log_2 n \rfloor, j = 1, 2, \ldots, m\right\}\,.$$

denote independent random subsets $\boldsymbol{S}_{t,j} \subseteq X$, where $\boldsymbol{S}_{t,j}$ is formed by sampling every point of X independently with probability 2^{-t}. Our embedding is

$$\begin{aligned} F(x) = \Big(&d(x, S_{1,1}), \ldots, d(x, S_{1,m}),\\ &d(x, S_{2,1}), \ldots, d(x, S_{2,m}),\\ &\cdots\\ &d(x, S_{\lfloor \log_2 n \rfloor,1}), \ldots, d(x, S_{\lfloor \log_2 n \rfloor,m})\Big)\,. \end{aligned}$$

From (6.3), we see that

$$\|F(x) - F(y)\|_2 \leqslant \sqrt{m\lfloor \log_2 n \rfloor} \cdot d(x,y) \qquad x, y \in X \tag{6.4}$$

(just as in in (6.2)).

We move on to the lower bound. To this end, we define the open and closed balls: For $R \geqslant 0$,

$$B(x,R) = \{y \in X : d(x,y) \leqslant R\}\,,$$
$$B^\circ(x,R) = \{y \in X : d(x,y) < R\}\,.$$

Fix $x, y \in X$ and for $t = 1, 2, \ldots, \lfloor \log_2 n \rfloor$, let r_t be the smallest radius such that

$$\max\left\{|B(x,r_t)|, |B(y,r_t)| \geqslant 2^t\right\}\,.$$

Let t^* be the smallest value of t such that $r_t \geqslant d(x,y)/4$ and reassign $r_{t^*} = d(x,y)/4$.

Note that

$$\frac{d(x,y)}{4} = r_1 + (r_2 - r_1) + (r_3 - r_2) + \cdots + (r_{t^*} - r_{t^*-1})\,. \tag{6.5}$$

We will use the sets $S_{t,j}$ to get a contribution of $r_t - r_{t-1}$ to the lower bound, and therefore (6.5) shows we will get a contribution of $\Omega(d(x,y))$.

So consider now some $t \in \{1, 2, \ldots, t^*\}$. For the sake of analysis, let $r_0 = 0$. Note that, by definition of r_t, we have at least one of $|B(x, r_{t-1})| \geqslant 2^{t-1}$ or $|B(y, r_{t-1})| \geqslant 2^{t-1}$. Without loss of generality, assume that it holds for x. It also true that $B^\circ(y, r_t)| < 2^t$. Let us summarize:

$$|B(x, r_{t-1})| \geqslant 2^{t-1}$$
$$|B^\circ(y, r_t)| < 2^t\,.$$

Let $S_t \subseteq X$ be a random subset where every point is sampled independently with probability 2^{-t}. Consider the event

$$\mathcal{E}_t = \left\{S_t \cap B(x, r_{t-1}) \neq \emptyset \text{ and } S_t \cap B^\circ(y, r_t) = \emptyset\right\}\,.$$

Notice that

$$\mathcal{E}_t \text{ occurs } \implies |d(x, S_t) - d(y, S_t)| \geqslant r_t - r_{t-1}\,. \tag{6.6}$$

Claim 6.2. $\mathbb{P}(\mathcal{E}_t) \geqslant \frac{1}{12}$.

Proof. Observe that $r_t \leqslant d(x,y)/4$, hence $B(x, r_t)$ and $B(y, r_t)$ are disjoint. In particular, the two events composing $\mathcal{E}_t$ are independent, and it suffices to lower bound their probabilities separately. First, note that

$$\begin{aligned}
\mathbb{P}\left(S_t \cap B(x, r_{t-1}) \neq \emptyset\right) &\geqslant 1 - \mathbb{P}\left(S_t \cap B(x, r_{t-1}) = \emptyset\right) \\
&= 1 - \left(1 - 2^{-t}\right)^{|B(x, r_{t-1})|} \\
&\geqslant 1 - \left(1 - 2^{-t}\right)^{2^{t-1}} \\
&\geqslant 1 - \frac{1}{\sqrt{e}} \geqslant \frac{1}{3}\,,
\end{aligned}$$

where we have used the fact that $(1 - \frac{1}{k})^k \leqslant \frac{1}{e}$ for $k \geqslant 1$. Next, calculate

$$\mathbb{P}\left(S_t \cap B^\circ(y, r_t) = \emptyset\right) = \left(1 - 2^{-t}\right)^{|B^\circ(y, r_t)|} \geqslant \left(1 - 2^{-t}\right)^{2^t} \geqslant \frac{1}{4}\,,$$

where we have used $(1 - \frac{1}{k})^k \geqslant 1/4$ for $k \geqslant 2$. ∎

Now let $\mathcal{E}_{t,j}$ be the event corresponding to (6.6) for the set $S_{t,j}$.

Corollary 6.3. *If $\Omega(m)$ of the events $\{\mathcal{E}_{t,j} : j = 1, \dots, m\}$ occur, then*

$$\|F(x) - F(y)\|_2^2 \geqslant \Omega(m)(r_t - r_{t-1})^2 .$$

We can say something more: If it holds that

$$\Omega(m) \text{ of the events } \{\mathcal{E}_{t,j} : j = 1, \dots m\} \text{ occur for every } t = 1, 2, \dots, \lfloor \log_2 n \rfloor, \tag{6.7}$$

then since the contributions come from disjoint sets of coordinates,

$$\|F(x)-F(y)\|_2^2 \geqslant \Omega(m) \sum_{t=1}^{t^*} (r_t - r_{t-1})^2 \geqslant \Omega\left(\frac{m}{t^*}\right)\left(\sum_{t=1}^{t^*}(r_t - r_{t-1})\right)^2 \geqslant \Omega\left(\frac{m}{t^*}\right) d(x,y)^2 \geqslant \Omega\left(\frac{m}{\log n}\right) d(x,y)^2 .$$

The second inequality is Cauchy-Schwarz, and the third is from (6.5).

Combining this with (6.4), our map has distortion $O(\log n)$ as long as we choose m large enough so that (6.7) holds with probability, say, $1 - 1/n^3$. That's because we can then take a union bound over all possible pairs $x, y \in X$. But since each event $\mathcal{E}_{t,j}$ occurs with probability at least $1/12$, a simple Chernoff bound shows that choosing some $m \leqslant O(\log n)$ suffices.

7 The curse of dimensionality and dimension reduction

The "curse" of dimensionality refers to the fact that many algorithmic approaches to problems in $\mathbb{R}^k$ become exponentially more difficult as k grows. This is essentially due to volume growth: The k-dimensional Euclidean ball of radius r grows exponentially in k: $\mathrm{vol}_k(B(0, 2r)) = 2^k \mathrm{vol}_k(B(0, r))$.

On the other hand, in high dimensions, the concentration of measure phenomenon can come to our aid: Sufficiently smooth functionals on $\mathbb{R}^k$ are tightly concentrated around their expected value. A prototypical example is the Johnson-Lindenstrauss dimension reduction lemma.

7.1 The Johnson-Lindenstrauss lemma

Lemma 7.1 (Johnson-Lindenstrauss)**.** *For every $n \geqslant 1$ and every n-point subset $X \subseteq \mathbb{R}^n$, the following holds. For every $\varepsilon > 0$, there is a linear map $A : \mathbb{R}^n \to \mathbb{R}^k$ such that*

$$(1-\varepsilon)^2\|x-y\|_2^2 \leqslant \|A(x) - A(y)\|_2^2 \leqslant (1+\varepsilon)^2\|x-y\|_2^2 \qquad \forall x, y \in X , \tag{7.1}$$

and $k \leqslant \frac{24 \ln n}{\varepsilon^2}$.

Proof. We may assume that $0 < \varepsilon < 1/2$. We will a define random linear map $A : \mathbb{R}^n \to \mathbb{R}^k$ and then argue that A satisfies (7.1) with high probability. Let $\{X_i^{(j)} : i = 1, \dots, k, j = 1, \dots, n\}$ be a family of i.i.d. $N(0,1)$ random variables, and define an $m \times n$ matrix A by $A_{ij} = \frac{1}{\sqrt{k}} X_i^{(j)}$.

Claim 7.2. *For every $u \in \mathbb{R}^n$ with $\|u\|_2 = 1$, we have*

$$\mathbb{P}\left[\|Au\|_2^2 \notin [1-\varepsilon, 1+\varepsilon]\right] \leqslant 2e^{-\varepsilon^2 k/8} .$$

Let us finish the proof of Lemma 7.1 and then prove the claim. By setting $u = \frac{x-y}{\|x-y\|_2}$ for every $x, y \in X$ and taking a union bound over these n^2 different pairs, we have

$$\mathbb{P}\left((1-\varepsilon)\|x-y\|_2^2 \leqslant \|A(x-y)\|_2^2 \leqslant (1+\varepsilon)\|x-y\|_2^2\right) \geqslant 1 - 2n^2 e^{-\varepsilon^2 k/8} .$$

Taking $k = \frac{24 \ln n}{\varepsilon^2}$ the latter probability is at least $1 - \frac{2}{n}$, showing that the desired map A exists (and can be found by a randomized algorithm).

Proof of Claim 7.2. Note first that

$$\|Au\|_2^2 = \frac{1}{k}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} u_j X_i^{(j)}\right)^2 .$$

By the 2-stability property of independent $N(0,1)$ random variables, we have $\sum_{j=1}^{n} u_j X_i^{(j)}$ again has distribution $N(0,1)$. Thus we have

$$\|Au\|_2^2 = \frac{1}{k}\sum_{i=1}^{k} Y_i^2$$

where $\{Y_i\}$ are i.i.d. $N(0,1)$.

Since $\|Au\|_2^2$ is a sum of independent random variables, it makes sense to use the method of Laplace transforms. So far we have only done this for bounded random variables, but since $N(0,1)$ variables have a quickly diminishing tail, we can hope that the method will work here as well.

For some parameter $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\|Au\|_2^2 > 1+\varepsilon\right) &= \mathbb{P}\left(\sum_{i=1}^{k}(Y_i^2-1) > \varepsilon k\right) \\
&= \mathbb{P}\left(e^{\lambda\sum_{i=1}^{k}(Y_i^2-1)} > e^{\varepsilon\lambda k}\right) \\
&\leqslant \frac{\mathbb{E}\left[e^{\lambda\sum_{i=1}^{k}(Y_i^2-1)}\right]}{e^{\varepsilon\lambda k}} \\
&= \frac{\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(Y_i^2-1)}\right]}{e^{\varepsilon\lambda k}} .
\end{aligned}$$

Thus our remaining goal is to bound $\mathbb{E}\left[e^{\lambda(Y^2-1)}\right]$ when Y is $N(0,1)$. First, observe that

$$\mathbb{E}\left[e^{\lambda Y^2}\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-t^2/2}e^{\lambda t^2}\,dt = \frac{1}{\sqrt{1-2\lambda}} \tag{7.2}$$

for $0 < \lambda < 1/2$. (Clearly the integral is divergent for $\lambda \geqslant 1/2$.) Thus in this range of λ,

$$\mathbb{E}\left[e^{\lambda(Y^2-1)}\right] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} .$$

Let us finally calculate (using the Taylor expansion $\log(1-x) = -\sum_{k\geqslant 1}\frac{x^k}{k}$),

$$\begin{aligned}
\left|\log\mathbb{E}\left[e^{\lambda(Y^2-1)}\right]\right| &= \left|-\lambda - \frac{1}{2}\log(1-2\lambda)\right| \\
&\leqslant 2\lambda^2(1+(2\lambda)+(2\lambda)^2+\cdots) \\
&= \frac{2\lambda^2}{1-2\lambda} .
\end{aligned}$$

Therefore:

$$\mathbb{P}\left(\|Au\|_2^2 > 1+\varepsilon\right) \leqslant e^{2\lambda^2 k/(1-2\lambda)-\varepsilon\lambda k} .$$

Choosing $\lambda = \varepsilon/4$ and using $0 < \varepsilon < 1/2$, we conclude that $\mathbb{P}\left(\|Au\|_2^2 > 1+\varepsilon\right) \leqslant e^{-\varepsilon^2 k/8}$, completing the proof. A similar argument shows that $\mathbb{P}\left(\|Au\|_2^2 < 1-\varepsilon\right) \leqslant e^{-\varepsilon^2 k/8}$. □

□

Remark 7.3. A simple volume bound shows that that $\Theta(\log n)$ dependence in the dimension is necessary. A linear algebraic argument of Alon shows that the dimension must be at least $\Omega(\frac{\log n}{\varepsilon^2 \log(1/\varepsilon)})$. Very recently (FOCS 2017), Larsen and Nelson established that there is a lower bound of $\Omega(\frac{\log n}{\varepsilon^2})$, showing that Lemma 7.1 is tight up to the constant factor.

Remark 7.4. Lemma 7.1 actually works for any independent family $X_i^{(j)}$ where each random variable satisfies a sub-Gaussian tail bound: $\mathbb{E}[e^{\alpha X}] \leqslant e^{C\alpha^2}$ for some constant $C \geqslant 1$. For instance, if X is a uniform ± 1 random variable, then $\mathbb{E}[e^{\alpha X}] = \frac{1}{2}(e^{\alpha} + e^{-\alpha}) \leqslant e^{\alpha^2/2}$ (recall that to prove this, you should Taylor expand both sides).

The proof uses a clever trick. Note that if Z is a $N(0,1)$ random variable (independent of X) then $\mathbb{E}[e^{\alpha Z}] = e^{\alpha^2/2}$ for any $\alpha > 0$. Now write

$$\mathbb{E}[e^{\lambda X^2}] = \mathbb{E}[e^{(\sqrt{2\lambda}X)^2/2}] = \mathbb{E}[e^{\sqrt{2\lambda}ZX}] \leqslant \mathbb{E}[e^{2C\lambda Z^2}] = \frac{1}{\sqrt{1-2\lambda C}},$$

for $0 < \lambda < \frac{1}{2C}$, where the last equality is exactly the equality we proved in (7.2). Given this bound, we can finish the proof just as in Claim 7.2.

8 Compressive sensing and the RIP

The reason that extreme compression of photographs or audio recordings is possible is that the corresponding images are often sparse in the correct basis (e.g., the Fourier or wavelet basis). Thus one can take a very detailed photo and then zero out all the small coefficients, vastly compressing the image while also preserving the bulk of the important information.

Problematically, despite only recording a small amount of information at the end (say, s large Fourier coefficients), in order to figure out which coefficients to save, we had to perform a very detailed measurement (making our camera pretty expensive). *Compressive sensing* is the idea that, if we do a few random linear measurements, then we can capture the large coefficients without first knowing what they are.

Sparse recovery. Let us formalize the sparse recovery problem. Our signal will be a point $x \in \mathbb{R}^n$, and we will have a linear measurement map $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ that makes m linear measurements, where hopefully $m \ll n$. Say that a signal $x \in \mathbb{R}^n$ is *s-sparse* if $\|x\|_0 \leqslant s$, where $\|\cdot\|_0$ denotes the number of non-zero coordinates in its argument. For s-sparse signals x to be uniquely recoverable from the measurements $\Phi(x)$, the following property is necessary and sufficient: For every pair of *distinct* s-sparse vectors $x, y \in \mathbb{R}^n$, it holds that $\Phi(x) \neq \Phi(y)$.

Given the measurements $M = \Phi(x)$, we might want to recover the unique corresponding s-sparse vector x. It would be natural to solve the following optimization: $\min \|y\|_0$ subject to $\Phi(y) = M$. Clearly the optimizer y^* satisifies $\|y^*\|_0 \leqslant s$, so by the unique encoding property for s-sparse vectors and the fact that $\Phi(x) = \Phi(y)$, it must be that $x = y$. Unfortunately, ℓ_0 optimization subject to linear constraints is an NP-hard problem.

Instead, one often solves the problem: $\min \|y\|_1$ subject to $\Phi(y) = M$. This is a linear program and can thus be solved efficiently. It is often referred to as the "basis pursuit" algorithm. Remarkably, if we choose the map Φ appropriately, then the optimum solution y^* satisfies $x = y^*$, yielding an efficient algorithm for sparse recovery.

8.1 The restricted isometry property

We will now formalize the properties of the map $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ that makes efficient sparse recovery possible. For $s > 1$, let $\delta_s = \delta_s(\Phi)$ be the smallest number such that for every s-sparse vector $x \in \mathbb{R}^n$, we have

$$(1 - \delta_s)^2 \|x\|_2^2 \leqslant \|\Phi(x)\|_2^2 \leqslant (1 + \delta_s)^2 \|x\|_2^2 \tag{8.1}$$

It will help to think about this parameter in a slightly dierent way as well. Let $T \subseteq [n]$ index a subset of $|T| = s$ columns of Φ (thought of as an $m \times n$ matrix). Let $\Phi_T : \mathbb{R}^s \to \mathbb{R}^m$ be the linear map corresponding to the matrix formed from the columns of Φ indexed by T. Then the above property is equivalent to the property that for every $|T| = s$ and $x \in \mathbb{R}^s$, we have

$$(1 - \delta_s)^2 \|x\|_2^2 \leqslant \|\Phi_T(x)\|_2^2 \leqslant (1 + \delta_s)^2 \|x\|_2^2. \tag{8.2}$$

Theorem 8.1. *If $\delta_{2s}(\Phi) < 1$, then Φ has the unique recovery property for s-sparse vectors. If $\delta_{2s}(\Phi) < \sqrt{2}-1$, then ℓ_1-minimization performs s-sparse recovery.*

Proof. We will prove only the first assertion. Suppose that $x, y \in \mathbb{R}^n$ are s-sparse vectors. Then $x - y$ is $2s$-sparse, hence if $\Phi(x) = \Phi(y)$, then (8.1) gives

$$0 = \|\Phi(x - y)\|_2^2 \geqslant (1 - \delta_{2s})^2 \|x - y\|_2^2,$$

and therefore $x = y$ when $\delta_{2s} < 1$. □

8.2 Random construction of RIP matrices

Let us dene the $m \times n$ random matrix Φ by setting $\Phi_{ij} = \frac{1}{\sqrt{m}} X_i^{(j)}$ where $\{X_i^{(j)}\}$ is a family of i.i.d. $N(0, 1)$ random variables. With high probability, this matrix will have the RIP or appropriately chosen parameters.

Theorem 8.2. *For every $n \geqslant s \geqslant 1$ and $0 < \delta < 1$, there is an $m \leqslant O(\frac{s}{\delta^2} \log \frac{n}{s} + s \log \frac{1}{\delta})$ such that with high probability, $\delta_s(\Phi) \leqslant \delta$.*

Proof. Fix a subset $T \subseteq [n]$ with $|T| = s$. Let $\mathcal{E}_T$ denote the event that $\|\Phi_T(x)\|_2 \in [1 - \delta, 1 + \delta]$ for all $x \in \mathbb{R}^s$ with $\|x\|_2 = 1$. We will show that

$$\mathbb{P}[\mathcal{E}_T] \geqslant 1 - 2\left(\frac{16}{\delta}\right)^s e^{-\delta^2 m/48}. \tag{8.3}$$

Assuming this is true, we can take a union bound over $|T| = s$, yielding

$$\mathbb{P}[\delta_s(\Phi) \leqslant \delta] = \mathbb{P}\left[\mathcal{E}_T \text{ for every } T \subseteq [n], |T| = s\right] \geqslant 1 - 2\left(\frac{16}{\delta}\right)^s e^{-\delta^2 m/48} \binom{n}{s}$$

Using the fact that $\log \binom{n}{s} \leqslant O(s \log \frac{n}{s})$, we can conclude by choosing m as in the theorem statement so that this probability is at least, say, $1 - 1/n$.

Thus we are left to prove (8.3). Let N be a $\delta/4$-net on the unit sphere in $\mathbb{R}^s$. This is a collection of unit vectors N such that for every $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, there is an $x' \in N$ with $\|x - x'\|_2 \leqslant \delta/4$. A simple volume argument shows we can choose such a net with $|N| \leqslant (16/\delta)^s$.

Now using Claim 1.2 from Lecture 10 and a union bound over N, we have

$$\mathbb{P}\left(\forall x \in N, \|\Phi_T(x)\|_2 \in \left[1 - \frac{\delta}{4}, 1 + \frac{\delta}{4}\right]\right) \geqslant 1 - 2\left(\frac{16}{\delta}\right)^s e^{-\delta^2 m/48}.$$

We are left to show that $\|\Phi_T(x)\|_2 \in [1-\frac{\delta}{4}, 1+\frac{\delta}{4}]$ for all $x \in N$ implies $\|\Phi_T(x)\|_2 \in [1-\delta, 1+\delta]$ or all $x \in \mathbb{R}^s$ with $\|x\|_2 = 1$.

This uses a clever trick. We will define a sequence of points $\{x_i : i \geqslant 0\}$ such that $\frac{x_i}{\|x_i\|} \in N$ for every $i \geqslant 0$. For any $y \in \mathbb{R}^s$, let $\Gamma(y) = y'\|y\|_2$ where $y' \in N$ is the closest point from N to $y/\|y\|_2$. Note that by the net property, we have $\|y - \Gamma(y)\|_2 \leqslant \frac{\delta}{4}\|y\|_2$.

Consider now some $\|x\|_2 = 1$. Define $x_0 := \Gamma(x)$, $x_1 := \Gamma(x - x_0)$, and so on:

$$x_{i+1} := \Gamma(x - (x_0 + \cdots + x_i))$$

Then:

$$x = x_0 + (x - x_0) = x_0 + x_1 + (x - x_0 - x_1) = \cdots = \sum_{i=0}^{\infty} x_i.$$

By a simple induction, we have $\|x_i\|_2 \leqslant (\delta/4)^i$ and by construction, $x_i/\|x_i\| \in N$ for every $i \geqslant 0$.

Now we can use our assumption that $\|\Phi(y)\|_2 \in [1-\delta/4, 1+\delta/4]$ for every $y \in N$ to write

$$\|\Phi_T(x)\|_2 \leqslant \sum_{i=0}^{\infty} \|\Phi_T(x_i)\|_2 \leqslant \left(1+\frac{\delta}{4}\right)\sum_{i=0}^{\infty}\left(\frac{\delta}{4}\right)^i = \frac{1+\delta/4}{1-\delta/4} \leqslant 1+\delta,$$

where the last inequality follows from $\delta < 1$. For the other side, write

$$\begin{aligned}\|\Phi_T(x)\|_2 \geqslant \|\Phi_T(x_0)\|_2 - \sum_{i=1}^{\infty}\|\Phi_T(x_i)\|_2 &\geqslant \left(1-\frac{\delta}{4}\right) - \frac{\delta}{4}\left(1+\frac{\delta}{4}\right)\sum_{i=0}^{\infty}(\delta/4)^i \\ &\geqslant 1 - \frac{\delta}{4} - \frac{\delta(1+\frac{\delta}{4})}{4(1-\delta/4)} \geqslant 1-\delta,\end{aligned}$$

where we again used $\delta < 1$. We have thus confirmed (8.3), completing the proof. □

Remark 8.3. Note that we must always perform s "measurements" even if we know exactly the s important coordinates. The preceding theorems says that we can do unique (and efficient) recovery with only $O(s \log n)$ measurements without knowing anything about the input signal except that it's s-sparse.

Remark 8.4. In a more realistic model, we might expect that our signal is of the form $x = x_s + y$ where x_s is s-sparse and $\|y\|_2 \leqslant \varepsilon\|x\|_2$. In other words, the signal has s large coordinates plus "noise." The RIP and basis pursuit algorithms can also be used to provide guarantees in this setting.

9 Concentration for sums of random matrices

Consider the *graph sparsification problem:* Given a graph $G = (V, E)$, we want to approximate G (in a sense to be defined later) by a sparse graph $H = (V, E')$. Generally we would like that $E' \subseteq E$ and moreover $|E'|$ is as small as possible—say $O(n)$ or $O(n \log n)$ where $n = |V|$. We will be able to do this by choosing a (nonuniform) random sample of the edges, but to analyze such a process, we will need a large-deviation inequality for sums of random *matrices.*

9.1 Symmetric matrices

If A is a $d \times d$ real symmetric matrix, then A has all real eigenvalues which we can order $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots \geqslant \lambda_d(A)$. The *operator norm* of A is

$$\|A\| := \max_{\|x\|_2=1}\|Ax\|_2 = \max\{|\lambda_i(A)| : i \in \{1, \ldots, d\}\}\ .$$

The trace of A is $\mathrm{Tr}(A) = \sum_{i=1}^d A_{ii} = \sum_{i=1}^d \lambda_i(A)$. The *trace norm* of A is $\|A\|_* = \sum_{i=1}^d |\lambda_i(A)|$. A symmetric matrix is *positive semidefinite* (PSD) if all its eigenvalues are nonnegative. Note that for a PSD matrix A, we have $\mathrm{Tr}(A) = \|A\|_*$. We also recall the matrix exponential $e^A = \sum_{k=0}^\infty \frac{A^k}{k!}$ which is well-defined for all real symmetric A and is itself also a real symmetric matrix. If A is symmetric, then e^A is always PSD, as the next argument shows.

Every real symmetric matrix can be diagonalized, writing $A = U^T D U$, where U is an orthogonal matrix, i.e. $UU^T = U^T U = I$ and D is diagonal. One can easily check that $A^k = U^T D^k U$ for any $k \in \mathbb{N}$, thus A^k and A are simultaneously diagonalizable. It follows that A and e^A are simultaneously diagonalizable. In particular, we have $\lambda_i(e^A) = e^{\lambda_i(A)}$.

Finally, note that for symmetric matrices A and B, we have $|\mathrm{Tr}(AB)| \leqslant \|A\| \cdot \|B\|_*$. To see this, let $\{u_i\}$ be an orthonormal basis of eigenvectors of B with $Bu_i = \lambda_i(B)u_i$. Then

$$|\mathrm{Tr}(AB)| = \left|\sum_{i=1}^d \langle u_i, ABu_i\rangle\right| = \left|\sum_{i=1}^d \lambda_i(B)\langle u_i, Au_i\rangle\right| \leqslant \sum_{i=1}^d |\lambda_i(B)| \cdot \|A\| = \|B\|_*\|A\|.$$

Many classical statements are either false or significantly more difficult to prove when translated to the matrix setting. For instance, while $e^{x+y} = e^x e^y = e^y e^x$ is true for arbitrary real numbers x and y, it is only the case that $e^{A+B} = e^A e^B$ if A and B are simultaneously diagonalizable. However, somewhat remarkably, the matrix analog does hold if we do it inside the trace.

Theorem 9.1 (Golden-Thompson inequality)**.** *If A and B are real symmetric matrices, then*

$$\mathrm{Tr}(e^{A+B}) \leqslant \mathrm{Tr}(e^A e^B)\,.$$

Proof. We can prove this using the non-commutative Hölder inequality: For any even integer $p \geqslant 2$ and real symmetric matrices $A_1, A_2, \ldots, A_p$:

$$\left|\mathrm{Tr}(A_1 A_2 \cdots A_p)\right| \leqslant \|A_1\|_{S_p}\|A_2\|_{S_p} \cdots \|A_p\|_{S_p},$$

where $\|A\|_{S_p} = \left(\mathrm{Tr}(A^T A)^{p/2}\right)^{1/p}$ is the Schatten p-norm. Consider real symmetric matrices U, V. Applying this with $A_1 = \cdots A_p = UV$ gives, for every even $p \geqslant 2$:

$$\mathrm{Tr}((UV)^p) \leqslant \|UV\|_{S_p}^p = \mathrm{Tr}((V^T U^T UV)^{p/2}) = \mathrm{Tr}((VU^2V)^{p/2}) = \mathrm{Tr}((U^2V^2)^{p/2}),$$

where the last inequality uses the cyclic property of the trace. Applying this inequality repeatedly now yields, for every even p,

$$\mathrm{Tr}((UV)^p) \leqslant \mathrm{Tr}(U^p V^p).$$

If we now take $U = e^{A/p}$ and $V = e^{B/p}$, this gives

$$\mathrm{Tr}((e^{A/p}e^{B/p})^p) \leqslant \mathrm{Tr}(e^A e^B). \tag{9.1}$$

For p large, we can use the Taylor approximation $e^{A/p} = 1 + A/p + O(1/p^2)$ and similarly for $e^{B/p}$. Thus: $e^{A/p}e^{B/p} \sim e^{(A+B)/p+O(1/p^2)}$. Therefore taking $p \to \infty$ in (9.1) gives

$$\mathrm{Tr}(e^{A+B}) \leqslant \mathrm{Tr}(e^A e^B). \qquad \square$$

9.2 The method of exponential moments for matrices

We will consider now a random $d \times d$ real matrix X. The entries (X_{ij}) of X are all (not necessarily independent) random variables. We have seen inequalities (like those named after Chernoff and Azuma) which assert that if $X = X_1 + X_2 + \cdots + X_n$ is a sum of *independent* random numbers, then X is tightly concentrated around its mean. Our goal now is to prove a similar fact for sums of independent random symmetric matrices.

First, observe that the trace is a linear operator; this is easy to see from the fact that it is the sum of the diagonal entries of its argument. If A and B are arbitrary real matrices, then $\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$. This implies that if X is a random matrix, then $\mathbb{E}[\mathrm{Tr}(X)] = \mathrm{Tr}(\mathbb{E}[X])$. Note that $\mathbb{E}[X]$ is the matrix defined by $(\mathbb{E}[X])_{ij} = \mathbb{E}[X_{ij}]$.

Suppose that $X_1, X_2, \ldots, X_n$ are independent random real symmetric matrices. Let $X = X_1 + X_2 + \cdots + X_n$. Let $S_k = X_1 + \cdots + X_k$ be the partial sum of the first k terms so that $X = S_n$. Our first goal will be to bound the probability that X has an eigenvalue bigger than t. To do this, we will try to extend the method of exponential moments to work with symmetric matrices, as discovered by Ahlswede and Winter. It is much simpler than previous approaches that only worked for special cases.

Note that for $\beta > 0$, we have $\lambda_i(e^{\beta X}) = e^{\beta \lambda_i(X)}$. Therefore:

$$\mathbb{P}\left[\max_i \lambda_i(X) > t\right] = \mathbb{P}\left[\max_i \lambda_i(e^{\beta X}) > e^{\beta t}\right] \leqslant \mathbb{P}\left[\mathrm{Tr}(e^{\beta X}) > e^{\beta t}\right], \tag{9.2}$$

where the last inequality uses the fact that all the eigenvalues of $e^{\beta X}$ are nonnegative, hence $\mathrm{Tr}(e^{\beta X}) = \sum_i \lambda_i(e^{\beta X}) \geqslant \max_i \lambda_i(e^{\beta X})$.

Now Markov's inequality implies that

$$\mathbb{P}[\mathrm{Tr}(e^{\beta X}) > e^{\beta t}] \leqslant \frac{\mathbb{E}[\mathrm{Tr}(e^{\beta X})]}{e^{\beta t}}. \tag{9.3}$$

As in our earlier uses of the Laplace transform, our goal is now to bound $\mathbb{E}[\mathrm{Tr}(e^{\beta X})]$ by a product that has one factor for each term X_i.

In the matrix setting, this is more subtle: Using Theorem 9.1,

$$\mathbb{E}[\mathrm{Tr}(e^{\beta X})] = \mathbb{E}[\mathrm{Tr}(e^{\beta(S_{n-1}+X_n)})] \leqslant \mathbb{E}[\mathrm{Tr}(e^{\beta S_{n-1}} e^{\beta X_n})].$$

Now we push the expectation over X_n inside the trace:

$$\mathbb{E}[\mathrm{Tr}(e^{\beta S_{n-1}} e^{\beta X_n})] = \mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}}\, \mathbb{E}[e^{\beta X_n} \mid X_1, \ldots, X_{n-1}])\right] = \mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}}\, \mathbb{E}[e^{\beta X_n}])\right],$$

and we have used independence to pull $e^{\beta S_{n-1}}$ outside the expectation and then to remove the conditioning. Finally, we use the fact that $\mathrm{Tr}(AB) \leqslant \|A\| \cdot \|B\|_*$ and $\|B\|_* = \mathrm{Tr}(B)$ when B has all nonnegative eigenvalues (as is the case for $e^{\beta S_{n-1}}$):

$$\mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}}\, \mathbb{E}[e^{\beta X_n}])\right] \leqslant \left\|\mathbb{E}[e^{\beta X_n}]\right\| \mathbb{E}\left[\mathrm{Tr}(e^{\beta S_{n-1}})\right].$$

Doing this n times yields

$$\mathbb{E}[\mathrm{Tr}(e^{\beta X})] \leqslant \mathrm{Tr}(I) \prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\| = d \prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\|.$$

Combining this with (9.2) and (9.3) yields

$$\mathbb{P}\left[\max_i \lambda_i(X) > t\right] \leqslant e^{-\beta t} d \prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\| .$$

We can also apply this to $-X$ to get

$$\mathbb{P}\left[\|X\| > t\right] \leqslant e^{-\beta t} d \left(\prod_{i=1}^{n} \left\|\mathbb{E}[e^{\beta X_i}]\right\| + \prod_{i=1}^{n} \left\|\mathbb{E}[e^{-\beta X_i}]\right\|\right) . \tag{9.4}$$

9.3 Large-deviation bounds

Let Y be a random, symmetric, psd $d \times d$ matrix with $\mathbb{E}[Y] = I$. Suppose that $\|Y\| \leqslant L$ with probability one.

Theorem 9.2. *If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. copies of Y, then for any $\varepsilon \in (0,1)$ the following holds. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the eigenvalues of $\frac{1}{n}\sum_{i=1}^{n} Y_i$. Then*

$$\mathbb{P}[\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \subseteq [1-\varepsilon, 1+\varepsilon]] \geqslant 1 - 2d \exp\left(-\varepsilon^2 n/4L\right).$$

There is a slightly nicer way to write this using the *Löwner ordering of symmetric matrices:* We write $A \succeq B$ to denote that the matrix $A - B$ is positive semidefinite. We can rewrite the conclusion of Theorem 9.2 as

$$\mathbb{P}\left[(1-\varepsilon)I \preceq \frac{1}{n}\sum_{i=1}^{n} Y_i \preceq (1+\varepsilon)I\right] \geqslant 1 - 2d \exp\left(-\varepsilon^2 n/4L\right) . \tag{9.5}$$

Proof of Theorem 9.2. Define $X_i := Y_i - \mathbb{E}[Y_i]$ and $X := X_1 + \cdots + X_n$. Then (9.5) is equivalent to

$$\mathbb{P}\left[\|X\| > \varepsilon n\right] \leqslant 2d \exp\left(-\varepsilon^2 n/4L\right) .$$

We know from (9.4) that it will suffice to bound $\left\|\mathbb{E}[e^{\beta X_i}]\right\|$ for each i. To do this, we will use the fact that

$$1 + x \leqslant e^x \leqslant 1 + x + x^2 \qquad \forall x \in [-1, 1] .$$

Note that if A is a real symmetric matrix, then since I, A, A^2, and e^A are simultaneously diagonalizable, this yields

$$I + A \preceq e^A \preceq I + A + A^2 \tag{9.6}$$

for any A with $\|A\| \leqslant 1$.

Observe that $\mathbb{E}[X_i] = 0$. To evaluate $\|X_i\|$, let us use the fact that for real symmetric A, we have $\|A\| = \max_{\|x\|_2=1} |x^T A x|$. So consider some $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$ and write

$$|x^T X_i x| = \frac{|x^T Y_i x - x^T \mathbb{E}[Y_i] x|}{n} \leqslant |x^T Y_i x| \leqslant L ,$$

where we have used the fact that since Y_i is PSD, so is $\mathbb{E}[Y_i]$, and thus $x^T \mathbb{E}[Y_i] x$ and $x^T Y_i x$ are both nonnegative. We also used our assumption that $\|Y_i\| \leqslant L$. We conclude that $\|X_i\| \leqslant L$.

Moreover, we have

$$\mathbb{E}[X_i^2] = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] = \left(\mathbb{E}[Y_i^2] - (\mathbb{E}[Y_i])^2\right) \preceq \mathbb{E}[Y_i^2] \preceq \mathbb{E}[\|Y_i\| Y_i] \preceq L\, \mathbb{E}[Y_i] ,$$

where in the final line we again used the assumption $\|Y_i\| \leqslant L$. Finally, since $\mathbb{E}[Y_i] = I$, conclude that $\|\mathbb{E}[X_i^2]\| \leqslant L$.

Therefore for any $\beta \leqslant 1/L$, we can apply (9.6), yielding

$$\mathbb{E}[e^{\beta X_i}] \preceq I + \beta\,\mathbb{E}[X_i] + \beta^2\,\mathbb{E}[X_i^2] = I + \beta^2\,\mathbb{E}[X_i^2] \preceq e^{\beta^2\,\mathbb{E}[X_i^2]}\,.$$

We conclude that for $\beta \leqslant 1/L$, we have $\|\mathbb{E}[e^{\beta X_i}]\| \leqslant e^{\beta^2 L}$.

Plugging this into (9.4), we see that

$$\mathbb{P}\left[\|X\| > \varepsilon n\right] \leqslant 2de^{-\varepsilon n\beta}e^{\beta^2 Ln}.$$

Choosing $\beta := \frac{\varepsilon}{2L}$ yields

$$\mathbb{P}\left[\|X\| > \varepsilon n\right] \leqslant 2de^{-\varepsilon^2 n/4L}\,,$$

completing the argument. □

Finally, we can prove a generalization of Theorem 9.2 for random matrices whose expectation is not the identity.

Theorem 9.3. *Let Z be a $d \times d$ random real, symmetric, PSD matrix. Suppose also that $Z \preceq L \cdot \mathbb{E}[Z]$ for some $L \geqslant 1$. If $Z_1, Z_2, \ldots, Z_n$ are i.i.d. copies of Z, then for any $\varepsilon > 0$, it holds that*

$$\mathbb{P}\left[(1-\varepsilon)\,\mathbb{E}[Z] \preceq \frac{1}{n}\sum_{i=1}^{n} Z_i \preceq (1+\varepsilon)\,\mathbb{E}[Z]\right] \geqslant 1 - 2de^{-\varepsilon^2 n/4L}\,.$$

In other words, the empirical mean $\frac{1}{n}\sum_{i=1}^n Z_i$ is very close (in a spectral sense) to the expectation $\mathbb{E}[Z]$.

Proof of Theorem 9.3. This is made difficult only because it may be that $A = \mathbb{E}[Z]$ is not invertible. Suppose that $A = UDU^T$ where D is a diagonal matrix $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0)$, where k is the rank of A and $\lambda_i \neq 0$ for $i = 1, \ldots, k$. Then the *pseudoinverse* of A is defined by

$$A^+ = U\mathrm{diag}(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_k^{-1}, 0, \ldots, 0)U^T\,.$$

Note that $AA^+ = A^+A = I_{\mathrm{im}(A)}$, where $I_{\mathrm{im}(A)}$ denotes the operator that acts by the identity on the image of A, and annihilates $\ker(A)$.

Since A is PSD, A^+ is also PSD, and we can define $A^{+/2}$ as the *square root* of A^+. One can write this explicitly as

$$A^{+/2} = U\mathrm{diag}(\lambda_1^{-1/2}, \lambda_2^{-1/2}, \ldots, \lambda_k^{-1/2}, 0, \ldots, 0)U^T\,.$$

Now to prove Theorem 9.3, it suffices to apply Theorem 9.2 to the matrices $Y_i = A^{+/2}Z_iA^{+/2}$. Verification is left as an exercise. □

10 Spectral sparsification

Consider the *graph sparsification problem:* Given a graph $G = (V, E)$, we want to approximate G (in a sense to be defined shortly) by a sparse graph $H = (V, E')$. Generally we would like that $E' \subseteq E$ and moreover $|E'|$ is as small as possible—say $O(n)$ or $O(n \log n)$ where $n = |V|$.

10.1 Laplacians of graphs

In everything that follows, we will consider n-vertex graphs with vertex set $V = \{1, 2, \ldots, n\}$. For an edge $e \in E$ with $e = \{i, j\}$ and $i < j$, we define the vector $x_e = e_i - e_j$ where $\{e_i\}$ are the standard basis vectors in $\mathbb{R}^n$. We also define the $n \times n$ matrix $L_e = x_e x_e^T$. Notice that this matrix has $(L_e)_{ii} = (L_e)_{jj} = 1$ and $(L_e)_{ij} = (L_e)_{ji} = -1$; the rest of the entries are zero. This matrix has rank one and is positive semidefinite: For every $v \in \mathbb{R}^n$, we have

$$v^T L_e v = (v_i - v_j)^2 .$$

Now for a graph $G = (V, E)$, we define the *(combinatorial) Laplacian of G* by

$$L_G = \sum_{e \in E} L_e .$$

It should be clear that L_G is also PSD (since it is a nonnegative sum of PSD matrices) and

$$v^T L_G v = \sum_{\{i,j\} \in E} (v_i - v_j)^2 . \tag{10.1}$$

If G is equipped with nonnegative edge weights $\{w_e \geqslant 0 : e \in E\}$, we define the corresponding weighted Laplacian by $L_G = \sum_{e \in E} w_e L_e$.

10.1.1 Spectral sparsification

Spielman and Teng introduced the following notion of *spectral* graph approximation. Consider (possibly weighted) graphs H and G. We say that H *ε-spectrally approximates G* for some $\varepsilon > 0$ if

$$(1 - \varepsilon) L_G \preceq L_H \preceq (1 + \varepsilon) L_G . \tag{10.2}$$

Recall that this equivalent to requiring that for every $v \in \mathbb{R}^n$, we have

$$(1 - \varepsilon) v^T L_G v \leqslant v^T L_H v \leqslant (1 + \varepsilon) v^T L_G v .$$

From this expression, we see that spectral approximation is stronger than cut approximation. Indeed, consider any subset of vertices $S \subseteq V$ and the corresponding characteristic vector $v = \mathbf{1}_S$. Then $v^T L_G v = w_G(E(S, \bar{S}))$ and $v_T L_H v = w_H(E(S, \bar{S}))$, where these two expressions are meant to represent the weight of the edges in G that cross the cut $(S, \bar{S})$ and the weight of the edges in H that cross the cut $(S, \bar{S})$, respectively. In particular, (10.2) entails that the weight of every cut in H should be within $1 \pm \varepsilon$ of the weight of the corresponding cut in G.

10.2 Random sampling

We will prove the following theorem.

Theorem 10.1 (Spielman-Srivastava 2008)**.** *For every $\varepsilon > 0$, the following holds. For every unweighted, connected graph $G = (V, E)$, there exists a weighted graph $H = (V, E')$ such that $E' \subseteq E$ and $|E| \leqslant O(\frac{n \log n}{\varepsilon^2})$ and H ε-spectrally approximates G.*

To see that weights are necessary in Theorem 10.1, consider the case when G is an n-clique. In that case, we will need to put large weights on any sparse graph H that approximates G (recalling that, in particular, the weight of every cut in H should be close to the size of the corresponding cut in G).

A sampling algorithm. Suppose we have some probabilities $p_e \geqslant 0$ such that $\sum_{e \in E} p_e = 1$. Then we can consider the following algorithm: Set all edge weights $w_e := 0$ for every $e \in E$. For $i = 1, 2, \ldots, k$, sample an edge e (independent from previous choices) with probability p_e and update

$$w_e := w_e + \frac{1}{k p_e} .$$

Let H be the corresponding (random) weighed graph.

It is easy to see that $L_H = \sum_{i=1}^{k} Z_i$ where $Z_i = \frac{1}{k p_{e^{(i)}}} L_{e^{(i)}}$ and $e^{(i)}$ is the edge that we sampled in the ith iteration. Moreover, we can calculate for any $i \in \{1, \ldots, k\}$:

$$\mathbb{E}[Z_i] = \sum_{e \in E} p_e \frac{1}{k p_e} L_e = \frac{1}{k} L_G .$$

Therefore $\mathbb{E}[L_H] = L_G$. The expectation of our "approximator" is equal to L_G. Moreover, our approximator L_H is a sum of i.i.d. random matrices. In order to achieve (10.2) using a small value of k, we need concentration for this sum. And in order to achieve concentration using the approach of Lecture 14, we need a bound on the eigenvalues of individual summands.

Claim 10.2. *For any $\kappa \geqslant 1$, if $Z_i \preceq \kappa \, \mathbb{E}[Z_i]$ (with probability one), then we can choose $k = O(\frac{\kappa}{\varepsilon^2} \log n)$ and achieve (10.2) with high probability.*

Proof. Theorem 1.3 of Lecture 14 states that under our assumptions, we have

$$\mathbb{P}\left[(1-\varepsilon) L_G \preceq \sum_{i=1}^{k} Z_i \preceq (1+\varepsilon) L_G \right] \geqslant 1 - 2n e^{-\varepsilon^2 k}/4\kappa .$$

Plugging in some $k = O(\frac{\kappa}{\varepsilon^2} \log n)$, we can get this expression to be at least $1 - 1/n$. □

So we are left to choose sampling probabilities p_e that can guarantee $Z_i \preceq \kappa \, \mathbb{E}[Z_i]$ for some reasonable value of κ. A natural choice would be *uniform* sampling: $p_e = 1/|E|$. But this can fail to give non-trivial bounds: Suppose that G is the union of two disjoint $n/2$-cliques connected by a single edge a. In order for H to spectrally approximate G, we had better include the edge a (otherwise the corresponding cut in H will have weight zero, while it has non-zero weight in G). But this would mean we need to sample $\Theta(n^2)$ edges if we do uniform random sampling (since G contains $\Theta(n^2)$ edges, only one of which is a). That doesn't yield a particularly sparse graph.

Remark 10.3. It is not too difficult to see that no matter what choice we make for $\{p_e\}$ we will need at least $\Omega(n \log n)$ edges to be sampled. Consider G to be the n-path. In that case, we will need to sample every edge at least once to approximate all $n - 1$ cuts in G. The standard coupon collector bound dictates that we will need at least $\Omega(n \log n)$ samples.

10.2.1 Effective resistances

Instead, we need to choose the probabilities $\{p_e\}$ so that important edges (like the single edge a in the preceding example) have a very good chance of being sampled. To do this, we will set p_e to be proportional to the *effective resistance* of the edge e. We'll see more about effective resistances (and their relationship to random walks) in the next lecture.

For now, we can simply give the definition: For every edge e, we set

$$\mathsf{R}_e := \mathrm{Tr}(L_e L_G^{+}) ,$$

where we recall that L_G^+ is the *pseudo-inverse of* L_G.

Since G is connected, L_G has rank $n-1$. That is easy to see from the formula (10.1). If v is a multiple of the all-ones vector $(1,1,\ldots,1)$, then $v^T L_G v = 0$. On the other hand, if $v^T L_G v = \sum_{\{i,j\}\in E}(v_i - v_j)^2 = 0$ and G is connected, it must be that $v_1 = v_2 = \cdots = v_n$. In other words, $L_G v = 0$ implies that v is a multiple of $(1,1,\ldots,1)$. Thus L_G has rank $n-1$. That means we can write $L_G = \sum_{i=2}^n \lambda_i w_i w_i^T$ for some orthonormal family of vectors $\{w_i\} \subseteq \mathbb{R}^n$ and $\lambda_i > 0$. One then defines $L_G^+ = \sum_{i=1}^n \frac{1}{\lambda_i} w_i w_i^T$. Now we have $L_G L_G^+ = I_{\mathrm{im}(L_G)}$. We will also need the positive square root: $L_G^{+/2} = \sum_{i=1}^n \frac{1}{\sqrt{\lambda_i}} w_i w_i^T$.

Finally we define our sampling probabilities $p_e = \frac{\mathsf{R}_e}{n-1}$. First, let's verify that these are indeed probabilities. Recalling that $L_e = x_e x_e^T$, we have $\mathsf{R}_e = \mathrm{Tr}(L_e L_G^+) = x_e^T L_G^+ x_e \geqslant 0$ since L_G^+ is PSD (because L_G is PSD). Moreover, using linearity of the trace and the definition of L_G:

$$\sum_{e\in E} \mathsf{R}_e = \mathrm{Tr}\left(\sum_{e\in E} L_e L_G^+\right) = \mathrm{Tr}(L_G L_G^+) = \mathrm{Tr}(I_{\mathrm{im}(L_G)}) = n-1\,,$$

where the last line follows because, as we discussed, L_G has rank $n-1$.

Note that with these sampling probabilities, we have

$$Z_i \preceq \kappa \cdot \mathbb{E}[Z_i] \iff \frac{n-1}{k\mathsf{R}_e} L_e \preceq \kappa \frac{1}{k} L_G \iff L_e \preceq \kappa \frac{\mathsf{R}_e}{n-1} L_G$$

By Claim 10.2 with $\kappa = n-1$, we are done with the proof of Theorem 10.1 after we verify the following.

Lemma 10.4. *For every edge $e \in E$, we have $L_e \preceq \mathsf{R}_e L_G$.*

Proof. We need to prove that $v^T L_e v \leqslant \mathsf{R}_e v^T L_G v$ for every $v \in \mathbb{R}^n$. But we need only prove this for $v \perp (1,\ldots,1)$ since $(1,\ldots,1) \in \ker(L_G) \cap \ker(L_e)$. Since $\mathrm{im}(L_G^{+/2})$ contains every vector orthogonal to $(1,\ldots,1)$, it suffices to prove the inequality for $v = L_G^{+/2} w$ for any $w \in \mathbb{R}^n$ with $w \perp (1,\ldots,1)$:

$$(L_G^{+/2} w)^T L_e (L_G^{+/2} w) \leqslant \mathsf{R}_e (L_G^{+/2} w)^T L_G (L_G^{+/2} w)$$

Using the fact that L_G^+ is symmetric, and $w \perp \ker(L_G)$, the RHS is $\mathsf{R}_e w^T w$. On the other hand, the LHS is $w^T L_G^{+/2} L_e L_G^{+/2} w \leqslant \|L_G^{+/2} L_e L_G^{+/2}\| w^T w \leqslant \mathrm{Tr}(L_G^{+/2} L_e L_G^{+/2}) w^T w = \mathsf{R}_e w^T w$, completing the proof. □

Remark 10.5. The proof is a bit easier to understand in the following way: We want to prove $L_e \preceq \mathsf{R}_e L_G$. It would be nice to simply multiply on the left and right by $L_G^{+/2}$ yielding

$$L_G^{+/2} L_e L_G^{+/2} \preceq \mathsf{R}_e I_{\mathrm{im(G)}}\,.$$

This latter statement is true because the maximum eigenvalue of $L_G^{+/2} L_e L_G^{+/2}$ is certainly at most its trace which is equal to R_e.

If A and B are symmetric matrices and C is invertible, then $CAC^T \preceq CBC^T \iff A \preceq B$ (this is an easy exercise). On the other hand, if C is singular (like $L_G^{+/2}$), then we need to be careful about what happens on $\ker(C)$.

11 Random walks and electrical networks

Let $G = (V, E)$ be an undirected graph. The random walk on G is a Markov chain on V that, at each time step, moves to a uniformly random neighbor of the current vertex.

For $x \in V$, use d_x to denote the degree of vertex x. Then more formally, *random walk on G* is the following process $\{X_t\}$. We start at at some node $X_0 = v_0 \in V$. Then if $X_t = v$, we put $X_{t+1} = u$ with probability $1/d_v$ for every neighbor u of v.

11.1 Hitting times and cover times

One can study many natural properties of the random walk. For two vertices $u, v \in V$, we define the *hitting time H_{uv} from u to v* as the expected number of steps for the random walk to hit v when started at u. Formally, define the random variable $T = \min\{t \geqslant 0 : X_t = v\}$. Then $H_{uv} = \mathbb{E}[T \mid X_0 = u]$.

The *cover time of G starting from u* is the quantity $\mathrm{cov}_u(G)$ which is the expected number of steps needed to visit every vertex of G started at u. Again, we can define this formally: Let $T = \min\{t \geqslant 0 : \{X_0, X_1, \ldots, X_t\} = V\}$. Then $\mathrm{cov}_u(G) = \mathbb{E}[T \mid X_0 = u]$. Finally, we define the *cover time of G* as $\mathrm{cov}(G) = \max_{u \in V} \mathrm{cov}_u(G)$.

11.2 Random walks and electrical networks

It turns out that random walks (on undirected graphs) are very closely related to electrical networks. We recall the basics of such networks now. Again, we let $G = (V, E)$ be a connected, undirected graph which we think of as an electrical circuit with unit resistors on every edge.

If we create a potential difference at two vertices (by, say, connecting the positive and negative terminals of a battery), then we induce an electrical flow in the graph. Between every two nodes u, v there is a *potential* $\phi_{u,v} \in \mathbb{R}$. Electrical networks satisfying the following three laws.

(K1) The flow into every node equals the flow out.

(K2) The sum of the potential differences around any cycle is equal to zero.

(Ohm) The current flowing from u to v on an edge $e = \{u, v\}$ is precisely $\frac{\phi_{u,v}}{r_{uv}}$ where r_{uv} is the resistance of $\{u, v\}$. [In other words, $V = iR$.]

In our setting, all resistances are equal to one, but one can define things more generally: If we put conductances c_{uv} on the edges $\{u, v\} \in E$, then the corresponding random walk would operate as follows: If $X_t = u$ then $X_{t+1} = v$ with probability $\frac{c_{uv}}{\sum_{v \in V} c_{uv}}$ for every neighbor v of u. In that case, we would have $r_{uv} = 1/c_{uv}$.

Remark 11.1. In fact, (K2) is related to a somewhat more general fact. The potential differences are given—naturally—by differences in a potential. There exists a map $\varphi : V \to \mathbb{R}$ such that $\phi_{u,v} = \varphi(u) - \varphi(v)$. If G is connected, then the potential φ is uniquely defined up to a tranlation.

To define the potential φ, put $\varphi(v_0) = 0$ for some fixed node v_0. Now for any $v \in V$ and any path $\gamma = \langle v_0, v_1, v_2, \ldots, v_k = v\rangle$ in G, we can define $\varphi(v) = \phi_{v_0,v_1} + \phi_{v_1,v_2} + \cdots + \phi_{v_{k-1},v_k}$. This is well-defined—independent of the choice of path γ—since by (K2), the potential differences around every cycle sum to zero.

Finally, we make an important definition: The *effective resistance $R_{\mathrm{eff}}(u, v)$ between two nodes $u, v \in V$* is defined to be the necessary potential difference created between u and v to induce a current of one unit to flow between them. If we imagine the entire graph G acting as a single "wire"

between u and v, then $R_{\text{eff}}(u,v)$ denotes the effective resistance of that single wire (recall Ohm's law). We now prove the following.

Theorem 11.2. *If $G = (V, E)$ has m edges, then for any two nodes $u, v \in V$, we have*

$$H_{uv} + H_{vu} = 2mR_{\text{eff}}(u,v)\,.$$

In order to prove this, we will set up four electrical networks corresponding to the graph G. We label these networks (A)-(D).

(A) We inject d_x units of flow at every vertex $x \in X$, and extract $\sum_{x\in V} d_x = 2m$ units of flow at vertex v.

(B) We inject d_x units of flow at every vertex $x \in X$, and extract $2m$ units of flow at vertex u.

(C) We inject $2m$ units of flow at vertex u and extract d_x units of flow at every vertex $x \in X$.

(D) We inject $2m$ units of flow at vertex u and extract $2m$ units of flow at vertex v.

We will use the notation $\phi^{(A)}_{x,y}, \phi^{(B)}_{x,y}$, etc. to denote the potential differences in each of these networks.

Lemma 11.3. *For any vertex $u \in V$, we have $H_{uv} = \phi^{(A)}_{u,v}$.*

Proof. Calculate: For $u \neq v$,

$$\begin{aligned} d_u &= \sum_{w\sim u} \phi^{(A)}_{u,w} \\ &= \sum_{w\sim u} \left(\phi^{(A)}_{u,v} - \phi^{(A)}_{w,v}\right) \\ &= d_u\phi^{(A)}_{u,v} - \sum_{w\sim u} \phi^{(A)}_{w,v}\,, \end{aligned}$$

where we have first use (K1), then (K2). Rearranging yields

$$\phi^{(A)}_{u,v} = 1 + \frac{1}{d_u}\sum_{w\sim u} \phi^{(A)}_{w,v}\,.$$

Now observe that the hitting times satisfy the same set of linear equations: For $u \neq v$,

$$H_{uv} = 1 + \frac{1}{d_u}\sum_{w\sim u} H_{wv}\,.$$

We conclude that $H_{uv} = \phi^{(A)}_{u,v}$ as long as this system of linear equations has a unique solution. But consider some other solution H'_{uv} and define $f(u) = H_{uv} - H'_{uv}$. Plugging this into the preceding family of equations yields

$$f(u) = \frac{1}{d_u}\sum_{w\sim u} f(w)\,.$$

Such a map f is called *harmonic*, and it is a well-known fact that every harmonic function f on a finite, connected graph is constant. Since $f(v) = H_{vv} = 0$, this implies that $f \equiv 0$, and hence the family of equations has a unique solution, completing the proof. □

Remark 11.4. To prove that every harmonic function on a finite, connected graph is constant, we can look at the corresponding Laplace operator: $(Lf)(u) = d_u f(u) - \sum_{w \sim u} f(u)$. A function f is harmonic if and only if $Lf = 0$. But we have already seen that, on a connected graph, the Laplacian has rank $n - 1$ and $\ker(L) = \operatorname{span}(1, \dots, 1)$, i.e., the only harmonic functions on our graph are multiples of the constant function.

Define now the *commute time between u and v* as the quantity $C_{uv} = H_{uv} + H_{vu}$. We restate and prove Theorem 11.2.

Theorem 11.5. *In any connected graph with m edges, we have $C_{uv} = 2mR_{\text{eff}}(u, v)$ for every pair of vertices $u, v \in V$.*

Proof. From Lemma 11.3, we have $H_{uv} = \phi^{(A)}_{u,v}$. By symmetry, $H_{vu} = \phi^{(B)}_{v,u}$ as well Since network C is the reverse of network B, this yields $H_{vu} = \phi^{(C)}_{u,v}$. Finally, since network D is the sum of networks A and C, by linearity we have

$$\phi^{(D)}_{u,v} = \phi^{(C)}_{u,v} + \phi^{(A)}_{u,v} = H_{uv} + H_{vu} = C_{uv} \,.$$

Finally, note that $R_{\text{eff}}(u, v) = 2m\phi^{(D)}_{u,v}$ by definition, since network D has exactly $2m$ units of current flowing from u to v. This yields the claim of the theorem. □

11.3 Cover times

We can now use Theorem 11.5 to give a universal upper bound on the cover time of any graph.

Theorem 11.6. *For any connected graph $G = (V, E)$, we have* $\operatorname{cov}(G) \leqslant 2|E|(|V| - 1)$.

Proof. Fix a spanning tree T of G. Then we have

$$\operatorname{cov}(G) \leqslant \sum_{\{x,y\} \in E(T)} C_{xy} \,.$$

The right-hand side can be interpreted as a very particular way of covering the graph G: Start at some node x_0 and "walk" around the edges of the spanning tree in order $x_0, x_1, x_2, \dots, x_{2(n-1)} = x_0$. If we require the walk to first go from x_0 to x_1, then from x_1 to x_2, etc., we get the sum $\sum_{i=0}^{2(n-1)-1} H_{x_i x_{i+1}} = \sum_{\{x,y\} \in E(T)} C_{xy}$. This is one particular way to visit every node of G, so it gives an upper bound on the cover time.

Finally, we note that if $\{x, y\}$ is an edge of the graph, then by Theorem 11.5, we have $C_{xy} = 2|E|R_{\text{eff}}(x, y) \leqslant 2|E|$. Here we use the fact that for every edge $\{x, y\}$ of a graph, the effective resistance is at most the resistance, which is at most one. This completes the proof. □

Remark 11.7. The last stated fact is a special case of the *Rayleigh monotonicity principle*. This states that adding edges to the graph (or, more generally, decreasing the resistance of any edge) cannot increase any effective resistance. In the other direction, removing edges from the graph (or, more generally, increasing the resistance of any edge) cannot decrease any effective resistance. A similar fact is false for hitting times and commute times, as we will see in the next few examples.

Examples.

1. **The path.** Consider first G to be the path on vertices $\{0, 1, \ldots, n\}$. Then $H_{0n} + H_{n0} = C_{0n} = 2nR_{\text{eff}}(0, n) = 2n^2$. Since $H_{0n} = H_{n0}$ by symmetry, we conclude that $H_{0n} = n^2$. Note that Theorem 11.6 implies that $\text{cov}(G) \leqslant 2n^2$, and clearly $\text{cov}(G) \geqslant H_{0n} = n^2$, so the upper bound is off by at most a factor of 2.

2. **The lollipop.** Consider next the "lollipop graph" which is a path of length $n/2$ from u to v with an $n/2$ clique attached to v. We have $H_{uv} + H_{vu} = C_{uv} = \Theta(n^2)R_{\text{eff}}(u, v) = \Theta(n^3)$. On the other hand, we have already seen that $H_{uv} = \Theta(n^2)$. We conclude that $H_{vu} = \Theta(n^3)$, hence $\text{cov}(G) \geqslant \Omega(n^3)$. Again, the bound of Theorem 11.6 is $\text{cov}(G) \leqslant O(n^3)$, so it's tight up to a constant factor here as well.

3. **The complete graph.** Finally, consider the complete graph G on n nodes. In this case, Theorem 11.6 gives $\text{cov}(G) \leqslant O(n^3)$ which is way off from the actual value $\text{cov}(G) = \Theta(n \log n)$ (since this is just the coupon collector problem is flimsy disguise).

11.4 Matthews' bound

The last example shows that sometimes Theorem 11.6 doesn't give such a great upper bound. Fortunately, a relatively simple bound gets us within an $O(\log n)$ factor of the cover time.

Theorem 11.8. *If $G = (V, E)$ is a connected graph and $R_{\max} := \max_{x,y \in V} R_{\text{eff}}(x, y)$ is the maximum effective resistance in G, then*

$$|E|R_{\max} \leqslant \text{cov}(G) \leqslant O(\log n)|E|R_{\max}\,.$$

Proof. One direction is easy:

$$\text{cov}(G) \geqslant \max_{u,v} H_{uv} \geqslant \frac{1}{2} \max_{u,v} C_{uv} = \frac{1}{2} 2|E| \max_{u,v} R_{\text{eff}}(u, v) = |E|R_{\max}\,.$$

For the other direction, we will examine a random walk of length $2c|E|R_{\max} \log n$ divided into $\log n$ epochs of length $2c|E|R_{\max}$. Note that for any vertex v and any epoch i, we have

$$\mathbb{P}[v \text{ unvisited in epoch } i] \leqslant \frac{1}{c}\,.$$

This is because no matter what vertex is the first of epoch i, we know that the hitting time to v is at most $\max_u H_{uv} \leqslant \max_u C_{uv} \leqslant 2|E|R_{\max}$. Now Markov's inequality tells us that the probability it takes more than $2c|E|R_{\max}$ steps to hit v is at most $1/c$.

Therefore the probability we don't visit v in any epoch is at most $c^{-\log n} = n^{-\log c}$, and by a union bound, the probability that there is some vertex left unvisited after all the epochs is at most $n^{1-\log c}$.

We conclude that

$$\text{cov}(G) \leqslant 2c|E|R_{\max} \log n + n^{1-\log c} 2n^3\,,$$

where we have used the weak upper bound on the cover time provided by Theorem 11.6. Choosing c to be a large enough constant makes the second term negligible, yielding

$$\text{cov}(G) \leqslant O(|E|R_{\max} \log n)\,,$$

as desired. □

One can make some improvements to this "soft" proof, yielding the following stronger bounds.

Theorem 11.9. *For any connected graph $G = (V, E)$, the following holds. Let t_{hit} denote the maximum hitting time in G. Then*

$$\text{cov}(G) \leqslant t_{\text{hit}} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).$$

Moreover, if we define for any subset $A \subseteq V$, the quantity $t^A_{\text{min}} = \min_{u,v \in A, u \neq v} H_{uv}$, then

$$\text{cov}(G) \geqslant \max_{A \subseteq V} t^A_{\text{min}} \left(1 + \frac{1}{2} + \cdots + \frac{1}{|A| - 1}\right). \tag{11.1}$$

For the proofs, consult Chapter 11 of the Levin-Peres-Wilmer book http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf.

Kahn, Kim, Lovász, and Vu showed that the best lower bound in (11.1) is within an $O(\log \log n)^2$ factor of $\text{cov}(G)$, improving over the $O(\log n)$-approximation in Theorem 11.8. In a paper with Jian Ding and Yuval Peres, we showed that one can compute an $O(1)$ approximation using a multi-scale generalization of the bound (11.1) based on Talagrand's majorizing measures theory.

12 Markov chains and mixing times

Consider a finite state space Ω and a *transition kernel* $P : \Omega \times \Omega \to [0, 1]$ such that for every $x \in \Omega$, $\sum_{y \in \Omega} P(x, y) = 1$. The *Markov chain* corresponding to the kernel P is the sequence of random variables $\{X_0, X_1, X_2, \ldots\}$ such that for every $t \geqslant 0$, we have $\mathbb{P}[X_{t+1} = y \mid X_t = x] = P(x, y)$. Note that we also have to specify a distribution for the initial state X_0.

Corresponding to every such process, one can consider the (weighted) directed graph $D = (\Omega, A)$ with $A = \{(x, y) : P(x, y) > 0\}$ and edge weights $w(x, y) = P(x, y)$. Then the random process $\{X_t\}$ corresponds precisely to random walk on D: At every time step, one moves from the current vertex x to a neighbor y with probability $P(x, y)$.

Convergence to stationarity. For every $t \geqslant 0$, let $P^t(x, y) = \mathbb{P}[X_t = x \mid X_0 = y]$. The Markov chain described by P is said to be *irreducible* if for every $x, y \in \Omega$, there is some t such that $P^t(x, y) > 0$; in other words, there is always some way to reach any state form any other. This corresponds precisely to the digraph D being strongly connected. The chain is *aperiodic* if for every $x, y \in \Omega$,

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

Theorem 12.1 (Fundamental Theorem of Markov Chains)**.** *If P is irreducible and aperiodic, then there is a unique probability measure $\pi : \Omega \to [0, 1]$ such that for every $x, y \in \Omega$, we have*

$$P^t(x, y) \to \pi(y) \quad \text{as} \quad t \to \infty.$$

In other words, the Markov chain "forgets" where it started and converges to a unique limiting distribution. This is referred to as the *stationary measure* π.

Reversibility. A Markov chain is said to be *reversible with respect to the measure μ* if for every $x, y \in \Omega$, we have $\mu(x)P(x, y) = \mu(y)P(y, x)$. (These are called the "detailed balance conditions.") The chain is said to be *reversible* if it is reversible with respect to some probability measure. Note that reversible chains correspond precisely to random walks on (weighted) *undirected graphs.*

Also, if P is irreducible and aperiodic—and hence has a unique stationary measure π by Theorem 12.1—then actually $\pi = \mu$. To see this, note that by the detailed balance conditions: For every $y \in \Omega$, we have

$$\sum_{x\in\Omega} \mu(x)P(x,y) = \sum_{x\in\Omega} \mu(y)P(y,x) = \mu(y)\sum_{x\in\Omega} P(y,x) = \mu(y)\,. \tag{12.1}$$

The right-hand side can be interpreted as the probability of going to y in one step started from the measure μ. Now Theorem 12.1 implies that if we start from distribution μ, then we converge to π; on the other hand, (12.1) says that if we start distributed according to μ, then we stay that way under the chain. Thus $\mu = \pi$. This provides a nice local way to check that some measure is the stationary measure of the chain.

Remark 12.2. If P is irreducible, but not necessarily aperiodic, then there is still a unique stationary distribution, i.e. a probability π such that for every $x \in \Omega$, $\sum_{y\in\Omega} P(x,y)\pi(y) = \pi(x)$. But it may not be the case that the chain converges to π from some starting states.

For instance, if the chain is given by a directed graph with two nodes $\Omega = \{x, y\}$ and arcs (x, y) and (y, x), then $\pi = (1/2, 1/2)$ is the unique stationary measure, but the chain does not converge to π when starting in either state x or y (because of periodicity).

For our purposes, aperiodicity is a rather weak obstruction to mixing. Given any chain P and number $\alpha \in (0,1)$, we can consider the chain $P' = \alpha I + (1-\alpha)P$. If P is irreducible, then so is P'. Moreover, for any such α, the chain P' is aperiodic (even if P was not). When measuring convergence to equilibrium, this α "self loop" probability does not slow down the chain too much.

12.1 The Fundamental Theorem

Let us sketch a proof of Theorem 12.1. We want to begin with a *stationary measure* π for P, i.e., a probability distribution π on Ω (interpreted as a row vector) such that $\pi P = \pi$. Suppose $\{X_t\}$ is the Markov chain with transition law P and for $x \in \Omega$, define

$$\tau_x^+ := \min\{t > 0 : X_t = x\}.$$

We do not prove the following lemma (see, e.g., Section 1.5.3 in the Levin-Peres-Wilmer book).

Lemma 12.3. *Suppose that P is irreducible and aperioidic. If we define*

$$\pi(x) = \frac{1}{\mathbb{E}[\tau_x^+ \mid X_t = 0]},$$

then π is a probability distribution satisfying $\pi P = \pi$.

Now let us argue that $P^t(x,y) \to \pi(y)$ for every $x, y \in \Omega$. Let Π denote the $|\Omega| \times |\Omega|$ matrix where every row is π.

Fact 12.4. *It holds that $\Pi P = \Pi$ and $Q\Pi = \Pi$ for* every *row-stochastic matrix Q.*

Using the fact that P is irreducible and aperiodic, choose r such that $P^r(x,y) > 0$ for every $x, y \in \Omega$. Let $\theta < 1$ be such that

$$P^r(x,y) \geqslant (1-\theta)\pi(y) \qquad \forall x, y \in \Omega.$$

Then we can write

$$P^r = (1-\theta)\Pi + \theta Q, \tag{12.2}$$

where Q is stochastic. Now we claim that for every $k \geqslant 1$:

$$P^{kr} = (1 - \theta^k)\Pi + \theta^k Q^k. \tag{12.3}$$

If this holds, then for $s < r$, we have

$$P^{kr+s} = P^{kr}P^s = (1 - \theta^k)\Pi + \theta^k Q^k P^s,$$

and thus $P^t \to \Pi$ as $t \to \infty$, completing the proof of Theorem 12.1.

We prove (12.3) by induction on k. The case $k = 1$ is (12.2). In the general case, write

$$\begin{aligned} P^{r(k+1)} &= P^{rk}P^r \\ &= \left[(1 - \theta^k)\Pi + \theta^k Q^k\right] P^r \\ &= (1 - \theta^k)\Pi P^r + \theta^k Q^k \left[(1 - \theta)\Pi + \theta Q\right], \end{aligned}$$

where in the second line we use (12.2). Now (12.4) implies that $\Pi P^r = \Pi$ and $Q^k\Pi = \Pi$, hence

$$P^{r(k+1)} = \left[(1 - \theta^k) + \theta^k(1 - \theta)\right] \Pi + \theta^{k+1}Q^{k+1} = (1 - \theta^{k+1})\Pi + \theta^{k+1}Q^{k+1},$$

completing the proof.

12.2 Eigenvalues and mixing

It will be useful to give a more quantitative proof of Theorem 12.1 in the reversible case. To do this, we again think of P as an $\Omega \times \Omega$ matrix. If we also think about a probability measure $\mu \in \mathbb{R}^\Omega$ as a row vector, then μP denotes the distribution that arises by starting at μ and taking one step of the chain associated to P.

If P is reversible with respect to π, then (12.1) implies that $\pi P = \pi$, i.e. π is a (left) eigenvector with eigenvalue 1. We now analyze the other eigenvalues of P.

Real eigenvalues. Note that P is not necessarily a symmetric matrix, but we can prove that P is similar to a symmetric matrix. Let D denote the diagonal matrix with $D_{xx} = \pi(x)$. Then

$$(\sqrt{D^{-1}}P\sqrt{D})_{xy} = \langle e_x, e_y\sqrt{D^{-1}}P\sqrt{D}\rangle = \langle \sqrt{D^{-1}}e_x, e_y P\sqrt{D}\rangle = \sqrt{\frac{\pi(x)}{\pi(y)}}P(x, y).$$

But by (12.1), this is equal to $\sqrt{\frac{\pi(y)}{\pi(x)}}P(y, x)$. Thus $\sqrt{D^{-1}}P\sqrt{D}$ is a real, symmetric matrix and hence has real eigenvalues. This implies that P also has real eigenvalues.

All eigenvalues in $[-1, 1]$. Now note that for any $v \in \mathbb{R}^\Omega$, we have

$$\|vP\|_1 \leqslant \||v|P\|_1 = \|v\|_1, \tag{12.4}$$

where $|v|$ denotes the vector whose entries are the absolute value of the corresponding entries in v. This is simply because P is an averaging operator.

Now suppose that $vP = \lambda v$. Then using (12.4)

$$|\lambda| \cdot \|v\|_1 = \|vP\|_1 \leqslant \|v\|_1,$$

implying that $|\lambda| \leqslant 1$.

Unique eigenvector with eigenvalue 1. Suppose now that $v = vP$ and consider the corresponding Laplacian matrix $L = D - PD$ (using our notation for "edge Laplacians," this is $\frac{1}{2}\sum_{x,y} \pi(x)P(x,y)L_{\{x,y\}}$). One can check that this matrix is symmetric since PD is symmetric by the detailed balance conditions. As we saw in Lectures 14-15, for any vector w we have

$$wLw^T = \frac{1}{2}\sum_{x,y} \pi(x)P(x,y)\,(w_x - w_y)^2\,.$$

This is easiest to see by writing $L = \sum_{\{x,y\}} c_{xy}L_{xy}$ where $c_{xy} := \pi(x)P(x,y)$ and L_{xy} is the (unweighted) Laplacian corresponding to the graph with a single edge $\{x,y\}$. (The factor $1/2$ is due to the fact that we are summing over all pairs x, y vs. all edges $\{x,y\}$.)

Let $w = vD^{-1}$. Then $vP = v \implies wL = 0 \implies wLw^T = 0$. Thus $w_x = w_y$ whenever $P(x,y) > 0$. But since the chain P is irreducible, we can connect every pair x, y by a chain of such implications, implying that $w = \alpha(1,1,\ldots,1)$ is a multiple of the all-ones vector. But this implies that $v = Dw$ is a multiple of π. Since P is an averaging operator, it preserves the ℓ_1 norm, hence $\alpha = 1$ and $v = \pi$.

Not a bipartite graph. Now we claim that if P is aperiodic, -1 cannot be an eigenvalue of P. Suppose, for the sake of contradiction, that $vP = -v$ for some $v \neq 0$. Again, let $|v|$ denote the vector whose entries are the absolute values of the corresponding entries in v. Then

$$\|v\|_2^2 = \big||v|Pv^T\big| \leqslant |v|P|v|^T \leqslant \|v\|_2^2\,,$$

where the last inequality follows from the fact that all the eigenvalues of P lie in $[-1,1]$. We conclude that $|v|P = |v|$, implying that $|v| = \pi$.

Finally, observe that $vP = -v$ implies that for every x, one has $v_x = -(Pv)_x$, hence

$$\pi(x)\mathrm{sgn}(v_x) = v_x = -(vP)_x = -\sum_y P(y,x)v_y = -\sum_y P(y,x)\pi(y)\mathrm{sgn}(v_y)\,.$$

But by the detailed balance conditions, we have $\pi(x) = \sum_y P(y,x)\pi(y)$. Hence it must be that $\mathrm{sgn}(v_x) = -\mathrm{sgn}(v_y)$ whenever $P(y,x) > 0$.

Thus if we set $L = \{x : v_x < 0\}$ and $R = \{x : v_x > 0\}$, then $P(x,y) > 0$ implies x and y are on different sides of the bipartition. (Note that this is a bipartition since $|v| = \pi$ implies that $v_x \neq 0$ for any $x \in \Omega$.) But this implies that for x, y on the same side of the bipartition, we have $P^t(x,y) = 0$ when t is odd, contradicting the fact that P was assumed aperiodic.

Convergence to stationarity (spectral argument). Let us fix an inner product in which the matrix P is self-adjoint:

$$\langle u, v\rangle_{L^2(\pi)} = \sum_{x\in\Omega} \pi(x)u_x v_x,$$

and let $\|u\|_{L^2(\pi)} = \sqrt{\langle u,u\rangle_{L^2(\pi)}}$ denote the corresponding Euclidean norm.

Consider any vector $w \in \mathbb{R}^\Omega$. Let $\lambda_1 = 1, \lambda_2, \ldots, \lambda_n$ denote the (left) eigenvalues of P arranged so that $1 = |\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$. Since P is self-adjoint with respect to $L^2(\pi)$, we can choose an $L^2(\pi)$-orthonormal basis $v^{(1)}, v^{(2)}, \ldots, v^{(n)}$ of corresponding eigenvectors with $v^{(1)}$ a multiple of π.

Recall that from the above reasoning, we have $|\lambda_i| < 1$ for $i > 1$. Write $w = \sum_{i=1}^n \alpha_i v^{(i)}$, and note that for any $t \geqslant 0$,

$$wP^t = \alpha_1 v^{(1)} + \sum_{i\geqslant 2} \lambda_i^t \alpha_i v^{(i)}\,.$$

In particular, we have

$$\left\| wP^t - \alpha_1 v^{(1)} \right\|_{L^2(\pi)}^2 = \sum_{i \geqslant 2} \lambda_i^{2t} |\alpha_i|^2 \leqslant \lambda_2^{2t} \|w\|_{L^2(\pi)}^2 . \tag{12.5}$$

Since $|\lambda_2| < 1$, this implies that $\|wP^t - \alpha_1 v^{(1)}\|_{L^2(\pi)} \to 0$ as $t \to \infty$, showing that $wP^t \to \alpha_1 v^{(1)}$, where we recall that $v^{(1)}$ is a multiple of π.

Note that if w has all non-negative entries, then since P is an averaging operator, we have $\|wP^t\|_1 = \|w\|_1$, hence $wP^t \to \|w\|_1 \pi$. Finally, observe that if $v = e_x$ then this implies $e_x P^t \to \pi$, which is exactly the claim of Theorem 12.1 (in the reversible case). For later use, we note the following consequence of (12.5): If $x \in \Omega$, then

$$\left\| e_x P^t - \pi \right\|_{L^2(\pi)}^2 \leqslant \lambda_2^{2t} \pi(x). \tag{12.6}$$

12.3 Mixing times

Now we have seen that any irreducible, aperiodic Markov chain P on a finite state space Ω converges to a unique stationary measure π. We are not only concerned with convergence, but also the rate of convergence—we would like to be able to sample efficiently from π.

To this end, we first introduce a metric on the space of probability measures on Ω: For any two measures μ and ν on Ω, the *total variation distance* is defined by

$$d_{TV}(\mu, \nu) \stackrel{\text{def}}{=} \frac{1}{2} \|\mu - \nu\|_1 = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| .$$

As an exercise, one can also show that $d_{TV}(\mu, \nu) = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|$.

For simplicity of notation, let us define $p_t^{(x)}$ to be the distribution given by $P^t e_x$ (i.e., the distribution of the chain started at x after t steps). For any $t \geqslant 0$ and $x \in \Omega$, define the quantity $\Delta_x(t) = d_{TV}(\pi, p_t^{(x)})$, and we set $\Delta(t) = \max_{x \in \Omega} \Delta_x(t)$. For $\varepsilon > 0$, we denote

$$\tau(\varepsilon) = \min\{t : \Delta(t) \leqslant \varepsilon\} .$$

In words, this is the first time t such that, starting from any initial state, the measure of the chain after t steps is within ε of the stationary measure. Finally, by convention, one takes $\tau_{\text{mix}} = \tau(1/2e)$ as the *mixing time* of the Markov chain P. Note that the precise value of ε is not so important; as the following lemma shows, once we have obtained the mixing time, further convergence to stationarity happens very fast.

Lemma 12.5. *For every $t \geqslant 0$, we have*

$$\Delta(t) \leqslant \exp\left(- \left\lfloor \frac{t}{\tau_{\text{mix}}} \right\rfloor \right) .$$

In particular, for every $\varepsilon > 0$, it holds that $\tau(\varepsilon) \leqslant \tau_{\text{mix}} \lceil \ln(1/\varepsilon) \rceil$.

We will not prove this, but it can be done using the coupling characterization of total variation distance that appears in Homework #6. Finally, we can use our proof of Theorem 12.1 in the reversible case to give an upper bound on τ_{mix} in terms of the spectral gap of the chain.

Theorem 12.6. *Let P be a reversible and irreducible, aperiodic Markov chain on the state space Ω. Suppose that P has eigenvalues $1 = \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, and let $\lambda(P) = \max\{|\lambda_2|, |\lambda_n|\}$. Then*

$$\tau_{\text{mix}} \leqslant \left\lceil \frac{1 + \ln(1/\pi_{\min})}{1 - \lambda(P)} \right\rceil,$$

where $\pi_{\min} := \min\{\pi(x) : x \in \Omega\}$.

Proof. Consider $x \in \Omega$ and $\varepsilon > 0$. Recall that

$$d_{TV}(p_t^{(x)}, \pi) = \frac{1}{2} \sum_{y \in \Omega} |P^t(x,y) - \pi(y)| = \frac{1}{2} \sum_{y \in \Omega} \pi(y) \left| \frac{P^t(x,y)}{\pi(y)} - 1 \right| \leqslant \frac{1}{2} \left(\sum_{y \in \Omega} \pi(y) \left| \frac{P^t(x,y)}{\pi(y)} - 1 \right|^2 \right)^{1/2},$$

where the last line uses $\mathbb{E}[X^2] \geqslant (\mathbb{E}[X])^2$. Observe that

$$\begin{aligned} \sum_{y \in \Omega} \pi(y) \left| \frac{P^t(x,y)}{\pi(y)} - 1 \right|^2 \leqslant \frac{1}{\pi_{\min}} \sum_{y \in \Omega} \pi(y)^2 \left| \frac{P^t(x,y)}{\pi(y)} - 1 \right|^2 &= \frac{1}{\pi_{\min}} \sum_{y \in \Omega} \pi(y) \left| P^t(x,y) - \pi(y) \right|^2 \\ &= \frac{1}{\pi_{\min}} \left\| e_x P^t - \pi \right\|_{L^2(\pi)}^2. \end{aligned}$$

Combining the preceding two inequalities with (12.6) gives

$$d_{TV}(p_t^{(x)}, \pi)^2 \leqslant \frac{1}{\pi_{\min}} \left\| e_x P^t - \pi \right\|_{L^2(\pi)}^2 \leqslant \frac{\pi(x)}{\pi_{\min}} \lambda(P)^{2t}$$

Now setting $t = \lceil \frac{1}{1-\lambda(P)} \ln(1/(\varepsilon \pi_{\min})) \rceil$ and using the fact that $(1-\delta)^{1/\delta} \leqslant e^{-1}$ for $\delta > 0$ yields

$$d_{TV}(p_t^{(x)}, \pi)^2 \leqslant \frac{\varepsilon^2}{4},$$

implying $d_{TV}(p_t^{(x)}, \pi) \leqslant \varepsilon/2$. Setting $\varepsilon = 1/e$ and recalling the definition of τ_{mix} yields the desired result. □

Finally, one should note that this bound is essentially tight up to the $O(\log(1/\pi_{\min}))$ factor.

Theorem 12.7. *Under the assumption of Theorem 12.6, we have*

$$\tau_{\text{mix}} \geqslant \frac{1}{1 - \lambda(P)} - 1.$$

Proof. Let v be a (left) eigenvector of P with eigenvalue $\lambda = \lambda(P) \neq 1$. In that case, since π is also an eigenvector of P, we see that v is orthogonal to the stationary measure π, i.e. $\sum_{y \in \Omega} \pi(y) v_y = 0$. It follows that for $t \geqslant 0$ and any $x \in \Omega$,

$$|\lambda^t v_x| = |(vP^t)_x| = \left| \sum_y P^t(x,y) v_y - \pi(y) v_y \right| \leqslant \|v\|_\infty \sum_y |P^t(x,y) - \pi(y)| = 2\|v\|_\infty \, d_{TV}(p_t^{(x)}, \pi).$$

Now choose x so that $|v_x| = \|v\|_\infty$, yielding

$$d_{TV}(p_t^{(x)}, \pi) \geqslant \frac{1}{2} \lambda(P)^t.$$

Therefore $\lambda(P)^{\tau_{\text{mix}}} \leqslant 1/e$, implying that

$$\tau_{\text{mix}} \geqslant \frac{-1}{\log(1 - (1 - \lambda(P)))} \geqslant \frac{1}{1 - \lambda(P)} - 1,$$

where in the final line we have used that $\log(1-a) \geqslant 1 + \frac{1}{a-1}$ for all $a \in [0,1]$. □

So we see that up to a $\log(1/\pi_{\min})$ factor, the spectral gap $1 - \lambda(P)$ controls the mixing time of the chain: If we set $\tau_{\text{rel}} := \frac{1}{1-\lambda(P)}$ (commonly called the "relaxation time" of the chain), then

$$\tau_{\text{rel}} - 1 \leqslant \tau_{\text{mix}} \leqslant O(\log(1/\pi_{\min}))\, \tau_{\text{rel}}.$$

12.4 Some Markov chains

One famous state space is the set of all permutations of n objects (for $n = 52$). In this case, $|\Omega| = n!$. Here are some shuffles:

1. **Random transposition.** At every step, we choose two uniformly random positions i and j (with replacement) and swap the cards at positions i and j.
2. **Top to random.** We take the top card and insert it at one of the n positions in the deck uniformly at random.
3. **Riffle shuffle.** We split the deck into two parts L and R uniformly at random, and then take a uniformly random interleaving of L and R.

And here's a combinatorial example: Let $G = (V, E)$ be a graph with degree at most Δ, and suppose we have q colors with $q \geqslant \Delta + 1$ (so we are assured that G is q-colorable). Let Ω be the set of all q-colorings on G. Here is a natural Markov chain: Suppose we have a proper coloring $\chi : V \to [q]$. We choose a uniformly random $v \in V$ and a uniformly random color $c \in [q]$. If no neighbor of v in χ has color c, then we color v with c. Otherwise, we stay at the current coloring.

This example demonstrates the complex structure of Markov chains on combinatorial state spaces. For what values of q (depending on Δ) is the chain irreducible? It turns out that if $q \geqslant \Delta + 2$, then the chain is always irreducible, and the stationary measure is uniform on proper q-colorings. A huge open problem in MCMC (Markov chain Monte Carlo) is to resolve the following conjecture.

Conjecture 12.8. *For all $q \geqslant \Delta + 2$, this Markov chain has mixing time $O(n \log n)$, where $n = |V|$.*

The best bound (due to Vigoda, 1999) is that this holds for $q \geqslant \frac{11}{6}\Delta$.[1]

13 Eigenvalues, expansion, and rapid mixing

Let P be the transition kernel of a reversible, irreducible, aperiodic Markov chain on the state space Ω. Suppose that P has stationary measure π (this exists and is unique by the Fundamental Theorem of Markov Chains). Let us also assume that all the eigenvalues of P lie in $[0, 1]$. In the last lecture, we proved that they must lie in $[-1, 1]$. Now by replacing P with $P' = \frac{1}{2}I + \frac{1}{2}P$, we can ensure that all eigenvalues are nonnegative while only changing the mixing time by a factor of 2.

Suppose the eigenvalues of P are $1 = \lambda_1 > \lambda_2 \geqslant \cdots \geqslant \lambda_{|\Omega|} \geqslant 0$. In the last lecture, we defined τ_{mix} and showed that

$$\frac{1}{1-\lambda_2} - 1 \leqslant \tau_{\text{mix}} \leqslant O(\log(1/\pi_{\min}))\frac{1}{1-\lambda_2},$$

where $\pi_{\min} := \min\{\pi(x) : x \in \Omega\}$ is the minimum stationary probability. In other words, up to a factor of $O(\log(1/\pi_{\min}))$, the mixing time is controlled by the inverse spectral gap of P.

[1] This year (2019), a group from MIT has improved this bound slightly.

The Gibbs distribution on matchings. To understand the phrase "rapid mixing," let us consider sampling from a particular measure on an exponentially large state space. Fix an n-vertex graph $G = (V, E)$ and consider the set $\mathcal{M}(G)$ of all matchings in G; these are precisely subsets of the edges E in which every vertex has degree at most one. It is clear that $\mathcal{M}(G)$ can be very large; for instance, in the complete graph on $2n$ vertices, we have $\log |\mathcal{M}(G)| \asymp n \log n$.

For a parameter $\lambda \geqslant 1$, let π_λ denote the measure on $\mathcal{M}(G)$ where a matching m has probability proportional to $\lambda^{|m|}$. Here, $|m|$ denotes the number of edges in m. Thus $\pi_\lambda(m) = \lambda^{|m|}/Z$, where

$$Z = \sum_{m \in \mathcal{M}(G)} \lambda^{|m|}$$

is the corresponding *partition function*, which can itself be very difficult to compute. (In fact, the ability to approximate Z efficiently is essentially equivalent to the ability to sample efficiently from π_λ.)

Our goal is to produce a sample from a distribution that is very close to π_λ. To do this, we will define a Markov chain on $\mathcal{M}(G)$ whose stationary distribution is π_λ. We will then show that $\tau_{\mathrm{mix}} \leqslant n^{O(1)}$, implying that there is a polynomial-time algorithm to sample via simulating the chain for $n^{O(1)}$ steps. In general, for such an exponentially large state space indexed by objects of size n, we say that the chain is "rapidly mixing" if the mixing time is at most $n^{O(1)}$.

13.1 Conductance

For a pair of states $x, y \in \Omega$, define $Q(x, y) = \pi(x)P(x, y)$ and note that since P is reversible, the detailed balance conditions give us $Q(x, y) = Q(y, x)$. For two sets $S, T \subseteq \Omega$, define $Q(S, T) = \sum_{x \in S} \sum_{y \in T} Q(x, y)$. Finally, given a subset $A \subseteq \Omega$, we define its *conductance* as the quantity

$$\Phi(A) = \frac{Q(A, \bar{A})}{\pi(A)}\,.$$

Note that $Q(A, \bar{A})$ represents the "ergodic flow" from A to $\bar{A}$—this is the probability of a transition going between A and $\bar{A}$ at stationarity. This quantity has a straightforward operational interpretation: It is precisely the probability that one step of the Markov chain leaves A when we start from the stationary measure restricted to A. Note that if $\Phi(A)$ is small, we expect that the chain might get "trapped" inside A, and thus perhaps such a "bottleneck" could be an obstruction to mixing. In fact, we will see momentarily that this is true, and moreover, these are the only obstructions to rapid mixing.

We define the *conductance of the chain* P to capture the conductance of the "worst" set

$$\Phi^* = \max_{\pi(A) \leqslant \frac{1}{2}} \Phi(A)\,.$$

Then we have the following probabilistic version of the discrete Cheeger inequality (proved independently by Jerrum-Sinclair and Lawler-Sokal in the context of Markov chains on discrete spaces).

Theorem 13.1. *It always holds that*

$$\frac{1}{2}(\Phi^*)^2 \leqslant 1 - \lambda_2 \leqslant 2\,\Phi^*\,.$$

This is a basic fact in spectral graph theory; we will not prove it here. Let us mention, though, that the right-hand side is straightforward—it verifies that indeed a low-conductance set is an obstruction to rapid mixing. The left-hand side, which claims that those are the only such obstructions, is more subtle.

The best way to prove the right-hand side is as follows: Recall the inner product

$$\langle u, v\rangle_{\ell_2(\pi)} = \sum_{x\in\Omega} \pi(x) u_x v_x$$

and the associated Euclidean norm $\|v\|_{\ell_2(\pi)} = \sqrt{\langle v, v\rangle_{\ell_2(\pi)}}$. Then using the variational principle for eigenvalues, we have

$$\lambda_2 = \max_{v:\langle v,\mathbf{1}\rangle_{\ell_2(\pi)}=0} \langle v, vP\rangle\,,$$

where $\mathbf{1}$ denotes the all-ones vector. Consider now any $A \subseteq \Omega$ with $\pi(A) \leqslant \frac{1}{2}$, and define

$$v_x = \begin{cases} \sqrt{\frac{1-\pi(A)}{\pi(A)}} & x \in A \\ -\sqrt{\frac{\pi(A)}{1-\pi(A)}} & x \notin A\,. \end{cases}$$

Note that $\langle v, \mathbf{1}\rangle_{\ell_2(\pi)} = \pi(A)\sqrt{\frac{1-\pi(A)}{\pi(A)}} - (1-\pi(A))\sqrt{\frac{\pi(A)}{1-\pi(A)}} = 0$, and

$$\|v\|^2_{\ell_2(\pi)} = 1 - \pi(A) + \pi(A) = 1\,.$$

Therefore

$$1 - \lambda_2 = \langle v, v(I-P)\rangle_{\ell_2(\pi)} = \frac{1}{2}\sum_{x,y} Q(x,y)(v_x - v_y)^2\,,$$

where the last equality is the usual one we have done with Laplacian matrices (like $I - P$) in preceding lectures. But the latter quantity is precisely

$$Q(A,\bar{A})\left(\sqrt{\frac{1-\pi(A)}{\pi(A)}} + \sqrt{\frac{\pi(A)}{1-\pi(A)}}\right)^2 \leqslant 2\frac{Q(A,\bar{A})}{\pi(A)} = 2\Phi(A)\,,$$

where the inequality uses the fact that $\pi(A) \leqslant \frac{1}{2}$.

13.2 Multi-commodity flows

Although Theorem 13.1 gives a nice characterization of rapid mixing in terms of conductance, the quantity Φ^* is NP-hard to compute, and can be difficult to get a handle on for explicit chains. Thus we now present another connection between conductance and multi-commodity flows.

We consider a multi-commodity flow instance on a graph with vertices corresponding to states Ω and edges $\{x, y\}$ with capacity $Q(x, y)$. The demand between x and y is $\pi(x)\pi(y)$. Let C^* be the optimal congestion that can be achieved by a multi-commodity flow satisfying all the demands (recalling that the congestion of an edge in a given flow is the ratio of the total flow over the edge to its capacity).

Theorem 13.2. *It holds that*

$$\frac{1}{2C^*} \leqslant \Phi^* \leqslant \frac{1}{C^*} O(\log|\Omega|)\,.$$

The right-hand side is due to Leighton and Rao (1988). We will only need the much simpler left-hand side inequality which can be proved as follows. Suppose there exists a flow achieving congestion C and consider some $A \subseteq \Omega$. Then

$$C \cdot Q(A, \bar{A}) \geqslant \pi(A)\pi(\bar{A}) .$$

This is because the left-hand side represents an upper bound on the total flow going across the cut—$Q(A, \bar{A})$ is the capacity across the cut $(A, \bar{A})$, and we have to rescale by C to account for the congestion. On the other hand, $\pi(A)\pi(\bar{A})$ represents the amount of flow that must be traveling across the cut to satisfy the demand. If $\pi(A) \leqslant \frac{1}{2}$, we conclude that

$$Q(A, \bar{A}) \geqslant \frac{\pi(A)\pi(\bar{A})}{C} \geqslant \frac{\pi(A)}{2C} ,$$

completing the proof.

Remark 13.3 (Proof sketch of RHS of Theorem 13.2). (This is related to HW#4(c), which would give the worse bound $O((\log|\Omega|)^2)$.) If we use linear programming duality to characterize C^*, it has the following dual representation:

$$\frac{1}{C^*} = \min_d \frac{\sum_{\{x,y\}} Q(x,y)d(x,y)}{\sum_{x,y\in\Omega} \pi(x)\pi(y)d(x,y)}, \tag{13.1}$$

where the minimum is over all symmetric distance functions $d(x, y)$ on $\Omega \times \Omega$ that satisfy the triangle inequality $d(x, y) \leqslant d(x, z) + d(z, y)$ for all $x, y, z \in \Omega$.

Recall that every finite metric space (X, d) admits a mapping $F : X \to \mathbb{R}^n$ with distortion $D \leqslant O(\log n)$, i.e.,

$$\frac{d(x,y)}{D} \leqslant \|F(x) - F(y)\|_2 \leqslant d(x,y) \qquad x, y \in X.$$

Now let us decompose the Euclidean distance on $\mathbb{R}^n$ into a convex combination over cuts. First, note that for any $a, b \in \mathbb{R}$, we have

$$|a - b| = \int_{-\infty}^{\infty} |\chi_s(a) - \chi_s(b)|\, ds,$$

where $\chi_s := \mathbf{1}_{(-\infty,s]}$. In other words, $\chi_s(a) = 1$ if $a \leqslant s$ and $\chi_s(a) = 0$ otherwise.

Let $\boldsymbol{g}$ denote a random n-dimensional Gaussian vector, i.e., $\boldsymbol{g} = (g_1, \ldots, g_n)$ where $\{g_i\}$ are i.i.d. $N(0, 1)$ random variables. Recall that for $u, v \in \mathbb{R}^n$, we have $\|u - v\|_2^2 = \mathbb{E}[\langle u - v, \boldsymbol{g}\rangle^2]$, because $\langle u - v, \boldsymbol{g}\rangle$ is an $N(0, \|u - v\|_2^2)$ random variable (by the 2-stability property of normal random variables). One can also calculate: If g_0 is an arbitrary normal random variable with mean zero, then

$$\mathbb{E}[|g_0|] = \sqrt{\frac{2}{\pi}}\sqrt{\mathbb{E}[g_0^2]}.$$

Therefore:

$$\|u - v\|_2 = \sqrt{\mathbb{E}[\langle u - v, \boldsymbol{g}\rangle^2]} = \sqrt{\frac{\pi}{2}}\, \mathbb{E}\left[|\langle u - v, \boldsymbol{g}\rangle|\right]$$

We thus arrive at the following "cut decomposition" for all of $\mathbb{R}^n$:

$$\|u - v\|_2 = \sqrt{\frac{\pi}{2}}\, \mathop{\mathbb{E}}_{\boldsymbol{g}} \left[\int_{-\infty}^{\infty} \left|\chi_s(\langle u, \boldsymbol{g}\rangle) - \chi_s(\langle v, \boldsymbol{g}\rangle)\right| \, ds\right]$$

Suppose now that d is the optimal metric in (13.1) and let $F : \Omega \to \mathbb{R}^n$ denote a distortion $D \leqslant O(\log n)$ embedding. The distortion condition yields

$$\frac{1}{C^*} \geqslant \frac{1}{D} \frac{\sum_{\{x,y\}} Q(x,y)\|F(x)-F(y)\|_2}{\sum_{x,y} \pi(x)\pi(y)\|F(x)-F(y)\|_2} = \frac{\mathbb{E}_{\boldsymbol{g}}\left[\int_{-\infty}^{\infty} \sum_{\{x,y\}} Q(x,y)\left|\chi_s(\langle F(x),\boldsymbol{g}\rangle) - \chi_s(\langle F(y),\boldsymbol{g}\rangle)\right| \, ds\right]}{\mathbb{E}_{\boldsymbol{g}}\left[\int_{-\infty}^{\infty} \sum_{x,y} \pi(x)\pi(y)\left|\chi_s(\langle F(x),\boldsymbol{g}\rangle) - \chi_s(\langle F(y),\boldsymbol{g}\rangle)\right| \, ds\right]}$$

Finally, we observe that

$$\frac{\int f(x)\,dx}{\int g(x)\,dx} \geqslant \min_x \frac{f(x)}{g(x)}.$$

Thus there exists some choice of $g \in \mathbb{R}^n$ and $s \in \mathbb{R}$ such that

$$\frac{1}{C^*} \geqslant \frac{1}{D} \frac{\sum_{\{x,y\}} Q(x,y)|\chi_s(\langle F(x),g\rangle) - \chi_s(\langle F(y),g\rangle)|}{\sum_{x,y} \pi(x)\pi(y)|\chi_s(\langle F(x),g\rangle) - \chi_s(\langle F(y),g\rangle)|},$$

but the latter ratio is precisely $\frac{1}{D}\frac{Q(A,\bar{A})}{\pi(A)\pi(\bar{A})}$ for the set $A = \{x \in \Omega : \langle F(x), g\rangle \leqslant s\}$, hence

$$\frac{1}{C^*} \geqslant \frac{1}{D}\frac{Q(A,\bar{A})}{\pi(A)\pi(\bar{A})} \geqslant \frac{1}{2D}\frac{Q(A,\bar{A})}{\min(\pi(A),\pi(\bar{A}))} \geqslant \frac{\Phi^*}{2D},$$

verifying the RHS of Theorem 13.2.

13.3 The Gibbs sampler

Recall now that our goal is to sample from the Gibbs measure π_λ introduced earlier. The following Markov chain is due to Jerrum and Sinclair. If we are currently at a matching $m \in \mathcal{M}(G)$, we define our local transition as follows.

1. With probability 1/2, we stay at m.
2. Otherwise, choose an edge $e = \{u, v\} \in E(G)$ uniformly at random and:
   (a) If both u and v are unmatched in m, set $m := m \cup \{e\}$.
   (b) If $e \in m$, then with probability $1/\lambda$, put $m := m \setminus \{e\}$, and otherwise stay at m.
   (c) If exactly one of u or v is matched in m, then let e' be the unique edge that contains one of u or v and put $m := m \setminus \{e'\} \cup \{e\}$.
   (d) If both u and v are matched, stay at m.

Exercise: Show that this chain is reversible with respect to the measure π_λ.

Now we would like to prove that this chain is rapid mixing by giving a low-congestion multi-commodity flow in the corresponding graph. In fact, we will give an "integral flow," i.e. we will specify for every pair of matchings $x, y \in \mathcal{M}(G)$, a path γ_{xy}.

To do this, consider the edges of x to be colored red and the edges of y to be covered blue. Then the colored union $x \cup y$ is a multi-graph where every node has degree at most 2. It is easy to see that every such graph breaks into a disjoint union of paths and even-length cycles. (Note also the trivial cycles of length two when x and y share an edge.)

The path γ_{xy} will "fix" each of these components one at a time (in some arbitrary order). The trivial cycles are already fine (we don't have to move those edges). To explain how to handle the

path components, we look at a simple example. Suppose the path is $e_1, e_2, e_3, e_4, e_5, e_6$. Then we define a path from the red matching to blue the matching (in this component as follows):

$$e_1, e_3, e_5 \to e_3, e_5 \to e_2, e_5 \to e_2, e_4 \to e_2, e_4, e_6 .$$

Note that each transition is a valid step of the chain. We can do a similar thing for a cycle by first deleting a red edge so that it becomes a path.

Congestion analysis. So now we have given a path γ_{xy} between every pair of states $x, y \in \mathcal{M}(G)$. In the flow, this path should have flow value $\pi_\lambda(x)\pi_\lambda(y)$ so that is satisfies the corresponding demand. We are left to analyze the weight of paths that use a given "edge" (a transition) of the chain. The interested reader is referred to the beautiful argument at its original source [**?**].