Randomness & Computation: CS 271

Professor Alistair Sinclair
Notes by Albert Zhang

Spring 2020

Contents

1 First Moment Method

In many situations where we want to show the existence of an object with some desired property, it may be easier to show that

$$\mathbb{P}[X \text{ has some property}] > 0,$$

which would imply that there exists some point in the probability space which has the property. If, say, our property is something of the form $\{X \geq x\}$, then it also suffices to show

$$\mathbb{E}[X] \geq x,$$

so that at least one sample point must have value $\geq x$. We may also have a sequence of random variables $\{X_n\}$, and we wish to show that the probability of some "bad event" $\mathcal{B}_n$ occurs with probability tending to 0. If X is a nonnegative, discrete/integer-valued random variable, we may apply Markov's inequality to get

$$\mathbb{P}[\mathcal{B}_n] = \mathbb{P}[X_n > x] \leq \frac{\mathbb{E}[X_n]}{x},$$

where we want to show that $\mathbb{E}[X_n]/x \to 0$.

This is known as the *probabilistic method*, and more generally falls into the class of *first moment methods*. Note that we have shown existence without having ever constructed the object explicitly. In many applications, we may want to find an explicit construction–we deal with this in Section 1.6.

1.1 Ramsey Theory

Definition 1.1. The k-th (diagonal) Ramsey number $R_k = R_{k,k}$ is the smallest number n such that any 2-coloring of the edges of the complete graph K_n must contain a monochromatic k-clique.

It has been shown that $R_3 = 6$ and $R_4 = 18$. Surprisingly for R_5, we only know it lies in the interval $[43, 49]$. In general, for larger Ramsey numbers, we only have rather course bounds. In fact, there's a comical quote by Erdős saying that if aliens were to threaten to invade earth unless we solved R_5, we should marshal the world's resources towards computing it. However, if it were R_6, we should instead marshal the world's resources towards a preemptive military attack.

Theorem 1.2

By the probabilistic method, we may show the following lower bound:

$$R_k > 2^{k/2}.$$

Proof. It suffices to show that for $n = 2^{k/2}$, there exists a 2-coloring which does not contain a monochromatic k-clique. Consider the model $G \sim \mathcal{G}(n, p)$, where we take

$p = 1/2$. Then given any k-clique C in G, we have

$$\mathbb{P}\left[C \text{ is monochromatic}\right] = 2 \cdot \left(\frac{1}{2}\right)^{\binom{k}{2}}.$$

Therefore, since the total number of k-cliques in G is $\binom{n}{k}$, we get by a union bound

$$\begin{aligned}
\mathbb{P}\left[G \text{ has a monochromatic } k\text{-clique}\right] &\leq \binom{n}{k} 2^{1-\binom{k}{2}} \\
&\leq \frac{n^k}{k!} \cdot 2^{1-\binom{k}{2}} \\
&= \frac{2^{\frac{k^2}{2}+1-\frac{k^2-k}{2}}}{k!} \\
&= \frac{2^{1+\frac{k}{2}}}{k!} \\
&< 1,
\end{aligned}$$

for $k \geq 3$. Thus there must exist a point in the probability space which has no monochromatic k-cliques, given $n = 2^{k/2}$. $\square$

It turns out the this lower bound is essentially the best known, in the sense that no bounds of the form $R_k \geq 2^{(1/2+\epsilon)k}$ or $R_k \leq 2^{(2-\epsilon)k}$ have been found.

1.2 Max Cut

Recall the Min-Cut problem, which can be solved as the dual to the Max Flow problem efficiently. On the other hand, we have the NP-hard Max Cut problem, in which we want to find the partition such that the number of cut edges is maximimzed.

Lemma 1.3

Given a graph $G = (V, E)$, there exists a cut containing at least $|E|/2$ edges.

Proof. Let $V_1 \cup V_2 = V$ be our partition. Assign each vertex to V_1 and V_2 with probability $1/2$. Define the random variable $X = \sum_{e \in E} X_e$ as the sum of indicators X_e determining whether the edge e is in the cut or not. Then

$$\mathbb{E}\left[X\right] = \sum_{e \in E} \mathbb{E}\left[X_e\right] = \frac{|E|}{2}.$$

Therefore, there must exist a partition such that the number of edges crossing the cut $X \geq |E|/2$. $\square$

1.3 Independent Set

Given a graph $G = (V, E)$, a subset $U \subset V$ is said to be an *independent set* if no two vertices $u_1, u_2 \in U$ are adjacent in G. The problem of determining the size of the largest independent set is NP-hard. However, we can achieve a good lower bound.

Theorem 1.4

Given a graph $G = (V, E)$, the size of the largest independent set V' is at least

$$|V'| \geq \sum_{v \in V} \frac{1}{\deg(v) + 1}.$$

Proof. To each vertex v, assign a weight $w_v \sim \mathrm{Unif}([0, 1])$. Call v a *local minimum* if $w_v < w_u$ for all neighbors u of v. Then clearly no two adjacent vertices can both be local minima (or at least, such an event has measure zero). Therefore the set of local minima forms an independent set. Furthermore, for each vertex v, we have

$$\mathbb{P}\left[v \text{ is a local minimum}\right] = \frac{1}{\deg(v) + 1},$$

so by linearity, we get

$$\mathbb{E}\left[X\right] = \sum_{v \in V} \mathbb{E}\left[X_v\right] = \sum_{v \in V} \frac{1}{\deg(v) + 1}.$$

Hence there must exist an independent set at least this size. □

1.4 Graph Crossing Number

Given a graph $G = (V, E)$, with $n = |V|$ and $m = |E|$, define the *crossing number* $c(G)$ as the minimum number of edge crossings in any planar embedding of G. So, a graph is planar if and only if $c(G) = 0$.

Note that by Euler's formula, if a graph is planar then

$$m \leq 3n = 6.$$

And so if a graph can be embedded in the plane without crossing edges, it must necessarily be quite sparse.

Lemma 1.5

For any graph G with n vertices and m edges, we have

$$c(G) \geq m - 3n + 6,$$

which generalizes Euler's formula.

Proof. The proof is purely deterministic. Consider the optimal embedding of G that achieves $c = c(G)$ edge crossings. Then this embedding must satisfy

1. No edge crosses itself.
2. No two edges cross more than once.
3. No two edges which share a vertex cross.

Now, construct a new graph $G' = (V', E')$ from G by inserting a vertex at each edge crossing. Note that the resulting graph is planar, so must satisfy Euler's formula. In particular, we have

$$\begin{aligned} m' &\leq 3n' - 6 \\ m + 2c &\leq 3(n + c) - 6 \\ c &\geq m - 3n + 6, \end{aligned}$$

where we have used the substitutions $m' = m + 2c$ (each edge crossing creates 2 new edges), and $n' = n + c$ (each edge crossing inserts 1 new vertex). □

The above result turns out to be reasonably tight for sparse graphs, where m is not much larger than $3n$. For denser graphs, where m is larger, we have the stronger lower bound using the probabilistic method:

Theorem 1.6

For any graph G with n vertices and m edges, where $m \geq 4n$, we have

$$c(G) \geq \frac{m^3}{64n^2}.$$

Proof. Consider an optimal planar embedding of G with $c = c(G)$ edge crossings. Now generate a random induced subgraph G_p of G by keeping each vertex with probability p, and keeping each edge both of whose endpoints are kept in G_p. The value of p will be optimized over later.

Denote by c_p, n_p, and m_p the respective quantities of G_p corresponding to those of G. Then by Lemma 1.5, we have

$$c_p \geq m_p - 3n_p + 6.$$

Taking expectations, we get

$$\mathbb{E}[c_p] \geq \mathbb{E}[m_p - 3n_p + 6] \geq \mathbb{E}[m_p] - 3\mathbb{E}[n_p].$$

Now, each crossing survives with probability p^4, each edge survives with probability p^2, and each vertex survives with probability p. Therefore the inequality becomes

$$cp^4 \geq mp^2 - 3np.$$

From this we get

$$c \geq \frac{m}{p^2} - \frac{3n}{p^3}.$$

Optimizing over p, we will set $p = 4n/m$ to yield

$$c \geq \frac{m^3}{64n^2},$$

as desired. □

1.5 Sample & Modify

In previous examples, we simply constructed a random object and computed first moments to show that some property exists in the sample space. We now provide two more sophisticated examples where we start with a randomized construction, and supplement it with a deterministic modification to ensure existence of the desired property.

1.5.1 Unbalancing Lights

In this example, we consider a square $n \times n$ array of lights, and a set of row and column switches. The n row switches each toggle the lights in one of the rows, and similarly for the column switches.

Note that naively, we can flip all the switches independently and u.a.r. Then each light will be on with probability $1/2$, and the states will all be pairwise independent. In particular, $\mathbb{E}[X] = \frac{n^2}{2}$, and $\mathrm{Var}(X) = \frac{n^2}{4}$, so the difference $|\#\text{on} - \#\text{off}|$ would be $\Omega(n)$. Thus there exists a setting of the switches which achieves $\frac{n^2}{2} + \Omega(n)$ lights on. We will now show using the probabilistic method, along with a deterministic modification, that we can do better.

Theorem 1.7

For any initial configuration of the lights, there exists a setting of the switches such that the number of lights on X is asymptotically $\Omega\left(\frac{n^2}{2} + \sqrt{\frac{1}{2\pi}} \cdot n^{3/2}\right)$ as $n \to \infty$.

Proof. First set the column switches randomly and independently. Define the indicator X_{ij} to be 1 if light (i, j) is on and -1 if off. Define, for row i, the variable $Z_i = \sum_j X_{ij}$. Due to our random flipping of the columns, the lights in a given row are i.i.d., so that by a CLT type result, we have

$$\mathbb{E}\left[|Z_i|\right] \sim \sqrt{\frac{2n}{\pi}}.$$

Now, we deterministically flip each row so as to get the majority of lights on. By linearity, we have

$$\mathbb{E}\left[\#\text{on} - \#\text{off}\right] = \sum_{i=1}^{n} \mathbb{E}\left[|Z_i|\right] \sim \frac{2}{\pi} \cdot n^{3/2}.$$

Then there must exist some setting of switches which achieves this difference, so that as $n \to \infty$, the number of lights on will be asymptotically

$$\frac{n^2}{2} + \sqrt{\frac{1}{2\pi}} \cdot n^{3/2}.$$

□

1.5.2 Large Girth & Chromatic Number

Given a graph $G = (V, E)$, we define the *girth* of G to be the length of the shortest cycle in G, and the *chromatic number* of G to be the smallest number of colors needed to color the graph so that no two adjacent vertices are of the same color. Intuitively, it makes sense that girth and chromatic number of inversely related, and that graphs with large girth should have small chromatic number. However, the following result by Erdős says this is not the case.

Theorem 1.8

For any positive integers k and l, there exists a graph with girth $\geq l$ and chromatic number $\geq k$.

Proof. Consider the model $G \sim \mathcal{G}(n, p)$. We will pick $p = n^{\frac{1}{l}-1}$, for reasons we will see later.

Denote by X the number of cycles of length less than l in G. Then we have

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=3}^{l-1} \binom{n}{3} \cdot \frac{i!}{2i} \cdot p^i \\ &\leq \sum_{i=3}^{l-1} \frac{n^{i/l}}{2i} \\ &= o(n), \end{aligned}$$

where the first line follows since $\binom{n}{i} \cdot \frac{i!}{2i}$ is the number of possible cycles of length i, and the second line follows from plugging in our choice of p. Therefore, by Markov, we get

$$\mathbb{P}\left[X \geq n/2\right] = o(1).$$

Now, for the chromatic number, note that

$$\text{chromatic } \# \geq \frac{|V|}{|\text{max independent set}|},$$

since the set of vertices that receive any given color is an independent set. Let Y be the size of the maximal independent set. Then by union bound,

$$\begin{aligned} \mathbb{P}\left[Y \geq y\right] &\leq \binom{n}{y} (1-p)^{\binom{y}{2}} \\ &\leq \left(ne^{-p(y-1)/2}\right)^y, \end{aligned}$$

which is $o(1)$ by setting $y = \frac{3}{p}\ln n$. Note that we used the inequalities $\binom{n}{y} \leq n^y$ and $1 + x \leq e^x$.

Together, these results say that we can take n large enough so that both $\mathbb{P}[X \geq n/2]$ and $\mathbb{P}[Y \geq \frac{3}{p}\ln n]$ are less than $1/2$. So by union bound, there exists a graph G with at most $n/2$ cycles of length $< l$, and containing a max independent set of size $< \frac{3}{p}\ln n$. Now, modify G by removing one vertex from each cycle of length at most l, and we get a graph G' satisfying

1. G' has girth $\geq l$,
2. G' has $\geq n/2$ vertices,
3. G' has chromatic number $\geq k$,

for n large enough. □

1.6 Construction

So far, the probabilistic method has only allowed us to prove the existence of an object, without giving us the object itself. In this section, we go over a simple example, discuss how to make the method algorithmic, and then how to derandomize it.

1.6.1 MAX3SAT

In the MAX3SAT problem we are given a boolean formula φ in 3CNF (conjunctive normal form) on variables $\{x_i\}_{1\leq i\leq n}$ and clauses $\{C_i\}_{1\leq i\leq m}$. We want to find the maximum number of clauses that can be satisfied with any assignment of T/F to the variables. This is an NP-hard optimization problem.

Theorem 1.9

For any formula φ, there exists an assignment satisfying at least $\frac{7m}{8}$ clauses.

Proof. Assign T/F to each variable with probability $1/2$ independently. Let X be the number of satisfied clauses in a random assignment. Then a simple argument by indicators (one for each clause C_i) gives

$$\mathbb{E}[X] = \sum_{i=1}^{m} \mathbb{E}[X_i] = \sum_{i=1}^{m} \frac{7}{8} = \frac{7m}{8}.$$

Since there must exist a point in the sample space achieving this, we are done. □

1.6.2 Monte Carlo Approach

Naively, we can directly apply the randomized construction by simply picking a random assignment and keep resampling until it satisfies a sufficient threshold of clauses. To analyze the behavior, we use Markov's inequality.

Lemma 1.10

Let X be the random variable from the proof of Theorem 1.9, i.e. the number of satisfied clauses in a random assignment. Then

$$\mathbb{P}\left[X \geq \frac{7}{8}m\right] \geq \frac{1}{m+1}.$$

Proof. First, we apply Markov to the random variable $m - X$ to get

$$\mathbb{P}\left[X \leq \left(1 - \frac{\alpha}{8}\right) m\right] = \mathbb{P}\left[m - X \geq \frac{\alpha}{8}m\right] \leq \frac{m - \mathbb{E}[X]}{\frac{\alpha}{8}m} = \frac{1}{\alpha}.$$

Now, let $\alpha = 1 + \frac{1}{m}$, which gives us

$$\begin{aligned}\mathbb{P}\left[X < \frac{7}{8}m\right] &= \mathbb{P}\left[X < \left\lfloor\frac{7}{8}m\right\rfloor\right] \\ &= \mathbb{P}\left[X \leq \frac{7}{8}m - \frac{1}{8}\right] \\ &= \mathbb{P}\left[X \leq \left(1 - \frac{\alpha}{8}\right) m\right] \\ &\leq \frac{1}{\alpha} \\ &= \frac{m}{m+1}.\end{aligned}$$

Note that we've used the crucial fact that X is integer-valued. Thus we have

$$\mathbb{P}\left[X \geq \frac{7}{8}m\right] \geq 1 - \frac{m}{m+1} = \frac{1}{m+1}.$$

□

Theorem 1.11

We can find an assignment satisfying at least $\frac{7}{8}m$ clauses in polynomial time with high probability.

Proof. By the lemma, we see that the Bernoulli random variable Z for producing an assignment satisfying at least $\frac{7}{8}m$ clauses stochastically dominates a geometric random variable with parameter $\frac{1}{m+1}$. Therefore in polynomial time we can achieve the expectation $\frac{7}{8}m$ with high probability. □

1.6.3 Method of Conditional Probabilities

Depending on the situation, we may be able to derandomize the random construction used in the probabilistic method, achieving the expected value or object with desired property deterministically.

For example, let's think of our random assignment of the variables $\{x_i\}$ of our 3CNF formula φ in a sequential fashion. First pick a T/F value for x_1, then for x_2, and so on. This process can be illustrated as a tree. We label each node of the tree with a formula Ψ, and denote by X_Ψ the number of clauses that are satisfed in the tree below Ψ given the fixed assignments of the variables above that node. So for instance, the root is just φ, with no variables fixed yet. The random variable X_0 is just X. The second level of the tree we have two nodes $\Psi_1 = \phi|_{x_1=T}$ and $\Psi_2 = \phi|_{x_1=F}$. The random variable X_1 counts the number of clauses that will be satisfied with a random assignment of variables $\{x_2, x_3, \ldots, x_n\}$, and likewise for X_2.

Note that we have

$$\mathbb{E}[X_\Psi] = \mathbb{P}\left[x_{i+1} = T\right] \cdot \mathbb{E}[X_{\Psi|_{x_{i+1}=T}}] + \mathbb{P}\left[x_{i+1} = F\right] \cdot \mathbb{E}[X_{\Psi|_{x_{i+1}=F}}] = \frac{1}{2}(\mathbb{E}[X_{\Psi_1}] + \mathbb{E}[X_{\Psi_2}])$$

where Ψ_1 and Ψ_2 are the children of Ψ here. Then at least one child must have expectation at least as large as $\frac{7}{8}m$. Since at the root, we started with $\mathbb{E}[X_\varphi] \geq \frac{7}{8}m$, there will be a leaf node with expectation at least $\frac{7}{8}m$. Furthermore, given a fixed assignment to some subset of the variables, we can explicitly compute $\mathbb{E}[X_\Psi]$, so that we can traverse down the tree in linear time to find the desired assignment.

This method will work whenever we can sequentially index our random choices and we have the ability to compute the conditional expectations when some of the random choices have already been made. Or we can compute the expectations approximately and proceed as before to obtain a final result that approximates the desired expectation.

2 Second Moment Method

In first moment method scenarios, we may be given a sequence of random variables $\{X_n\}$, and we wish to show that in the limit, the probability of some property goes to 0. It is much the same for the second moment method, although we are now given the ability to compute second moments as well. Intuitively speaking, second moments are usually harder to compute than first moments, so naively we should always attempt to use first moment bounds. But, there will be situations where these bounds are too weak, and we will necessarily have to turn to higher moments.

To see that in general, having access to higher moments gives us more power, consider the following example.

Example 2.1. Define the collection of random variables

$$X_n = \begin{cases} n^2 & \text{w.p. } 1/n \\ 0 & \text{o.w.} \end{cases}$$

Then $\mathbb{E}[X_n] = n \to \infty$, however $\mathbb{P}[X_n > 0] \to 0$. Thus any first moment techniques will fail here.

First, we list some of the basic tools of second moment methods. Recall the classical inequality:

Theorem 2.2 (Chebyshev's Inequality)
Let X be any random variable. Then

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

As immediate corollaries, we get

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \beta\mathbb{E}[X]\right] \leq \frac{\mathrm{Var}(X)}{\beta^2\mathbb{E}[X]^2}, \tag{1}$$

as well as

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \beta\sigma\right] \leq \frac{1}{\beta^2}, \tag{2}$$

where $\sigma = \sqrt{\mathrm{Var}(X)}$ is the standard deviation of X.

In many applications, we will set $\beta = 1$ in equation 1 to obtain:

Lemma 2.3
For a nonnegative, discrete random variable X,

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

In some situations, the vanilla inequality above might not be enough. As such, the following application of Cauchy-Schwarz provides an improved variant of the second moment method:

Theorem 2.4 (Paley-Zygmund Inequality)
Let X be a nonnegative random variable. For $0 < \theta < 1$,

$$\mathbb{P}\left[X \geq \theta \mathbb{E}[X]\right] \geq (1-\theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{3}$$

Proof. We have

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

where in the second line we have used Cauchy-Schwarz to obtain the second term. Rearranging the inequality gives the result. □

As a corollary, we obtain another variant of the second moment method:

Lemma 2.5
Let X be a nonnegative random variable that is not identically 0. Then

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

Proof. Take $\theta \downarrow 0$ in the Paley-Zygmund inequality. By monotone or dominated convergence of the indicators $\mathbf{1}_{X \geq \theta \mathbb{E}[X]} \uparrow \mathbf{1}_{X>0}$, we see that

$$\mathbb{P}\left[X > 0\right] = \mathbb{E}[\mathbf{1}_{X>0}] \geq \lim_{\theta \to 0} (1-\theta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} = \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

□

Note that since

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

compared to the vanilla second moment 2.3,

$$\mathbb{P}\left[X > 0\right] \geq 1 - \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} \leq 1 - \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2 + \mathrm{Var}(X)},$$

the one deduced from Paley-Zygmund in 2.5 is indeed stronger.

2.1 Thresholds in Random Graphs

Recall the $\mathcal{G}_{n,p}$ model where we sample a graph G of n vertices where each edge is included with probability p. We are concerned with questions such as

- Is $G \in \mathcal{G}_{n,p}$ connected?
- Does $G \in \mathcal{G}_{n,p}$ contain a Hamilton cycle?
- Does $G \in \mathcal{G}_{n,p}$ contain a 4-clique?

It turns out that for properties such as these, there exists a "point" where the answer to these questions transitions from yes to no (or vice versa) as we cross this point. More concretely, we call $p(n)$ a threshold for a property Q if as $n \to \infty$,

$$p \ll p(n) \Rightarrow \mathbb{P}\left[G \in \mathcal{G}_{n,p} \text{ has } Q\right] \to 0,$$
$$p \gg p(n) \Rightarrow \mathbb{P}\left[G \in \mathcal{G}_{n,p} \text{ has } Q\right] \to 1.$$

In this section, we will answer the third question. Let X denote the number of 4-cliques in G. For each subset C of 4 vertices in G, define the indicator X_C. Then we have

$$\mathbb{E}[X] = \sum_C \mathbb{E}[X_C] = \binom{n}{4} p^6 = \Theta(n^4 p^6).$$

Therefore, we see that

- If $p \ll n^{-2/3}$, then $\mathbb{E}[X] \to 0$.
- If $p \gg n^{-2/3}$, then $\mathbb{E}[X] \to \infty$.

Based on this observation, we guess that $p(n) = n^{-2/3}$ is the threshold for 4-cliques. Indeed, using second moment methods, we have the following result:

Theorem 2.6

The value $p(n) = n^{-2/3}$ is a threshold for G containing a 4-clique.

Proof. Let X and X_C be defined as above. The first direction follows easily from Markov. In particular, since X is integer-valued, we have

$$\mathbb{P}\left[X > 0\right] = \mathbb{P}\left[X \geq 1\right] \leq \mathbb{E}[X] \to 0$$

for $p \ll n^{-2/3}$.

For the other direction, note that $\mathbb{E}[X] \to \infty$ is not enough so show that

$$\mathbb{P}\left[G \in \mathcal{G}_{n,p} \text{ has a 4-clique}\right] \to 1,$$

since we could have $X = 0$ half the time, and X growing with n the other half of the time. Therefore we look to apply Lemma 2.3.

First, we compute

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(\sum_C X_C) \\ &= \sum_C \mathrm{Var}(X_C) + \sum_{C \neq D} \mathrm{Cov}(X_C, X_D).\end{aligned}$$

The first term is a sum over $\binom{n}{4}$ Bernoulli random variables, so we have

$$\sum_C \mathrm{Var}(X_C) = \binom{n}{4}(p^6 - p^{12}) = O(n^4 p^6).$$

The second term requires some casework.

- Case 1: $|C \cap D| \leq 1$. In this case X_C and X_D are independent, so $\mathrm{Cov}(X_C, X_D) = 0$.
- Case 2: $|C \cap D| = 2$. In this case we compute

$$\begin{aligned}\mathrm{Cov}(X_C, X_D) &\leq \mathbb{E}[X_C X_D] \\ &= \mathbb{P}\left[C,\ D \text{ are both cliques given } |C \cap D| = 2\right] \\ &= p^{11}.\end{aligned}$$

  Since there are $\binom{n}{6}\binom{6}{2}$ such pairs (C, D), the total contribution of this case is $O(n^6 p^{11})$.
- Case 3: $|C \cap D| = 3$. Here $\mathrm{Cov}(X_C, X_D) \leq p^9$, and there are $\binom{n}{5}\binom{5}{2}$ such pairs. Thus the total contribution of this case is $O(n^5 p^9)$.

Altogether, we get

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

Using the prior computation that $\mathbb{E}[X] = \Theta(n^4 p^6)$, we apply Lemma 2.3 to get

$$\mathbb{P}\left[X = 0\right] \leq \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} = O\left(\frac{1}{n^4 p^6}\right) + O\left(\frac{1}{n^2 p}\right) + O\left(\frac{1}{n^3 p^3}\right),$$

which vanishes to 0 as $n \to \infty$ assuming $p \gg n^{-2/3}$. Thus the probability of G having a 4-clique tends to 1, and this concludes the proof of the theorem. $\square$

Remark. It's possible to generalize the above proof for containment of general k-cliques. In fact, it turns out that we can generalize it to any subgraph H that is *balanced*. We call H balanced if the average degree of H is greater than or equal to the average degree of any induced subgraph of H. In particular, if this is the case, then we would expect the threshold to be $p = n^{-v/e}$, where v and e are the number of vertices and edges of H respectively.

2.2 Clique Number of Random Graphs

Given a graph G, we are concerned with its clique number, the size of a largest clique in G. Finding the clique number is NP-hard. However, if we are given a random graph $G \in \mathcal{G}_{n,p}$, then the clique number is known asymptotically.

Theorem 2.7

For $G \in \mathcal{G}_{n,p}$ and any constant $p \in (0,1)$, the clique number of G is close to $2\log_{1/p} n$ with probability tending to zero (the meaning of "close to" will be clarified in the proof).

Proof. For simplicity, restrict to the case where $p = 1/2$. Define X_k to be the number of k-cliques in a graph sampled from $\mathcal{G}_{n,p}$. Let $k_0(n)$ be the largest value of k such that $g(k) := \mathbb{E}[X_k] = \binom{n}{k} 2^{-\binom{k}{2}} \geq 1$. A calculation shows that $k_0(n) \sim 2\log n$. We will show that for any integer constant c:

1. For $k_1(n) = k_0(n) + c$, $\mathbb{P}\left[X_{k_1(n)} > 0\right] \to 0$ as $n \to \infty$.
2. For $k_2(n) = k_0(n) - c$, $\mathbb{P}\left[X_{k_2(n)} > 0\right] \to 1$ as $n \to \infty$.

Now, to get the behavior of $\mathbb{E}[X_k]$ around $k_0(n) \sim 2\log n$, we observe that

$$\frac{g(k+1)}{g(k)} = \frac{n-k}{k+1} \cdot 2^{-k} \sim \frac{n}{2\log n} \cdot n^{-2} \to 0, \quad n \to \infty,$$

for $k = k_0$. A similar computation shows that the ratio $g(k-1)/g(k)$ goes to ∞. Therefore, in any c-neighborhood of $k_0(n)$, the graph of $g(k)$ decreases sharply as $n \to \infty$. We deduce the following first moment behaviors:

- $\mathbb{E}[X_{k_1(n)}] \to 0$ as $n \to \infty$.
- $\mathbb{E}[X_{k_2(n)}] \to \infty$ as $n \to \infty$.

Claim (1) follows by Markov, since

$$\mathbb{P}\left[X_{k_1(n)} > 0\right] = \mathbb{P}\left[X_{k_1(n)} \geq 1\right] \leq \mathbb{E}[X_{k_1(n)}] \to 0, \quad n \to \infty.$$

For similar reasons as in the previous section, claim (2) requires the second moment method. In particular, by Lemma 2.3,

$$\mathbb{P}\left[X_{k_2(n)} = 0\right] \leq \mathbb{P}\left[|X_{k_2(n)} - \mathbb{E}[X_{k_2(n)}]| \geq \mathbb{E}[X_{k_2(n)}]\right] \leq \frac{\mathrm{Var}(X_{k_2(n)})}{\mathbb{E}[X_{k_2(n)}]^2}.$$

So, it suffices to show that $\frac{\mathrm{Var}(X_{k_2(n)})}{\mathbb{E}[X_{k_2(n)}]^2} \to 0$ as $n \to \infty$. To ease the notation, from now on we write X for $X_{k_2(n)}$, and for every subset S of the vertex set of size $k_2(n)$, we define the indicator X_S, so that $X = \sum_S X_S$. Also write $S \sim T$ if X_S and X_T are not independent–this happens whenever $S \neq T$ and $|S \cap T| \geq 2$. Then we have

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_S \mathrm{Var}(X_S) + \sum_{S\sim T} \mathrm{Cov}(X_S, X_T) \\
&\leq \sum_S \mathbb{E}[X_S^2] + \sum_{S\sim T} \mathbb{E}[X_S X_T] \\
&= \sum_S \mathbb{E}[X_S] + \sum_{S\sim T} \mathbb{E}[X_S X_T] \\
&= \mathbb{E}[X] + \sum_{S\sim T} \mathbb{E}[X_S X_T].
\end{aligned}$$

Therefore, we have

$$\frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} \leq \frac{1}{\mathbb{E}[X]} + \frac{1}{\mathbb{E}[X]^2} \sum_{S\sim T} \mathbb{E}[X_S X_T],$$

so that it suffices to show

$$\sum_{S\sim T} \mathbb{E}[X_S X_T] = o(\mathbb{E}[X]^2).$$

We compute

$$\begin{aligned}
\sum_{S\sim T} \mathbb{E}[X_S X_T] &= \sum_{S\sim T} \mathbb{P}[X_S = 1, X_T = 1] \\
&= \sum_{S\sim T} \mathbb{P}[X_S = 1] \cdot \mathbb{P}[X_T = 1 | X_S = 1] \\
&= \sum_S \mathbb{P}[X_S = 1] \sum_{T:T\sim S} \mathbb{P}[X_T = 1 | X_S = 1] \\
&= \left(\sum_S \mathbb{P}[X_S = 1]\right) \left(\sum_{T:T\sim S_0} \mathbb{P}[X_T = 1 | X_{S_0} = 1]\right) \\
&= \mathbb{E}[X] \sum_{T:T\sim S_0} \mathbb{P}[X_T = 1 | X_{S_0} = 1]
\end{aligned}$$

where we have fixed a S_0 by symmetry. Now, after some counting arguments, we get

$$\begin{aligned}
\frac{\sum_{T:T\sim S_0} \mathbb{P}[X_T = 1 | X_{S_0} = 1]}{\mathbb{E}[X]} &= \frac{\sum_{i=2}^{k_2-1} \binom{k_2}{i}\binom{n-k_2}{k_2-i} 2^{-\left[\binom{k_2}{2} - \binom{i}{2}\right]}}{\binom{n}{k_2} 2^{-\binom{k_2}{2}}} \\
&= \sum_{i=2}^{k_2-1} \frac{\binom{k_2}{i}\binom{n-k_2}{k_2-i} 2^{\binom{i}{2}}}{\binom{n}{k_2}} \\
&= \sum_{i=2}^{k_2-1} f(i) \\
&\leq k_2 \cdot \max_{2\leq i\leq k_2-1} f(i).
\end{aligned}$$

It can be shown (through some nasty analysis) that $f(i)$ is maximized at $i = 2$, so that

$$k_2 f(2) \sim \frac{k_2^5}{n^2} \sim \frac{(2\log n)^5}{n^2} \to 0, \qquad \text{as } n \to \infty.$$

Thus we have shown claim (2), so this concludes the proof of the theorem. $\square$

2.3 Pairwise Independence

Definition 2.8. A collection $\{X_i\}_{i=1}^n$ of discrete random variables over the same probability space is said to be k-wise independent if for every subset $I \subseteq \{1, \dots, n\}$ with $|I| \le k$, and for every set of values $\{a_i\}_{i=1}^n$, we have

$$\mathbb{P}\left[\bigcap_{i\in I} X_i = a_i\right] = \prod_{i\in I} \mathbb{P}[X_i = a_i].$$

We say the collection is mutually independent if they are n-wise independent.

There are many examples of random variables that are pairwise independent but not mutually (or k-wise, for $k > 2$) independent. It could be instructive to try and construct such an example.

Although pairwise independence is a weaker condition than mutual independence, in many applications pairwise independence is good enough for applying second moment methods. The main benefit is computational–it's possible to represent pairwise independence more compactly than mutual independence. Intuitively, this is because there is less randomness, and constructing randomness is costly.

Suppose we have a Monte Carlo algorithm $\mathcal{A}$ with one-sided error probability $\le 1/2$ (it is always correct on 'yes', but wrong on 'no' with probability at most $1/2$). Then we can achieve an error probability of $\le 2^{-t}$ if we use t independent trials. Assuming $\mathcal{A}$ requires m random bits, this implies we'll need $m \log r$ random bits to achieve an error probability of $1/r$. But we can do better.

Theorem 2.9

For any $r \le 2^m$, we can achieve error probability $\le 1/r$ using only $2m$ random bits, and runtime $O(rm)$.

Proof. Since $\mathcal{A}$ requires m random bits, we can represent the possible executions of $\mathcal{A}$ with bit strings from $\{0,1\}^m$. Then pick $r < 2^m$ pairwise independent uniform samples from $\{0,1\}^m$, and let X_i be the outcome of the algorithm on the i^{th} sample:

$$X_i = \begin{cases} 1 & \text{if } \mathcal{A} \text{ outputs yes on } i^{\text{th}} \text{ sample,} \\ 0 & \text{otherwise.} \end{cases}$$

Now, note that since X_i are pairwise independent, we have

$$\mathrm{Var}(X) = \sum_{i=1}^r \mathrm{Var}(X_i).$$

Furthermore, let $\mathbb{E}[X_i] = p$ be our one-sided error. Since each X_i is Bernoulli, we see that

$$\frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} = \frac{rp(1-p)}{(rp)^2} = \frac{1-p}{p} \cdot \frac{1}{r} \leq \frac{1}{r},$$

since $p \geq 1/2$. It follows by Lemma 2.3 that

$$\mathbb{P}[X = 0] \leq \frac{1}{r}.$$

It remains to show how we can achieve a collection of r pairwise independent variables with $2m$ random bits. Let q be a prime such that $2^m < q < 2^{m+1}$. Then pick a, b uniformly at random from the field $\mathbb{Z}_q$. Consider the function $f : \mathbb{Z}_q \to \mathbb{Z}_q$ given by

$$f_{a,b}(x) = ax + b.$$

We will show that the collection

$$\mathcal{B} = \{f_{a,b}(x) : x \in \mathbb{Z}_q\}$$

is a pairwise independent family over $\mathbb{Z}_q$. Note that our collection is indexed by the input x, and not the random integers a and b.

First, note that for all $x, z \in \mathbb{Z}_q$, we have

$$\mathbb{P}_{a,b}[f_{a,b}(x) = z] = \mathbb{P}_{a,b}[ax + b = z] = \frac{1}{q}.$$

Now, for $x \neq y \in \mathbb{Z}_q$, in order to have $f_{a,b}(x) = z_1$ and $f_{a,b}(y) = z_2$, we must satisfy the linear system

$$\begin{bmatrix} x & 1 \\ y & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

which is invertible since the 2×2 is just a Vandermonde, which is invertible. Therefore there is a unique solution for a and b, so that

$$\mathbb{P}_{a,b}[f_{a,b}(x) = z_1, f_{a,b}(y) = z_2] = \frac{1}{q^2}.$$

It follows that the family $\mathcal{B}$ is indeed pairwise independent. Note that we only use $2m$ random bits for the values of a and b, so that this concludes the proof. $\square$

Remark. Note that the above construction can be easily extended to k-wise independence. We just get a $k \times k$ Vandermonde, which is still invertible, and the rest of the proof is largely the same.

Theorem 2.10

Let A be a random $m \times n$ *Toeplitz matrix*, constructed by picking the entries of the first row and first column u.a.r. from $\{0, 1\}$, and then copying its values along each corresponding diagonal. For example, one such matrix might look like this:

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Then the family

$$\mathcal{T} = \{h_{A,b}(x) = Ax + b : x \in \{0, 1\}^n\}$$

consists of pairwise independent uniform random variables over $\{0, 1\}^m$, using only $2m + n$ random bits.

Proof. Left as an exercise. □

Remark. We close off this section with a brief discussion about derandomization using k-wise independent random variables. In a previous section, we talked about the method of conditional probabilities for derandomization. This method is inherently sequential, and hence hard to parallelize. Instead, using k-wise independent families, which can be constructed in polynomial space, we can simply do an exhaustive search through the probability space. This yields a polynomial algorithm that can also be easily parallelized.

3 Chernoff/Hoeffding Bounds

Suppose we have i.i.d. random variables $(X_i)_{1\leq i\leq n}$ with $\mathbb{E}[X_i] = \mu$, and $\mathrm{Var}(X_i) = \sigma^2$.Then the Central limit theorem says that as $n \to \infty$, the variable $\frac{X-n\mu}{\sqrt{n}\sigma}$ approaches a standard normal distribution $\mathcal{N}(0,1)$. As $n \to \infty$, this gives us the approximation

$$\mathbb{P}\left[|X - n\mu| > \beta\sqrt{n}\sigma\right] \to \frac{2}{\sqrt{2\pi}}\int_\beta^\infty e^{-t^2/2}dt \approx \frac{2}{\sqrt{2\pi}\beta}e^{-\beta^2/2}.$$

So, this gives us an "exponential bound" of the tail probability. However, note that this is just an approximation, and might not even give us a valid bound. Furthermore, this result is asymptotic only, and says nothing about the rate of convergence or behavior for finite n.

Chernoff/Hoeffding bounds will deal with these deficiencies. We first deal with the case where the X_i are $\{0,1\}$ valued. This will then generalize to variables which are $[0,1]$ valued, and then $[a,b]$ valued. It turns out that similar results also hold for unbounded random variables provided their distributions vanish quick enough (e.g. geometric random variables).

The following is the general form for 0-1 independent coin flips, though its complexity makes it less frequently used than more refined versions we will develop later.

Theorem 3.1 (Raw Chernoff)
Let $X_1, \ldots, X_n$ be independent 0-1 random variables with $\mathbb{E}[X_i] = p_i$. Let $X = \sum_{i=1}^n$, $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$, and $p = \mu/n$. Then

$$\mathbb{P}[X \geq \mu + \lambda] \leq \exp\left(-nH_p\left(p + \frac{\lambda}{n}\right)\right), \tag{4}$$

$$\mathbb{P}[X \leq \mu - \lambda] \leq \exp\left(-nH_{1-p}\left(1 - p + \frac{\lambda}{n}\right)\right), \tag{5}$$

where $H_p(x) = x\ln\left(\frac{x}{p}\right) + (1-x)\ln\left(\frac{1-x}{1-p}\right)$ is the relative entropy of x with respect to p.

Proof. The bounds in (4) and (5) are symmetric. To see this, replace x with $n - x$. Therefore it suffices to prove the first one.

The general strategy we will use is to exponentiate both sides with a dummy variable t, apply Markov, use independence, evaluate each expectation in the product, then use concavity to replace each distinct p_i with their mean p. Finally, we can optimize over the dummy variable t to achieve our desired bound.

Let $m = \mu + \lambda$. Then we have

$$\begin{aligned}
\mathbb{P}[X \geq m] &= \mathbb{P}[e^{tX} \geq e^{tm}] \quad \text{for any } t > 0 \\
&\leq e^{-tm}\mathbb{E}[e^{tX}] \quad \text{by Markov's inequality} \\
&= e^{-tm}\prod_{i=1}^{m}\mathbb{E}[e^{tX_i}] \quad \text{by independence of } X_i \\
&= e^{-tm}\prod_{i=1}^{m}(e^t p_i + 1 - p_i) \quad \text{by the distributions of the } X_i \\
&\leq e^{-tm}(e^t p + 1 - p)^n \quad \text{by concavity, or AM-GM.}
\end{aligned}$$

At this point, we minimize the RHS over $t > 0$. Some calculus tells us to pick t so that

$$e^t = \frac{m(1-p)}{(n-m)p},$$

which gives

$$\mathbb{P}[X \geq \mu + \lambda] \leq \exp\left(n\ln\left(\frac{m(1-p)}{n-m} + 1 - p\right) - m\ln\left(\frac{m(1-p)}{(n-m)p}\right)\right),$$

which can then massage this into the desired result

$$= \exp\left(-nH_p\left(p + \frac{\lambda}{n}\right)\right).$$

□

Corollary 3.2

Under the same hypotheses of Theorem 3.1, except instead assuming that the X_i's now take values in the interval $[0,1]$, the bounds from 3.1 still hold.

Proof. Suppose we are given a convex function f. Let X be a $\{0,1\}$ valued r.v. and Y be a $[0,1]$ valued r.v. such that $\mathbb{E}[X] = \mathbb{E}[Y] = p$. We will show that

$$\mathbb{E}[f(X)] \geq \mathbb{E}[f(Y)]. \tag{6}$$

By convexity, we have for any $y \in [0,1]$,

$$(1-y)f(0) + yf(1) \geq f(y).$$

Taking expectations of both sides, this becomes

$$(1-p)f(0) + pf(1) \geq \mathbb{E}[f(Y)].$$

Note that the LHS is just $\mathbb{E}[f(X)]$, so we have shown (6).

Now, to complete the proof, apply (6) with the function $f(x) = e^{tx}$ and replace the $=$ with a $\leq$ in the fourth line of the string of equalities/inequalities from the proof of 3.1. □

Here is a more useful version of the bound given in 3.1.

Corollary 3.3

Let X be as before. Then

$$\left.\begin{matrix}\mathbb{P}[X \leq \mu - \lambda] \\ \mathbb{P}[X \geq \mu + \lambda]\end{matrix}\right\} \leq \exp\left(-\frac{2\lambda^2}{n}\right).$$

Proof. The proofs for the upper and lower tails are symmetric. Let $z = \frac{\lambda}{n}$. We may take log's so that it suffices to show that for $0 \leq z \leq 1 - p$,

$$f(z) := (p+z)\ln\left(\frac{p+z}{p}\right) + (1-p-z)\ln\left(\frac{1-p-z}{1-p}\right) - 2z^2 \geq 0.$$

This is an easy exercise in calculus. ☐

Corollary 3.4

For r.v.s X_i taking values in $[a_i, b_i]$, we have the following generalization of 3.3:

$$\left.\begin{matrix}\mathbb{P}[X \leq \mu - \lambda] \\ \mathbb{P}[X \geq \mu + \lambda]\end{matrix}\right\} \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

An alternative corollary, due to Angluin & Valiant, is slightly worse when $\mu \sim O(n)$, but much sharper when $\mu \ll n$.

Corollary 3.5

Let X be as before. Then for $0 < \beta < 1$ we have the lower tail bound:

$$\begin{aligned}\mathbb{P}[X \leq (1-\beta)\mu] &\leq \exp(-\mu(\beta + (1-\beta)\ln(1-\beta))) \\ &\leq \exp\left(-\frac{\beta^2\mu}{2}\right).\end{aligned}$$

For $\beta > 0$ we have the upper tail bounds:

$$\begin{aligned}\mathbb{P}[X \geq (1+\beta)\mu] &\leq \exp(-\mu(-\beta + (1+\beta)\ln(1+\beta))) \\ &\leq \begin{cases}\exp\left(-\frac{\beta^2\mu}{2+\beta}\right) & \beta > 1 \\ \exp\left(-\frac{\beta^2\mu}{3}\right) & 0 < \beta \leq 1.\end{cases}\end{aligned}$$

Proof. The proof follows by plugging in $\lambda = \beta\mu = \beta np$ into the bounds from Theorem 3.1, and then repeatedly massaging it using concavity arguments such as

$$\ln(1-x) \leq -x.$$

☐

3.1 Simple Examples

Example 3.6 (Fair Coin Flips)**.** Suppose we have fair coin flips $X_1, \dots, X_n$ so that $p_i = p = 1/2$. Then the Chernoff bound 3.3 gives us

$$\mathbb{P}[|X - \mu| \geq \lambda] \leq 2e^{-\frac{2\lambda^2}{n}},$$

which for $\lambda = \beta\sigma$, where $\sigma = \sqrt{np(1-p)} = \sqrt{n}/2$ is the standard deviation, gives the deviation bound $2e^{-\beta^2/2}$. Compare this to the asymptotic bound provided by CLT at the beginning of the section.

Example 3.7 (Biased Coin Flips)**.** We will use this later on for the "median trick" for fully polynomial randomized approximation schemes. Suppose $p_i = p = 3/4$. Then by the Chernoff bound 3.5 we get that the probability of at most half the flips coming up heads is

$$\begin{aligned}\mathbb{P}[X \leq n/2] &= \mathbb{P}[X \leq (1 - 1/3)\mu] \\ &\leq \exp(-n/24).\end{aligned}$$

3.2 Randomized Routing

As in this example and many later ones, a primary cue to use Chernoff style bounds is an exponential sized sample space, since inherently this implies we have some built in structure resembling independent coin flips.

In the randomized routing problem, we consider a directed network in the n dimensional hypercube, denoting the vertices of the cube by bitstrings in $\{0,1\}^n$. Let $N = 2^n$ be the number of vertices. If π is any permutation of the vertex set, our goal is to send each packet located at a distinct vertex i to its corresponding end vertex $\pi(i)$ simultaneously for each i. We use a model in which the routing occurs in discrete time steps; in each time step, and for each edge e, at most one packet may be sent along e, whereas all other packets are held in a queue at the tail vertex of e. The goal is to minimize the total number of time steps before all packets reach their destination.

Naively, each packet only needs to travel $O(n)$ steps to reach its destination, but due to the potential for congestion, the overall time could be much longer. It turns out that for any deterministic, oblivious routing strategy (oblivious here means the packets' paths are mutually independent), there exists a permutation that requires $\Omega(\sqrt{N/n})$ steps. But with randomization, we can actually achieve a linear time strategy.

Theorem 3.8

There exists a randomized, oblivious routing strategy that terminates in $O(n)$ steps with high probability.

Proof. We sketch the proof. Here is our algorithm:

1. For each packet i, choose an intermediate destination $\delta(i)$ u.a.r. using a bit-fixing path (i.e. iterate through the bits and fix them from left to right).

2. Send each of the packets from $\delta(i)$ to its final destination $\pi(i)$ using a bit-fixing path.

Note that these two phases are symmetric, so it suffices to show that the first phase takes linear time in n. We will take a union bound over the 2^n packets, which is why we will need a Chernoff bound. In particular, let $D(i)$ be the total delay suffered by packet i, so that the total time taken is at most $n + \max_i D(i)$. Then it suffices to show for every i that

$$\mathbb{P}[D(i) > cn] \leq e^{-2n}, \tag{7}$$

so that by union bound we get

$$\mathbb{P}[\exists i : D(i) > cn] \leq 2^n e^{-2n} < 2^{-n}.$$

Through some tedious combinatorial arguments, we may deduce (7) by the Chernoff bound from 3.5. □

3.3 Chernoff for Poisson

As we have mentioned before, there exist chernoff type bounds for unbounded random variables, provided their distribution falls off quickly enough. One such distribution is the Poisson. We will derive bounds similar in spirit to the ones in 3.3 and 3.5.

Theorem 3.9

Suppose $X \sim \text{Poi}(\mu)$. Then for $\lambda > 0$, we have the upper tail bound

$$\mathbb{P}[X \geq \mu + \lambda] \leq \exp\left\{-(\mu+\lambda)\ln\frac{\mu+\lambda}{\mu} + \lambda\right\}.$$

For $\mu > \lambda > 0$, we have the lower tail bound,

$$\mathbb{P}[X \leq \mu - \lambda] \leq \exp\left\{-(\mu-\lambda)\ln\frac{\mu-\lambda}{\mu} - \lambda\right\}.$$

Proof. By Markov's inequality, we have

$$\begin{aligned}\mathbb{P}[X \geq m] &= \mathbb{P}[e^{Xt} \geq e^{mt}] \quad \text{for } t > 0 \\ &\leq e^{-mt}\mathbb{E}[e^{Xt}].\end{aligned}$$

We may compute

$$\begin{aligned}\mathbb{E}[e^{Xt}] &= \sum_{k=0}^{\infty} e^{kt} \cdot \frac{\mu^k e^{-\mu}}{k!} \\ &= \frac{e^{-\mu}}{e^{-\mu e^t}} \cdot \sum_{k=0}^{\infty} \frac{(\mu e^t)^k e^{-\mu e^t}}{k!} \\ &= e^{\mu e^t - \mu},\end{aligned}$$

where in the second line the series sums to 1 since it is the distribution of a $\mathrm{Poi}(\mu e^t)$ random variable. Therefore, our bound becomes

$$e^{-mt+\mu e^t-\mu}. \tag{8}$$

The first derivative is

$$e^{-mt+\mu e^t-\mu}(-m+\mu e^t). \tag{9}$$

The second derivative is

$$e^{-mt+\mu e^t-\mu}(\mu e^t+(-m+\mu e^t)^2)\geq 0.$$

Hence the function is convex in t, so we may set (9) to 0, which implies

$$\mu e^t = m,$$

so that

$$t=\ln\frac{m}{\mu}.$$

Plugging this back into (8) gives us

$$e^{-m\ln\frac{m}{\mu}+m-\mu}.$$

So, if we take $m=\mu+\lambda$, this gives us

$$\mathbb{P}[X\geq\mu+\lambda]\leq\exp\left\{-(\mu+\lambda)\ln\frac{\mu+\lambda}{\mu}+\lambda\right\}. \tag{10}$$

The proof for the lower tail bound proceeds similarly. We have

$$\begin{aligned}
\mathbb{P}[X\leq m] &= \mathbb{P}[e^{Xt}\leq e^{mt}] \quad \text{for } t>0\\
&= \mathbb{P}[e^{-Xt}\geq e^{-mt}]\\
&\leq e^{mt}\mathbb{E}[e^{-Xt}]\\
&= e^{mt}\sum_{k=0}^{\infty} e^{-kt}\frac{\mu^k e^{-\mu}}{k!}\\
&= e^{mt}e^{-\mu+\mu e^{-t}}\sum_{k=0}^{\infty}\frac{(\mu e^{-t})^k e^{-\mu e^{-t}}}{k!}\\
&= e^{mt+\mu e^{-t}-\mu}.
\end{aligned}$$

The first derivative is

$$e^{mt+\mu e^{-t}-\mu}(m-\mu e^{-t}). \tag{11}$$

The second derivative is

$$e^{mt+\mu e^{-t}-\mu}(\mu e^{-t}+(m-\mu e^{-t})^2)\geq 0,$$

so our bound is convex in t. Optimizing over t using (11), we take $t=-\ln\frac{m}{\mu}$. Plugging this back in, along with $m=\mu-\lambda$, gives us our lower tail bound

$$\mathbb{P}[X\leq\mu-\lambda]\leq\exp\left\{-(\mu-\lambda)\ln\frac{\mu-\lambda}{\mu}-\lambda\right\}. \tag{12}$$

□

Corollary 3.10

If $X \sim \mathrm{Poi}(\mu)$ as before, we have the Angluin type bounds, for any $\beta > 0$,

$$\begin{aligned}
\mathbb{P}[X \geq (1+\beta)\mu] &\leq \exp\{-\mu(-\beta + (1+\beta)\ln(1+\beta))\} \\
\mathbb{P}[X \leq (1-\beta)\mu] &\leq \exp\{-\mu(\beta + (1-\beta)\ln(1-\beta))\}.
\end{aligned}$$

Proof. Following immediately by taking $\lambda = \mu\beta$ in Theorem 3.9. $\square$

4 Balls & Bins

In the standard balls & bins model, we throw m balls into n bins independently and u.a.r. We are concerned with the issue of load-balancing. In particular, what is the maximum load of any bin? With the Poisson paradigm, we will be able to prove an asymptotic result under the standard model. But we will be able to obtain a much better result using the "power of two choices".

4.1 Poisson Paradigm

The idea behind the Poisson paradigm is essentially that of stochastic dominance. Since we only care about bounds and asymptotic behavior, it suffices to work in a Poisson model which "dominates" the standard balls and bins model. This will work to our advantage since multinomial coefficients tend to be much harder to work with than terms appearing in a Poisson distribution.

To get some intuition for why we want to use Poisson to dominate our balls and bins model, recall that in the limit, if we keep $m/n = \lambda$ constant, $\text{Bin}(m, 1/n) \xrightarrow{d} \text{Poi}(\lambda)$.

Lemma 4.1

Let $\mathcal{E}$ be any event that depends only on the bin loads such that $\mathbb{P}[\mathcal{E}]$ is monotonically increasing with m. Then

$$\mathbb{P}_X[\mathcal{E}] \leq 4\mathbb{P}_Y[\mathcal{E}],$$

where $\mathbb{P}_X$ is the standard balls and bins model, and $\mathbb{P}_Y$ is the Poisson model, where for each bin we associate an independent $\text{Poi}(m/n)$ variable.

Proof. First, we note that the joint distribution of the balls and bins model is equal to that of n independent $Y_i \sim \text{Poi}(\lambda)$ variables conditioned on $\sum Y_i = m$. Explicitly,

$$\mathbb{P}[X_{1:n} = k_{1:n}] = \frac{1}{n^m} \cdot \frac{m!}{k_1! \dots k_n!} = \frac{\prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!}}{\frac{e^{-n\lambda}(n\lambda)^m}{m!}} = \frac{\prod_{i=1}^n \mathbb{P}[Y_i = k_i]}{\mathbb{P}[\sum_{i=1}^n Y_i = m]}.$$

So, for an event $\mathcal{E}$, this tells us

$$\mathbb{P}_X[\mathcal{E}] = \mathbb{P}_Y\left[\mathcal{E} | \sum_{i=1}^n Y_i = m\right]$$

Now, we have

$$\begin{aligned}
\mathbb{P}_Y[\mathcal{E}] &= \sum_{k=0}^{\infty} \mathbb{P}_Y\left[\mathcal{E} \middle| \sum_{i=1}^{n} Y_i = k\right] \mathbb{P}_Y\left[\sum_{i=1}^{n} Y_i = k\right] \\
&\geq \mathbb{P}_Y\left[\mathcal{E} \middle| \sum_{i=1}^{n} Y_i = m\right] \mathbb{P}_Y\left[\sum_{i=1}^{n} Y_i \geq m\right] \\
&\geq \mathbb{P}_Y\left[\mathcal{E} \middle| \sum_{i=1}^{n} Y_i = m\right] \cdot \frac{1}{4} \\
&= \mathbb{P}_X[\mathcal{E}] \cdot \frac{1}{4}.
\end{aligned}$$

In the second line we used monotonicity, and in the third we used the fact that for any $Y \sim \mathrm{Poi}(\lambda)$, we have

$$\mathbb{P}[Y \geq \mathbb{E}[Y]] \geq 1/4,$$

and applied to the sum of independent Poissons $Y = \sum Y_i$, which is itself Poisson. □

Armed with this lemma, we can now show:

Theorem 4.2

For the balls and bins model with $m = n$ (so $\lambda = 1$), the maximum load of any bin is

$$\Theta\left(\frac{\ln n}{\ln \ln n}\right)$$

asymptotically almost surely (a.a.s.).

Proof. It suffices to show that the maximum load lies in

$$\left((1-\epsilon)\frac{\ln n}{\ln \ln n}, (1+\epsilon)\frac{\ln n}{\ln \ln n}\right)$$

for any $\epsilon > 0$ a.a.s. For notation, define $c_1 = 1 + \epsilon$ and $c_2 = 1 - \epsilon$, and the events

$$\begin{aligned}
\mathcal{E}_1 &:= \text{some bin contains more than } c_1 \tfrac{\ln n}{\ln \ln n} \text{ balls}, \\
\mathcal{E}_2 &:= \text{no bin contains more than } c_2 \tfrac{\ln n}{\ln \ln n} \text{ balls}.
\end{aligned}$$

We will show that $\mathbb{P}[\mathcal{E}_i] = o(1)$ for $i = 1, 2$.

Note that since $\mathcal{E}_i$ are monotonic (for decreasing, a similar bound holds). Therefore, by Lemma 4.1, we may work with independent Poi(1) variables. We have the following useful bounds:

$$\frac{1}{ek!} \leq \mathbb{P}[Y_i \geq k] \leq \frac{1}{ek!}\left(1 + \frac{1}{k+1} + \frac{1}{(k+1)(k+2)} + \cdots\right) \leq \frac{1}{k!}.$$

Setting $k = c_1 \frac{\ln n}{\ln \ln n}$, we get

$$\ln \mathbb{P}[Y_i \geq k] \leq -\ln k! \sim -k \ln k = -c_1 \cdot \frac{\ln n}{\ln \ln n}(\ln \ln n + \ln c_1 - \ln \ln \ln n) \sim -c_1 \ln n,$$

which implies $\mathbb{P}[Y_i \geq k] = o(n^{-1})$. Taking a union bound, gives us $\mathbb{P}[\mathcal{E}_1] = o(1)$.

Next, to show $\mathbb{P}[\mathcal{E}_2] = o(1)$, we note that

$$\mathbb{P}_Y[\mathcal{E}_2] = (1 - \mathbb{P}[Y_i \geq k])^n \leq \left(1 - \frac{1}{ek!}\right)^n \leq e^{-\frac{n}{ek!}}.$$

Taking $k = c_2 \frac{\ln n}{\ln\ln n}$, we can show that the exponent $\frac{n}{ek!} \to \infty$ as $n \to \infty$. It follows that $\mathbb{P}[\mathcal{E}_2] = o(1)$ as well. □

4.2 Power of Two Choices

The idea behind the power of two choices is that if, instead of choosing a single random bin, we choose d random bins for each ball, and place it in the bin with lowest load. We will show that if d is just 2, we can achieve a significant cut on the maximum load from before (Theorem 4.2).

Theorem 4.3

With $d = 2$ choices, the maximum load is a.a.s. at most

$$\frac{\ln\ln n}{\ln 2} + \Theta(1).$$

Proof. Let B_i be the number of bins with load at least i at the end of the process. We wish to find upper bounds β_i such that $B_i \leq \beta_i$ w.h.p. For then

$$\mathbb{P}[\text{a given ball is placed in a bin with load} \geq \text{i}] \leq \left(\frac{\beta_i}{n}\right)^2.$$

Furthermore, this tells us the distribution of B_{i+1} is dominated by $\text{Bin}(n, (\beta_i/n)^2)$, since the number of bins with load $i+1$ is at most an indicator for each ball with probability $(\beta_i/n)^2$.

First, set $\beta_6 = \frac{n}{2e}$. Note that $\mathbb{P}[B_6 \leq \beta_6] = 1$ since there can be at most $\frac{n}{6} \leq \frac{n}{2e}$ bins with ≥ 6 balls in them. Then, for $i > 6$, set

$$\beta_{i+1} = \frac{e\beta^2}{n}. \tag{13}$$

Define the event $\mathcal{E} = \{B_i \leq \beta_i\}$. We have by a Chernoff bound (plug $\beta = e - 1$ into the second bound in 3.5)

$$\begin{aligned}
\mathbb{P}[\neg\mathcal{E}_{i+1}|\mathcal{E}_i] &= \mathbb{P}[B_{i+1} > \beta_{i+1}|\mathcal{E}_i] \\
&\leq \frac{\mathbb{P}[\text{Bin}(n, (\beta_i/n)^2) \geq \beta_{i+1}]}{\mathbb{P}[\mathcal{E}_i]} \\
&\leq \frac{e^{-\beta_i^2/n}}{\mathbb{P}[\mathcal{E}_i]} \\
&\leq \frac{1/n^2}{\mathbb{P}[\mathcal{E}_i]},
\end{aligned}$$

where the last line holds for $\beta_i^2/n \geq 2\ln n$. To remove the conditioning, we prove by induction on i that

$$\mathbb{P}[\neg\mathcal{E}_i] \leq \frac{i}{n^2}.$$

For the base case, note that $\mathbb{P}[\neg\mathcal{E}_6] = 0 \leq 6/n^2$. For the step, write

$$\begin{aligned}
\mathbb{P}[\neg\mathcal{E}_{i+1}] &\leq \mathbb{P}[\neg\mathcal{E}_{i+1}|\mathcal{E}_i]\mathbb{P}[\mathcal{E}_i] + \mathbb{P}[\neg\mathcal{E}_i] \\
&\leq \frac{1/n^2}{\mathbb{P}[\mathcal{E}_i]} \cdot \mathbb{P}[\mathcal{E}_i] + \frac{i}{n^2} \\
&\leq \frac{i+1}{n^2}.
\end{aligned}$$

Now, let i^* be the first i for which $\beta_i^2 < 2n\ln n$. Using (13) we get that $i^* = \frac{\ln\ln n}{\ln 2} + O(1)$. We now show that

$$\mathbb{P}[B_{i^*+2} \geq 1] \leq O\left(\frac{\ln^2 n}{n}\right),$$

which will finish the proof, since this tells us that for $i \geq i^* + 2$ the number of bins with load i will be a.a.s. 0.

Note that w.h.p. there are $\leq \sqrt{2n\ln n}$ bins with load $\geq i^*$, so the expected number of balls falling in bins with load $\geq i^* + 1$ is at most $2\ln n$. Then we have by another Chernoff bound,

$$\begin{aligned}
\mathbb{P}[B_{i^*+1} \geq 6\ln n] &\leq \mathbb{P}[B_{i^*+1} \geq 6\ln n|\mathcal{E}_{i^*}] \cdot \mathbb{P}[\mathcal{E}_i^*] + \mathbb{P}[\neg\mathcal{E}_{i^*}] \\
&\leq \frac{\mathbb{P}[\text{Bin}(n, 2\ln n/n) \geq 6\ln n]}{\mathbb{P}[\mathcal{E}_{i^*}]} \cdot \mathbb{P}[\mathcal{E}_{i^*}] + \frac{1}{n} \\
&\leq \frac{1}{n^2} + \frac{1}{n} = O(n^{-1}).
\end{aligned}$$

Doing this for another step, we get

$$\begin{aligned}
\mathbb{P}[B_{i^*+2} \geq 1] &\leq \mathbb{P}[B_{i^*+2} \geq 1|B_{i^*+1} < 6\ln n] \cdot \mathbb{P}[B_{i^*+1} < 6\ln n] + \mathbb{P}[B_{i^*+1} \geq 6\ln n] \\
&\leq \frac{\mathbb{P}[\text{Bin}(n, (6\ln n/n)^2) \geq 1]}{\mathbb{P}[B_{i^*+1} < 6\ln n]} \cdot \mathbb{P}[B_{i^*+1} < 6\ln n] + O(n^{-1}) \\
&\leq \left(\frac{6\ln n}{n}\right)^2 \cdot n + O(n^{-1}) \\
&= O\left(\frac{(\ln n)^2}{n}\right),
\end{aligned}$$

where instead of a Chernoff bound, we have used a union bound for the probability that the binomial is nonzero. It follows that as $n \to \infty$, w.h.p. the maximum load is at most

$$i^* + 2 = \frac{\ln\ln n}{\ln 2} + \Theta(1).$$

□

5 The Galton-Watson Branching Process

The Galton-Watson process is used to model population growth and decay. We assume the population starts out with one node. At each time step, each node from the previous time step gives rise to some number of children distributed as the random variable X. We define Z_i to be the number of nodes at time i. So, $Z_0 = 1$ and Z_i is distributed as the sum of Z_{i-1} independent copies of X.

We are interested in the probability of extinction:

$$\mathbb{P}[\text{extinction}] = \lim_{n\to\infty} \mathbb{P}[Z_n = 0].$$

Theorem 5.1

For a branching process defined with nonnegative integer-valued random variable X satisfying $\mathbb{P}[X = 1] < 1$ and $\mathbb{P}[X = 0] > 0$ (these are to rule out trivial cases), we have:

(i) If $\mathbb{E}[X] \leq 1$ then $\lim_{n\to\infty} \mathbb{P}[Z_n = 0] = 1$.

(ii) If $\mathbb{E}[X] > 1$ then $\lim_{n\to\infty} \mathbb{P}[Z_n = 0] = p^* < 1$, where $p^* \in (0, 1)$ is the unique fixed point of the probability generating function

$$f(x) = \sum_{i\geq 0} \mathbb{P}[X = i]x^i.$$

Proof. Similar to the probability generating function for X, we define the probability generating function f_n for Z_n to be

$$f_n(x) = \sum_{i\geq 0} \mathbb{P}[Z_n = i]x^i.$$

Therefore $f_1(x) = f(x)$ since $Z_1 \sim X$. By comparing coefficients, we note that

$$f_n(x) = f(f_{n-1}(x)).$$

For notational convenience, we let the probability of extinction at time n be

$$q_n := \mathbb{P}[Z_n = 0] = f_n(0).$$

Note that the probability of extinction at time n is at least as large as that of time $n - 1$, so we have a monotonic increasing sequence

$$0 = q_0 < q_1 \leq q_2 \leq \cdots \leq 1,$$

which converges to some fixed point q^* of f. Furthermore, observe that f is a strictly increasing convex function from $[0, 1] \to [0, 1]$, with $f(0) > 0$ and $f(1) = 1$. Now, note that

$$\mathbb{E}[X] = f'(1) = \sum_{i\geq 0} \mathbb{P}[Z_n = i] \cdot i.$$

So we have two cases:

(i) If $\mathbb{E}[X] = f'(1) \leq 1$, then f only has a fixed point at 1, so that

$$\lim_{n\to\infty} \mathbb{P}[Z_n = 0] = q^* = 1.$$

(ii) If $\mathbb{E}[X] = f'(1) > 1$, then f has a unique fixed point in $(0, 1)$, so that

$$\lim_{n\to\infty} \mathbb{P}[Z_n = 0] = q^* \in (0, 1).$$

□

6 Geometric Embeddings

Recall a *metric space*, given by (X, d), where X is a set and $d : X \times X \to \mathbb{R}$ is the metric or distance function, satisfying for every $x, y, z \in X$:

(i) (Positive Definite) $d(x, y) \geq 0$ with equality iff $x = y$.
(ii) (Symmetric) $d(x, y) = d(y, x)$.
(iii) (Triangle Inequality) $d(x, y) \leq d(x, z) + d(z, y)$.

For this section, we will restrict $|X|$ and d to be finite. We are interested in finding a way to map (X, d) to a nicer space (Y, d') by some mapping φ which preserves distances up to a small distortion,

$$d'(\varphi(x), \varphi(y)) \approx d(x, y) \quad \forall x, y \in X.$$

6.1 Dimensionality Reduction

Theorem 6.1 (Johnson-Lindenstrauss Lemma)
Let X be any set of n points in $\mathbb{R}^d$. For any desired level of approximation $\epsilon \in (0, 1)$, there exists a mapping $\varphi : \mathbb{R}^d \to \mathbb{R}^k$ such that for every $u, v \in X$,

$$(1 - \epsilon)\|u - v\|_2^2 \leq \|\varphi(u) - \varphi(v)\|_2^2 \leq (1 + \epsilon)\|u - v\|_2^2,$$

where

$$k \geq \left\lceil \frac{4 \ln n}{\epsilon^2/2 - \epsilon^3/3} \right\rceil \leq \left\lceil \frac{24 \ln n}{\epsilon^2} \right\rceil.$$

Proof. We use a random construction and prove with the probabilistic method that an ϵ-approximation exists for the chosen k. In particular, let φ be the linear map which projects the point v onto a random k-dimensional hyperplane, obtaining the projection v', and then scaling appropriately to $\sqrt{\frac{d}{k}}v'$. So, we wish to show that

$$1 - \epsilon \leq \frac{\|\varphi(u) - \varphi(v)\|_2^2}{\|u - v\|_2^2} \leq 1 + \epsilon$$

for every pair $(u, v) \in X \times X$ with positive probability. Note that due to linearity, we may assume WLOG that $\|u - v\|_2 = 1$, so that we may focus on the random variable

$$\|\varphi(v)\|_2^2,$$

where v is some unit vector. Now, projecting v onto a random k-dimensional hyperplane is equivalent to projecting a random vector onto a fixed k-dimensional hyperplane, say the one spanned by the first k coordinates. So generate a vector u.a.r. from the unit hypersphere by first taking the random variable

$$X = (X_1, \ldots, X_d) \sim \mathcal{N}(0, I),$$

and normalizing it. So, we may consider φ as the mapping given by

$$\frac{X}{\|X\|_2} = \frac{(X_1, \ldots, X_d)}{\sqrt{X_1^2 + \cdots + X_d^2}} \mapsto \sqrt{\frac{d}{k}} \cdot \frac{(X_1, \ldots, X_k)}{\sqrt{X_1^2 + \cdots + X_d^2}}$$

Then we wish to analyze the distribution of

$$L := \|\varphi(v)\|_2^2 = \frac{d}{k} \cdot \frac{X_1^2 + \cdots + X_k^2}{X_1^2 + \cdots + X_d^2}.$$

We know that

$$\mathbb{E}\left[\frac{X_1^2 + \cdots + X_d^2}{X_1^2 + \cdots + X_d^2}\right] = 1,$$

so by linearity and symmetry we have

$$\mathbb{E}\left[\frac{X_i^2}{X_1^2 + \cdots + X_d^2}\right] = \frac{1}{d},$$

which tells us by symmetry that $\mathbb{E}[L] = \frac{d}{k} \cdot \frac{k}{d} = 1$. This is why we scaled our projection by $\sqrt{\frac{d}{k}}$ initially, because we want our variable L to be in $(1-\epsilon, 1+\epsilon)$ with sufficiently high probability.

We now use the following Chernoff bound specialized for our current setting:

Lemma 6.2

For L defined as above, we have

(i) $\mathbb{P}[L \leq 1-\epsilon] \leq \exp(\frac{-\epsilon^2 k}{4})$

(ii) $\mathbb{P}[L \geq 1+\epsilon] \leq \exp(\frac{-k}{2}(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}))$

The proof follows in similar spirit to the proof of the chernoff bound for independent $0-1$ variables, so we will instead focus finishing up the proof of the theorem.

Note that if we take $k \geq \left\lceil \frac{4 \ln n}{\epsilon^2/2 - \epsilon^3/3} \right\rceil$, we have

$$\mathbb{P}[|L-1| \geq \epsilon] \leq 2\exp(-2\ln n) = \frac{2}{n^2}.$$

So by a union bound,

$$\mathbb{P}[|L-1| \geq \epsilon \text{ for any pair } (u,v)] \leq \frac{2}{n^2} = 1 - \frac{1}{n}.$$

Therefore the random embedding φ is an ϵ-distortion with probability at least $1/n$. But then an ϵ-distortion exists, and we can find one by repeatedly sampling with $O(n)$ expected trials. $\square$

6.2 Embedding into ℓ^p

Instead of embedding into a smaller dimensional space, we now wish to embed an arbitrary metric space (X, d) into a space equipped with the ℓ_p metric with minimal distortion. For this section, we will only prove the case where $p = 1$, but much of the proof carries over to $p = 2$ (and these are the two that are most frequently used).

Theorem 6.3 (Bourgain's Embedding Theorem)
Let (X, d) be a metric space with $|X| = n$. Then there exists an embedding $\varphi : (X, d) \to (\mathbb{R}^k, \ell_1)$ such that for every $x, y \in X$,

$$\frac{1}{c \log n} d(x, y) \leq \|\varphi(x) - \varphi(y)\|_1 \leq d(x, y),$$

where $k = O(\log^2 n)$.

We will use a random construction, where we pick $m = r \log^2 n$ sets $A_i \subseteq X$ chosen as follows: For each $t \in \{1, 2, \ldots, \log n\}$, construct $r \log n$ sets $\{A_i^t\}_{i=1}^{r \log n}$ such that for each i, include each element X in A_i^t independently with probability 2^{-t}.

Then we define the embedding as

$$\varphi(x) = \frac{1}{m}(d(x, A_1), \ldots, d(x, A_m)),$$

where $d(x, A_i) = \min_{y \in A_i} d(x, y)$. So in a way we are treating the subsets $\{A_i\}$ as a sort of coordinates in our new space.

Proof of Upper Bound. We have

$$\|\varphi(x) - \varphi(y)\|_1 = \frac{1}{m} \sum_{i=1}^{m} |d(x, A_i) - d(y, A_i)|.$$

By triangle inequality, we have

$$|d(x, A_i) - d(y, A_i)| \leq d(x, y),$$

so it follows that

$$\|\varphi(x) - \varphi(y)\|_1 \leq d(x, y).$$

□

Proof of Lower Bound. The intuition is that we want a substantial portion of the

$$|d(x, A_i^T) - d(y, A_i^t)|$$

terms to contribute to

$$\|\varphi(x) - \varphi(y)\|_1$$

so that it becomes at least $\frac{1}{c\log n}d(x,y)$. To do this, we first fix a pair $x,y \in X$ (we can take a union bound later). We want to come up with a notion of "good" for the sets A_i^t, so that a set A_i^t is good when its contribution is high.

To make this precise, first define $B_\rho(x)$ and $B_\rho^o(x)$ to be the closed and open balls of radius ρ centered at x, respectively. Then define the increasing sequence of radii

$$0 = \rho_0 < \rho_1 < \cdots$$

by setting ρ_t to be the smallest $\rho > 0$ such that $B_\rho(x)$ and $B_\rho(y)$ both have at least 2^t points of X. We continue this sequence as long as $\rho_t < \frac{1}{4}d(x,y)$, and if t^*-1 is the last such t, we set $\rho_{t^*} = \frac{1}{4}d(x,y)$. Note that the balls centered at x are disjoint from the ones centered at y.

Now, we say a set A_i^t is good if it intersects $B_{\rho_{t-1}}(y)$ but does not intersect $B_{\rho_t}^o(x)$, given that WLOG the x-ball defines this radius (therefore $|B_{\rho_t}^o(x)| < 2^t$). Then if A_i^t is good,

$$d(x,A_i^t) \geq \rho_t, \quad d(y,A_i^t) \leq \rho_{t-1},$$

so that its contribution to $\|\varphi(x)-\varphi(y)\|_1$ is at least

$$\frac{1}{m}(\rho_t - \rho_{t-1}).$$

Now, using the way in which we defined our ρ_t's, we may show that w.h.p. enough of our sets are good. For any particular set A_i^t, we have

$$\begin{aligned}
\mathbb{P}[A_i^t \text{ is good for } x,y] &= \mathbb{P}[A_i^t \text{ hits } B_{\rho_{t-1}}(y) \text{ and misses } B_{\rho_t}^o(x)] \\
&\geq \mathbb{P}[A_i^t \text{ hits } B_{\rho_{t-1}}(y)] \cdot \mathbb{P}[B_{\rho_t}^o(x)] \quad \text{(positively correlated)} \\
&\geq \left(1-(1-2^{-t})^{2^{t-1}}\right)\left((1-2^{-t})^{2^t}\right) \\
&\geq \left(1-\frac{1}{\sqrt{e}}\right)\cdot\frac{1}{4} \\
&\geq \frac{1}{12}.
\end{aligned}$$

So, since for each value of t there are $r\log n$ sets A_i^t, we have for each t

$$\mu := \mathbb{E}[\#\text{ of good sets for } x,y] \geq \frac{r\log n}{12}.$$

Then applying a Chernoff bound, we have for each t

$$\mathbb{P}[\#\text{ of good sets for } x,y \leq \frac{1}{2}\mu] \leq \exp\left(-\frac{\mu}{8}\right) = \exp\left(-\frac{r\log n}{96}\right) \leq \frac{1}{n^3},$$

if we choose $r = 288$. A union bound over all pairs x,y and values of t tells us that with probability $1 - O(\frac{n^2\log n}{n^3}) = 1 - o(1)$ every pair x,y has at least $\frac{1}{2}\mu = \frac{r\log n}{24}$ good sets for all t. Then there exists an embedding for which this holds, so that

$$\begin{aligned}
\|\varphi(x)-\varphi(y)\|_1 &= \frac{1}{m}\sum_{t=1}^{t^*}\sum_{i=1}^{r\log n}|d(x,A_i^t)-d(y,A_i^t)| \\
&\geq \frac{1}{m}\frac{r\log n}{24}((\rho_1-\rho_0)+(\rho_2-\rho_1)+\cdots+(\rho_{t^*}-\rho_{t^*-1})) \\
&= \frac{1}{m}\cdot\frac{r\log n}{24}(\rho_{t^*}-\rho_0) \\
&= \frac{1}{m}\cdot\frac{r\log n}{24}\cdot\frac{1}{4}d(x,y) \\
&= \frac{1}{96\log n}d(x,y),
\end{aligned}$$

which is what we wanted, where $c = 96$. $\square$

Remark. Note that we've actually shown something much stronger than we needed. In addition to mere existence of a good embedding into ℓ_1, we've also shown that w.h.p. our random construction will produce such an embedding, so we can actually compute the embedding explicitly through sampling.

7 Martingales

We look at martingales from the perspective that they are tools for obtaining large deviation bounds when Chernoff/Hoeffding bounds fail. In particular, in Chernoff-type bounds we will often assume independence of the random variables, but martingales are more flexible.

Martingales are motivated by fair gambling games, in which a gambler's fortune remains the same in expectation from one time step to the next. In what follows, think of Z_i as the outcome of the i-th game and X_i as the gambler's capital after game i.

Definition 7.1. Let (Z_i) and (X_i) be sequences of random variables on the same probability space such that

$$\mathbb{E}[X_i|Z_{1:i-1}] = X_{i-1} \quad \text{for all } i.$$

We say that (X_i) is a *martingale* w.r.t. (Z_i).

The sequence

$$Y_i = X_i - X_{i-1}$$

is called a *martingale difference sequence*. So, in particular if X_i is a martingale we have

$$\mathbb{E}[Y_i|Z_{1:i-1}] = 0 \quad \text{for all } i.$$

7.1 Examples

Example 7.2 (Doob Martingale)**.** Let A and (Z_i) be *arbitrary* random variables on a common probability space. Then

$$X_i := \mathbb{E}[A|Z_{1:i}]$$

is called the *Doob martingale*. This follows from the tower property of conditional expectation:

$$\begin{aligned}\mathbb{E}[X_i|Z_{1:i-1}] &= \mathbb{E}[\mathbb{E}[A|Z_{1:i}]|Z_{1:i-1}] \\ &= \mathbb{E}[A|Z_{1:i-1}] \\ &= X_{i-1}.\end{aligned}$$

We can think of $A = f(Z_{1:n})$ as some function of the games' results. Then the martingale is just revealing more and more information at every time step. So, we begin with no information, and the value of the martingale is $\mathbb{E}[A]$. Then by the end we have the deterministic value $f(Z_{1:n})$.

Example 7.3 (Coin Tosses)**.** Let A be the number of heads after N tosses, where N is a fixed constant. If (Z_i) are the outcomes of the tosses, then we have the martingale

$$X_i = \mathbb{E}[A|Z_{1:i}].$$

Here we can come up with some intuition for the tower property. Say $N = 3$. The tower property says

$$\mathbb{E}[\mathbb{E}[A|Z_1, Z_2]|Z_1] = \mathbb{E}[A|Z_1].$$

We can think of Z_1 as partitioning up the probability space into two regions, one where $Z_1 = 0$ and another where $Z_1 = 1$. So the RHS says we average A based on which region we're in. The LHS says that after averaging out based on which of the four regions we're in (partitioned by Z_1, Z_2), we then average over the coarser partition given by just Z_1. In coin flipping terms, if we want to know the mean of A given the result of the first coin flip, it suffices to first determine the mean of A given the results of the first two coin flips, and then averaging that out with only the first result fixed, and the second flip allowed to vary over its distribution. Explicitly, if we have fair coin tosses, the RHS is

$$\mathbb{E}[A|Z_1] = Z_1 + \frac{1}{2} + \frac{1}{2}.$$

The LHS is

$$\begin{aligned}\mathbb{E}[\mathbb{E}[A|Z_1, Z_2]|Z_1] &= \mathbb{E}[Z_1 + Z_2 + \frac{1}{2}|Z_1] \\ &= Z_1 + \mathbb{E}[Z_2|Z_1] + \mathbb{E}[\frac{1}{2}|Z_1] \\ &= Z_1 + \frac{1}{2} + \frac{1}{2}.\end{aligned}$$

Example 7.4 (Balls & Bins)**.** Consider m balls being thrown into n bins independently and u.a.r. For $1 \le i \le m$ let $Z_i \in \{1, \ldots, n\}$ be the destination of the i-th ball. Let A be the number of empty bins, and we have the corresponding Doob martingale

$$X_i = \mathbb{E}[A|Z_{1:i}].$$

Example 7.5 (Vertex & Edge Exposure Martingales)**.** Consider the $\mathcal{G}_{n,p}$ random graph model. Let $Z_i \in \{0,1\}^{n-i}$ be a vector of indicators denoting whether edges between vertex i and vertices $j > i$ are present. For any graph property $A = f(Z_1, \ldots, Z_n)$, the corresponding Doob martingale $X_i = \mathbb{E}[A|Z_{1:i}]$ is called the *vertex exposure martingale*. On the other hand, let Z_i be an indicator for whether the i-th pair of vertices has an edge. Then for any graph property $A = f(Z_1, \ldots, Z_n)$, the corresponding Doob martingale $X_i = \mathbb{E}[A|Z_{1:i}]$ is called the *edge exposure martingale*.

7.2 Azuma's Inequality

The following is akin to the Chernoff-type bounds we've seen before. However, it can be used without assuming independence. In fact, we can recover Chernoff's bound by applying Azuma's inequality to the coin tossing martingale with $c_i = 1$ for each i, giving the bound $\exp(-\lambda^2/2n)$.

Theorem 7.6 (Azuma's Inequality)

Let (X_i) be a martingale w.r.t. (Z_i), and let $Y_i = X_i - X_{i-1}$ be the corresponding difference sequence. If we have bounded increments $c_i > 0$ such that $|Y_i| \le c_i$ for all i, then

$$\left.\begin{array}{r}\mathbb{P}[X_n \ge X_0 + \lambda] \\ \mathbb{P}[X_n \le X_0 - \lambda]\end{array}\right\} \le \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}\right).$$

Proof. First, we show that for a random variable $Y \in [-1, +1]$ with $\mathbb{E}[Y] = 0$, if $t \geq 0$ then

$$\mathbb{E}[e^{tY}] \leq e^{t^2/2}. \tag{14}$$

By convexity, we have for any $y \in [-1, +1]$,

$$e^{ty} \leq \frac{1}{2}(1+y)e^t + \frac{1}{2}(1-y)e^{-t}.$$

Taking expectations,

$$\begin{aligned}
\mathbb{E}[e^{tY}] &\leq \frac{1}{2}e^t + \frac{1}{2}e^{-t} \\
&= \frac{1}{2}\left[\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right) + \left(1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \cdots\right)\right] \\
&= 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \cdots \\
&= \sum_{n=0}^{\infty} \frac{t^{2n}}{(2n)!} \\
&\leq \sum_{n=0}^{\infty} \frac{t^{2n}}{2^n n!} \\
&= \sum_{n=0}^{\infty} \frac{(t^2/2)^n}{n!} = e^{t^2/2}.
\end{aligned}$$

Now, in a similar fashion as the proof of the Chernoff bounds before, we have

$$\begin{aligned}
\mathbb{P}[X_n - X_0 \geq \lambda] &= \mathbb{P}[e^{t(X_n - X_0)} \geq e^{t\lambda}] \quad \text{for } t > 0 \\
&\leq e^{-t\lambda}\mathbb{E}[e^{t(X_n - X_0)}] \\
&= e^{-t\lambda}\mathbb{E}[e^{t(Y_n + X_{n-1} - X_0)}] \\
&= e^{-t\lambda}\mathbb{E}[\mathbb{E}[e^{t(Y_n + X_{n-1} - X_0)}|\mathcal{F}_{n-1}]],
\end{aligned}$$

where in last line we used law of iterated expectation (here $\mathcal{F}_{n-1}$ is the $(n-1)$-th filtration, which is essentially just all the information from random variables $Z_{1:n-1}$). Note that given $\mathcal{F}_{n-1}$, we can factor out $X_{n-1} - X_0$ as constants. Then we may apply (14) to the random variable Y_n/c_n to get

$$\begin{aligned}
\mathbb{E}[e^{t(Y_n + X_{n-1} - X_0)}|\mathcal{F}_{n-1}] &= e^{t(X_{n-1} - X_0)}\mathbb{E}[e^{tY_n}|\mathcal{F}_{n-1}] \\
&\leq e^{t(X_{n-1} - X_0)}e^{t^2 c_n^2/2}.
\end{aligned}$$

Note that the proof of (14) still holds if we assume conditioning on $\mathcal{F}_{n-1}$, so the second step above was justified. Putting this together, we have

$$\mathbb{P}[X_n - X_0 \geq \lambda] \leq e^{-t\lambda}e^{t^2 c_n^2/2}\mathbb{E}[e^{t(X_{n-1} - X_0)}].$$

Now, we can keep expanding the last term inductively to get

$$\mathbb{P}[X_n - X_0 \geq \lambda] \leq e^{-t\lambda + t^2 \sum_{i=1}^n c_i^2/2}.$$

Then, optimizing over $t > 0$, we take $t = \frac{\lambda}{\sum c_i^2}$ to get

$$\mathbb{P}[X_n - X_0 \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}\right).$$

$\square$

Corollary 7.7 (Generalized Azuma's Inequality)

Suppose instead $Y_i \in [a_i, b_i]$. Through a standard change of variables, we can derive the following variation of Azuma's inequality:

$$\left.\begin{array}{l}\mathbb{P}[X_n \geq X_0 + \lambda] \\ \mathbb{P}[X_n \leq X_0 - \lambda]\end{array}\right\} \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

7.2.1 Simple Applications of Azuma

Example 7.8 (Gambling)**.** If Z_i is the outcome of the i-th game, and X_i is the gambler's capital at time i, then assuming the gambler can go into debt, we have

$$\mathbb{P}[|X_n - X_0| \geq \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2nM^2}\right),$$

where X_0 is some deterministic initial capital, and in each game the gambler can only win or lose at most M capital.

Example 7.9 (Coin tossing)**.** Let Z_i be the outcome of the i-th coin toss and X the number of heads after n tosses. Then we have bounded increments

$$|X_i - X_{i-1}| \leq 1,$$

so Azuma's inequality gives us

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

For our next example, we will first need the following definition and result.

Definition 7.10. A function of integer-valued variables $f(Z_1, \ldots, Z_n)$ is said to be *c-lipschitz* if changing the value of any one coordinate of f causes f to change by at most $\pm c$.

Lemma 7.11

If f is c-Lipschitz and Z_i is independent of the future $Z_{i+1:n}$ when conditioned on the past $Z_{1:i-1}$. Then the Doob martingale X_i given by $A = f(Z_1, \ldots, Z_n)$ has the bounded increments $|X_i - X_{i-1}| \leq c$.

Proof. Let $\hat{Z}_i$ be a random variable with the same distribution as Z_i conditioned on $Z_{1:i-1}$ but independent of $Z_{i:n}$. To see that such a random variable even exists, note that, given a distribution, we may construct a random variable with that distribution that is independent to any number of other random variables. Now, since expectations are preserved as long as distributions stay the same, we have

$$\begin{aligned} X_{i-1} &= \mathbb{E}[f(Z_1, \dots, Z_i, \dots, Z_n)|Z_{1:i-1}] \\ &= \mathbb{E}[f(Z_1, \dots, \hat{Z}_i, \dots, Z_n)|Z_{1:i-1}] \\ &= \mathbb{E}[f(Z_1, \dots, \hat{Z}_i, \dots, Z_n)|Z_{1:i}], \end{aligned}$$

where the third line follows because Z_i is independent of $Z_{i+1:n}$ when conditioned on $Z_{1:i-1}$ by assumption. Therefore, by the c-Lipschitz assumption,

$$|X_i - X_{i-1}| = |\mathbb{E}[f(Z_1, \dots, Z_i, \dots, Z_n) - f(Z_1, \dots, \hat{Z}_i, \dots, Z_n)|Z_{1:i}]| \leq c,$$

as desired. □

What this lemma is saying is basically that if each individual gambling game has bounded increments while the results of every other game is fixed, then we have bounded increments overall as long as the future is not affected by the past.

For suppose the result of game i, conditioned on all the previous games, had some effect on the future games. Then it's possible that the increments blow up over time. As a counterexample, consider the following gambling game:

- The first round is a fair coin toss ± 1. So winning is $+1$ and losing is -1.
- Suppose at round $i-1$ the result is distributed as a fair $k \pm 1$. If the gambler won last round then round i will be distributed as a fair $(k+1) \pm 1$, and if the gambler lost last round then round i will be distributed as a fair $(k-1) \pm 1$.

Clearly our function $f(Z_1, \dots, Z_n) = Z_1 + \cdots + Z_n$ is 2-Lipschitz. However, we don't have bounded increments $|X_i - X_{i-1}| \leq 2$. For depending on how the game goes, we could be winning up to $+n$ by the last round. As we can see, the independence assumption of the lemma was indeed necessary, as it ensures thatnothing which happens in the present can affect our increments in the future.

Example 7.12 (Balls & Bins)**.** Consider the m balls and n bins model from before. We let Z_i be the bin selected by the i-th ball, and $X = f(Z_1, \dots, Z_n)$ be the number of empty bins once all balls have been thrown.

Since each ball can change the number of empty bins by at most 1, f is 1-Lipschitz, so by Lemma 7.11, we have bounded increments $|X_i - X_{i-1}| \leq 1$, and by Azuma's inequality we have

$$\mathbb{P}[|X - X_0| \geq \lambda] = \mathbb{P}[|X - \mathbb{E}[X]| \geq \lambda] \leq 2 \exp\left(-\frac{\lambda^2}{2m}\right).$$

Note that this bound is only really useful if $\lambda \gg \sqrt{m}$, hence the classification of Azuma as a large deviations-type bound. Furthermore, we couldn't have applied Chernoff bounds here, as the increments are not independent.

7.2.2 The Chromatic Number of Random Graphs

Given a graph $G \in \mathcal{G}_{n,p}$, we are interested in a probabilistic estimate of its chromatic number $X = \chi_G$, which is the minimum number of colors needed to color all vertices of G so that no two endpoints of an edge share the same color. An equivalent definition is the size of the minimal partition of the vertex set into independent sets.

Theorem 7.13 (Large Deviation Bound of χ_G)
Let X be the chromatic number of $G \in \mathcal{G}_{n,p}$. Thn

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

Proof. Recall the vertex exposure martingale from 7.5, where Z_i encodes the edges between vertex i and vertices $i+1, \ldots, n$. The reason we define it this way is so that the sequence (Z_i) uniquely determines G, so that there exists a function f such that $X = f(Z_1, \ldots, Z_n)$.

So suppose $X = f(Z_1, \ldots, Z_n)$. Then note that f is 1-Lipschitz, for if we add edges from a fixed vertex i to other vertices, all we have to do is color i a new color, so that the chromatic number only increases by 1. Similarly, if we remove edges incident to i we can only decrease the chromatic number by at most 1. Furthermore, since all edges are generated independently, we know that Z_i is independent of $Z_{i+1:n}$ when conditioned on $Z_{1:i-1}$ since all the possible edge sets of each Z_i are disjoint. Therefore we may apply Lemma 7.11 and then subsequently Azuma's inequality to the Doob martingale of X with each $c_i = 1$ to obtain

$$\mathbb{P}[|X_n - X_0| \geq \lambda] = \mathbb{P}[|X - \mathbb{E}[X]| \geq \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2n}\right),$$

where we note that $X_n = \mathbb{E}[X|Z_{1:n}] = f(Z_1, \ldots, Z_n) = X$ and $X_0 = \mathbb{E}[X]$. $\square$

Before we go over how to compute $\mathbb{E}[X]$ for general p, it's useful to first describe a quick sanity check for the case where $p = 1/2$. Note that if $p = 1/2$, then the problem of finding the maximal clique is complementary to the problem of finding independent sets. We've shown using second moment methods that the largest clique has size $2\log_2 n$ a.a.s. Therefore the chromatic number is a.a.s. at least

$$\frac{n}{2\log_2 n}(1 + o(1)).$$

Since the large deviation bound tells us that deviations of $\omega(\sqrt{n})$ are unlikely, and

$$\frac{n}{\log_2 n} \gg \sqrt{n},$$

it follows that the large deviation bound implies a tight concentration of the chromatic number.

We will need a more sophisticated martingale argument to actually compute $\mathbb{E}[X]$.

Theorem 7.14

For $G \in \mathcal{G}_{n,p}$, we have a.a.s. that

$$\mathbb{E}[X] \sim \frac{n}{2\log_{1/(1-p)} n}.$$

Proof. The lower bound is immediate from the clique argument above. In particular we get that a.a.s. the chromatic number is at least

$$\frac{n}{2\log_{1/(1-p)} n},$$

which gives us an asymptotic lower bound on $\mathbb{E}[X]$.

Note, however, that this does not give us an upper bound on the chromatic number, for the maximal clique size says nothing about all the smaller independent sets. For simplicity we will restrict our attention to the case where $p = 1/2$, but all the arguments will carry over for general p.

Recall from the proof of Theorem 2.7 that we defined

$$g(k) = \binom{n}{k} 2^{-\binom{k}{2}}$$

to be the expected number of k-cliques, and

$$k_0(n) = \max_k \{g(k) \geq 1\}$$

to be the largest integer such that $g(k)$ is still at least 1. We showed that $k_0(n) \sim 2\log_2 n$. Now, set $k_2(n) = k_0(n) - 3$. Then we can show that $g(k_2(n)) = n^{3+o(1)}$ by just plugging in $k = 2\log_2 - 3$ into $g(k)$.

With the following lemma, which we'll prove later, we can prove the theorem.

Lemma 7.15

For $G \in \mathcal{G}_{n,p}$,

$$\mathbb{P}[G \text{ contains no independent set of size} \geq k_2(n)] \leq \exp(-n^{2-o(1)}).$$

We already know from the proof of Theorem 2.7 that this probability goes to 0; the point of this lemma is to show that this probability converges steeply to zero.

The idea is the following algorithm:

while $\exists$ more than m uncolored vertices in G:
    pick an arbitrary uncolored subset $S \subseteq V(G)$ of size m
    pick a new color to apply to the largest independent set of S
color each remaining vertex of G a different new color

In particular, we pick $m = |S| = \frac{n}{(\log_2 n)^2}$. Let $G|_S \in \mathcal{G}_{m,1/2}$ be the restriction of G to a given S. Then by Lemma 7.15 $G|_S$ contains an independent set of size $k_2(m) \sim 2\log_2 m \sim 2\log_2 n$ with probability at least $1 - \exp(-m^{2-o(1)}) = 1 - \exp(-n^{2-o(1)})$, so by a union bound over all $\binom{n}{m}$ choices of S we have

$$\begin{aligned}\mathbb{P}[\exists S \text{ s.t. } G|_S \text{ contains no independent set of size } k_2(m)] &\le \binom{n}{m}\exp(-n^{2-o(1)})\\ &\le 2^n \exp(-n^{2-o(1)})\\ &= o(1).\end{aligned}$$

Therefore, in our algorithm, a.a.s. every iteration we will color at least $k_2(m)$ vertices, so the number of colors used is a.a.s. at most

$$\frac{n}{k_2} + m = \frac{n}{2\log_2 n}(1 + o(1)),$$

which is the desired upper bound on χ_G. □

Proof of Lemma 7.15. Here we will use a martingale argument with Azuma's inequality. In particular, let Y be the size of a maximal family of edge-disjoint $k_2(n)$-cliques of G. Then $Y = 0$ if and only if G contains no cliques of size $\ge k_2(n)$. Then, given the indicators $(Z_1, \ldots, Z_{\binom{n}{2}})$ for each edge, we have

$$Y = f(Z_1, \ldots, Z_{\binom{n}{2}})$$

for some function f. Note that since we assumed edge disjointness of the cliques, our function f is 1-Lipschitz. Furthermore the Z_i's are all independent, so we may apply Lemma 7.11 to the Doob martingale of Y to obtain

$$\begin{aligned}&\mathbb{P}[G \text{ contains no independent set of size } \ge k_2(n)]\\ &\le \mathbb{P}[G \text{ contains no clique of size } k_2(n)]\\ &\le \mathbb{P}[Y = 0]\\ &= \mathbb{P}[Y - \mathbb{E}[Y] \le -\mathbb{E}[Y]].\end{aligned}$$

It remains to compute $\mathbb{E}[Y]$. We use a probabilistic method type argument. Let K_2 be the set of all $k_2(n)$-cliques in G and let $\mu = \mathbb{E}[|K_2|] = g(k_2(n)) = n^{3+o(1)}$. Let P be the set of all pairs of $k_2(n)$-cliques with non-trivial intersection, i.e. contains at least 2 but no more than $k_2(n) - 1$ shared vertices. From the second moment calculations in 2.7, recall that

$$\frac{\mathbb{E}[|P|]}{\mu^2} \sim \frac{1}{2} \cdot \frac{k_2(n)^4}{n^2}.$$

Now let K' be a random subset of K_2 obtained by including each clique with probability $q \in (0, 1)$, and let P' be the remaining pairs of P once these cliques are chosen. Then

$$\begin{aligned}\mathbb{E}[|K'|] &= q\mu\\ \mathbb{E}[|P'|] &= q^2 \cdot \frac{1}{2} \cdot \frac{k_2(n)^4}{n^2}\mu^2\end{aligned}$$

If we remove from K' one element of each pair of P', this gives us an edge disjoint family of $k_2(n)$-cliques, so we get

$$\mathbb{E}[Y] \geq \mathbb{E}[|K'|] - \mathbb{E}[|P'|] \sim q\mu - q^2 \cdot \frac{1}{2} \cdot \frac{k_2(n)^4}{n^2}\mu^2.$$

Optimizing over q, we pick $q = \frac{n^2}{\mu k_2(n)^4} < 1$, so we get a lower bound on $\mathbb{E}[Y]$:

$$\mathbb{E}[Y] \geq \frac{n^2}{2k_2(n)^4}(1 + o(1)).$$

Finally, applying Azuma's inequality to the Doob martingale of Y with $c_i = 1$ for all i to obtain

$$\begin{aligned}
\mathbb{P}[Y - \mathbb{E}[Y] \leq -\mathbb{E}[Y]] &\leq \exp\left(-\frac{(\mathbb{E}[Y])^2}{2\binom{n}{2}}\right) \\
&\leq \exp\left(-\frac{n^2}{4k_2(n)^8}(1 + o(1))\right) \\
&= \exp(-n^{2-o(1)}).
\end{aligned}$$

□

7.2.3 Random Geometric TSP

In the geometric traveling salesman problem, we are given n points in the unit hypercube, and we wish to find L_n, the length of the shortest tour which visits each point $z_1, \ldots, z_n$ exactly once. In general, this is NP-complete. However, when each point $z_n \sim Z_n$ is chosen independently and u.a.r. in the cube, we can prove tight concentration bounds for L_n around its mean $\mathbb{E}[L_n]$. This is known as the random geometric traveling salesman problem (RGTSP).

First we state some results without proof, that we'll use later.

Theorem 7.16

For the d-dimensional RGTSP, we have

$$\mathbb{E}[L_n] \sim \gamma_d n^{\frac{d-1}{d}},$$

where γ_d is a constant depending on d.

Theorem 7.17 (Rhee '92)

For λ_d given above, we have

$$\lim_{d\to\infty} \frac{\gamma_d}{\sqrt{d}} = \frac{1}{\sqrt{2\pi e}} \approx 0.242.$$

We now illustrate the aspect of the problem which uses martingales.

Theorem 7.18

For RGTSP in $d = 2$ dimensions, we have

$$\mathbb{P}[|L_n - \mathbb{E}[L_n]| \geq \lambda] \leq 2\exp\left(-\frac{A\lambda^2}{\log n}\right),$$

for some universal constant A.

Therefore deviations of size $\omega(\sqrt{\log n})$ are unlikely. Since $\mathbb{E}[L_n] \sim \Theta(\sqrt{n})$, this implies we have a tight concentration about the mean.

Proof. Let $f(Z_1, \ldots, Z_n)$ be the length of the shortest TS tour given points $\{Z_i\}_{i=1}^n$. As usual, consider the Doob martingale

$$X_i = \mathbb{E}[f(Z_1, \ldots, Z_n)|Z_{1:i}].$$

It's easy to see that our function f is c-Lipschitz for some constant c, but it turns out this is not enough. In particular, we'd only get a tail bound of the form $\exp(-\frac{A\lambda^2}{n})$, which only tells us that deviations of $\omega(\sqrt{n})$ are unlikely. But this is on the same order as $\mathbb{E}[L_n]$, so it wouldn't give us a tight concentration we are looking for.

So, to get a better bound on the differences, we write

$$X_i - X_{i-1} = \mathbb{E}[f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, \hat{Z}_i, \ldots, Z_n)|Z_{1:i}],$$

where $\hat{Z}_i$ has the same distribution as Z_i but is independent of all the $Z_{1:n}$. Then define

$$\Delta_i = |f(Z_1, \ldots, Z_i, \ldots, Z_n) - f(Z_1, \ldots, \hat{Z}_i, \ldots, Z_n)|.$$

Observe that by a standard triangle inequality argument, we have for any set of points S,

$$f(S) \leq f(S \cup \{z\}) \leq f(S) + 2\min_{y \in S} |y - z|.$$

So, with $S = \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\}$, we have

$$\Delta_i \leq 2[q(Z_i) - q(\hat{Z}_i)],$$

where $q(z)$ is the shortest distance from the point z to the set $\{Z_{i+1}, \ldots, Z_n\}$. Taking conditional expectations, we have

$$X_i - X_{i-1} \leq \mathbb{E}[\Delta_i|Z_{1:i}] \leq 2\mathbb{E}[q(Z_i) + q(\hat{Z}_i)|Z_{1:i}] \leq 4\mathbb{E}[Q_i],$$

where we define the random variable Q_i to be the shortest distance from a fixed point z to $n - i$ randomly selected points in the square. So, by symmetry of Z_i and $\hat{Z}_i$, we have

$$|X_i - X_{i-1}| \leq 4\mathbb{E}[Q_i].$$

We may compute the RHS as follows:

$$\begin{aligned}
\mathbb{E}[Q_i] &= \int_0^\infty \mathbb{P}[Q_i > r]dr \\
&\leq \int_0^{\sqrt{2}} (1 - Cr^2)^{n-i} dr \\
&\leq \int_0^{\sqrt{2}} \exp\{-Cr^2(n-i)\} dr \\
&\leq \frac{D}{\sqrt{n-i}},
\end{aligned}$$

where C is some constant chosen so that the area of the r-disc centered at z has at most Cr^2 area inside the unit square, and D is just some other constant that comes out when integrating. Therefore we have the bounded deviations

$$|X_i - X_{i-1}| \leq \frac{D}{\sqrt{n-i}} =: c_i,$$

for $1 \leq i < n$. For $i = n$, we can just use the trivial bound of $|X_n - X_{n-1}| \leq 4\sqrt{2} =: c_n$. Finally, by Azuma's inequality, we get

$$\begin{aligned}
\mathbb{P}[|L - \mathbb{E}[L_n]| \geq \lambda] &\leq 2\exp\left(-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}\right) \\
&= 2\exp\left(-\frac{\lambda^2}{2\left[(4\sqrt{2})^2 + \sum_{i=1}^{n-1} \frac{D^2}{n-i}\right]}\right) \\
&\leq 2\exp\left(-\frac{A\lambda^2}{\log n}\right),
\end{aligned}$$

for some constant A, as desired. □

7.3 The Optional Stopping Theorem

Definition 7.19. Let $(\mathcal{F}_i)$ be a filtration. A random variable $T \in \mathbb{N}_0 \cup \{\infty\}$ is a stopping time with respect to $(\mathcal{F}_i)$ if the event $\{T = i\}$ is $\mathcal{F}_i$ measurable. In more concrete words, this means there is no look-ahead. That is, whether we stop at time t only depends on the past and present, but not future times $\{t+1, t+2, \ldots\}$.

We know that in general, given a martingale (X_i) with respect to a filtration $(\mathcal{F}_i)$, we have

$$\mathbb{E}[X_i] = \mathbb{E}[X_0],$$

which follows from the tower property. But we may instead be interested in a random stopping time T, and whether we still have

$$\mathbb{E}[X_T] = \mathbb{E}[X_0].$$

It turns out there are certain sufficient conditions for this to hold, but first we give a counterexample.

Example 7.20. Consider a sequence of fair coin tosses, and let $X_i = \#\text{heads} - \#\text{tails}$ of the first i tosses. Then X is a martingale, and $\mathbb{E}[X_0] = 0$. If T is the first time such that $X_i \geq 17$, then we have

$$\mathbb{E}[X_T] = 17 \neq \mathbb{E}[X_0].$$

The reason equality fails here is because $\mathbb{E}[T] = \infty$.

Theorem 7.21 (Optional Stopping Theorem)

Let (X_i) be a martingale and T be a stopping time, both w.r.t. the filter $(\mathcal{F}_i)$. Then

$$\mathbb{E}[X_T] = \mathbb{E}[X_0],$$

provided the following conditions hold:

1. $\mathbb{P}[T < \infty] = 1$.
2. $\mathbb{E}[|X_T|] < \infty$.
3. $\mathbb{E}[X_i \mathbf{1}_{\{T>i\}}] \to 0$ as $i \to \infty$.

Alternatively, we can also use the following set of stronger conditions, which are often easier to verify in practice:

1. $\mathbb{E}[T] < \infty$.
2. $\mathbb{E}[|X_i - X_{i-1}||\mathcal{F}_i] \leq c$ for all i and some uniform constant c.

We illustrate the power of the Optional Stopping Theorem on a classic example.

Example 7.22 (Gambler's Ruin)**.** Consider a gambling game starting at 0 capital. At each time step, the gambler flips a fair coin and gains 1 capital of heads and loses 1 capital of tails. If the gambler reaches $-a$ capital he loses, and if he reaches $+b$ capital he wins. We are interested in the probability of winning,

$$p := \mathbb{P}[\text{win}],$$

as well as the expected time it takes to play this game,

$$\mathbb{E}[T].$$

First, let's see what we can do if the Optional Stopping Theorem holds. If X_t denotes the gambler's capital at time t, then note that (X_t) is a martingale w.r.t. the sequence of coin flips (Z_t). So we'd get $\mathbb{E}[X_T] = \mathbb{E}[X_0] = 0$. But this implies

$$p(b) + (1-p)(-a) = 0 \to p = \frac{a}{a+b}.$$

So it remains to verify the conditions of Theorem 7.21. Note that

$$\mathbb{P}[X_t \text{ hits } -a \text{ or } b \text{ within } \max\{a,b\} \text{ steps}] \geq \frac{1}{2^{\max\{a,b\}}}.$$

Then if we consider the random variable which indicates whether X_t has hit $-a$ or b in $k \cdot \max\{a, b\}$ time steps, we see that it is stochastically dominated by a $\mathrm{Ber}(2^{-\max\{a,b\}})$ random variable, which has finite expectation. Therefore $\mathbb{E}[T] < \infty$. Furthermore, it's clear that $|X_i - X_{i-1}| \leq 1$ for all i. Therefore we were justified in using the equality $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.

Next, we aim to compute $\mathbb{E}[T]$. The idea is still to use the Optional Stopping Theorem, but with a neat trick. Consider the random variable

$$Y_i = X_i^2 - i.$$

The intuition behind this choice is we want to be able to use a random variable which is still a martingale, but also allows us to extract T from the subscript:

$$\mathbb{E}[Y_T] = \mathbb{E}[X_T^2] - \mathbb{E}[T].$$

Indeed (Y_i) is a martingale w.r.t. (Z_i) because

$$\begin{aligned}
\mathbb{E}[Y_i | Z_{1:i-1}] &= \mathbb{E}[X_i^2 - i | Z_{1:i-1}] \\
&= \frac{1}{2}((X_{i-1} + 1)^2 - i) + \frac{1}{2}((X_{i-1} - 1)^2 - i) \\
&= X_{i-1}^2 - (i - 1) = Y_{i-1}.
\end{aligned}$$

Furthermore, we have

$$\mathbb{E}[|Y_i - Y_{i-1}| Z_{1:i-1}] \leq 2 \max\{a, b\} + 1.$$

So, we may apply Theorem 7.21 to obtain

$$0 = \mathbb{E}[Y_0] = \mathbb{E}[Y_T] = \mathbb{E}[X_T^2] - \mathbb{E}[T] = \frac{a}{a+b} \cdot b^2 + \frac{b}{a+b} \cdot (-a)^2 - \mathbb{E}[T],$$

which tells us that

$$\mathbb{E}[T] = ab.$$

Note how simple these derivations were! If we were to go about solving for p and $\mathbb{E}[T]$ in the usual way with markov chains, we'd come up with a nasty recurrence relation and a system of equations that scales in size with $a + b$. Indeed, most of the work is actually encoded in the proof of the Optional Stopping Theorem.

Definition 7.23. A sequence of random variables (X_i) is a *submartingale* w.r.t. a filter $(\mathcal{F}_i)$ if

$$\mathbb{E}[X_i | \mathcal{F}_{i-1}] \geq X_{i-1}.$$

Likewise, (X_i) is called a *supermartingale* if

$$\mathbb{E}[X_i | \mathcal{F}_{i-1}] \leq X_{i-1}.$$

It turns out the Optional Stopping Theorem has generalizations to supermartingales and submartingales. In particular, if the same conditions holds, we have

$$\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$$

if (X_i) is a supermartingale, and

$$\mathbb{E}[X_T] \geq \mathbb{E}[X_0]$$

if (X_i) is a submartingale.

Example 7.24 (Gambler's Ruin with Drift)**.** We can generalize the previous example to the case where there is drift in the random walk. We will also consider the additional property where there is a reflecting barrier at one end of the interval, and we want to know how long it takes to reach the end without the barrier. Let $D_i = X_i - X_{i-1}$ be the difference sequence, and consider the supermartingale (X_i) defined on $[0, n]$ with $X_0 = s$. We assume

$$\begin{aligned} \mathbb{E}[D_i | X_{1:i-1}] &\leq 0 \\ \mathbb{E}[D_i^2 | X_{1:i-1}] &\geq \sigma^2. \end{aligned}$$

The first is just the supermartingale property, and the second gives us a lower bound on the jump sizes. Additionally, if $X_{i-1} = n$, we assume $X_i = n - 1$ with probability 1. We are interested in a bound for $\mathbb{E}[T]$, where T is the number of time steps to reach 0.

Again, we will pick an auxilliary sequence of random variables

$$Y_i = X_i^2 + \lambda X_i + \mu i.$$

The best way to gain intuition for these sorts of choices is probably to reverse-engineer it, i.e. think about what we want and make some guesses as to what could produce the desired quantities. It's also illustrative to run through the proof using only a linear functions of X_i to see why we really do need a quadratic function. In particular, we will pick λ and μ so that (Y_i) is a submartingale. We write

$$\begin{aligned} \mathbb{E}[Y_i | X_{1:i-1}] &= \mathbb{E}[(X_{i-1} - D_i)^2 + \lambda(X_{i-1} + D_i) + \mu i | X_{1:i-1}] \\ &= X_{i-1}^2 + \lambda X_{i-1} + \mu i + (2X_{i-1} + \lambda) \cdot \mathbb{E}[D_i | X_{1:i-1}] + \mathbb{E}[D_i^2 | X_{1:i-1}] \\ &= Y_{i-1} + (2X_{i-1} + \lambda) \cdot \mathbb{E}[D_i | X_{1:i-1}] + (\mathbb{E}[D_i^2 | X_{1:i-1}] + \mu). \end{aligned}$$

Given our assumptions, it suffices to take $\lambda = -2n$ and $\mu = -\sigma^2$ (of course, any $\mu > -\sigma^2$ works too, but this won't give us as tight of a bound). By the Optional Stopping Theorem for submartingales, we have

$$\begin{aligned} \mathbb{E}[Y_T] &\geq \mathbb{E}[Y_0] \\ \mathbb{E}[X_T^2] - 2n\mathbb{E}[X_T] - \sigma^2 \mathbb{E}[T] &\geq s^2 - 2ns \\ \mathbb{E}[T] &\leq \frac{2ns - s^2}{\sigma^2} \leq \frac{n^2}{\sigma^2}. \end{aligned}$$

It's easy to check the conditions of the Optional Stopping Theorem with similar arguments as before, so we're done.

Example 7.25 (2-SAT)**.** Recall that 2-SAT can be solved in poly-time using strongly connected components of directed graphs. Here we illustrate a randomized algorithm whose analysis we can recycle from the previous example.

In particular, consider the following algorithm. Suppose we're given a 2-CNF formula φ with n variables, and suppose that a satisfying assignment exists. Then start with an arbitrary initial assignment a_0. At each step, pick an arbitrary unsatisfied clause C_i and flip one of its literals uniformly at random. Proceed for $O(n^2)$ iterations.

If a^* is some satisfying assignment of φ, let X_i denote the Hamming distance between a_i and a^*. Then note that

$$\begin{aligned}\mathbb{P}[|X_i - X_{i-1}| = 1] &= 1\\ \mathbb{P}[X_i - X_{i-1} = -1] &\geq \frac{1}{2}.\end{aligned}$$

As before, define $D_i = X_i - X_{i-1}$. Then we have

$$\begin{aligned}\mathbb{E}[D_i|X_{1:i-1}] &\leq 0\\ \mathbb{E}[D_i^2|X_{1:i-1}] &= \sigma^2 = 1.\end{aligned}$$

So we can recycle the analysis from before to get

$$\mathbb{E}[\text{steps to go from } a_0 \text{ to } a^*] \leq \frac{n^2}{\sigma^2} = n^2.$$

Example 7.26 (The Ballot Problem)**.** Suppose we have two candidates A and B, each with a and b votes respectively. We assume $a > b$, and we want to find the probability that A always stays strictly ahead (aside from the beginning) if the votes are counted in any random order. The answer turns out to be

$$\frac{a-b}{a+b},$$

which can be found combinatorially using reflection arguments. We give a martingale proof.

Let S_k be the number of votes for A minus the number of votes for B after k votes are counted. So $S_0 = 0$ and $S_n = a - b$. Define the auxiliary variable

$$X_k = \frac{S_{n-k}}{n-k}.$$

Then (X_k) will be our "backwards" martingale with respect to the vote counting. The reason we chose backwards rather than the more obvious choice of forwards, is because it's easier to mold the backward conditional expectation into a martingale. In particular, through standard combinatorial arguments, we get that

$$\mathbb{E}[S_{n-k}|B_{n-k+1:n}] = S_{n-k+1} \cdot \frac{n-k}{n-k+1}.$$

Therefore we can easily mold this by introducing a factor of $\frac{1}{n-k}$ and our sequence of random variables becomes a martingale! Now, let $T = \min\{k < n : X_k = 0\}$ or

$T = n - 1$ if no such k exists. Then if A is always ahead, $X_T = X_{n-1} = S_1 = 1$, and if A is not always ahead, then $X_T = 0$. Clearly the conditions of Theorem 7.21 hold, so we have

$$p = \mathbb{E}[X_T] = \mathbb{E}[X_0] = \frac{a-b}{a+b}.$$

Note the recurring theme of tinkering with the random variable and molding it so that it becomes a martingale, from which we may then apply Optional Stopping Theorem. Very rarely will our random variables already be a martingale. Indeed, the transforms for the Gambler's Ruin problems look like

$$Y_i = X_i^2 - i, \quad Y_i = X_i^2 + \lambda X_i + \mu i,$$

and for the Ballot problem we used

$$X_k = \frac{S_{n-k}}{n-k}$$

which is even a backwards martingale.

7.3.1 A Proof of Wald's Identity

Theorem 7.27

Suppose (X_i) are i.i.d. random variables and T is a stopping time so that $|\mathbb{E}[X_i]| < \infty$ and $\mathbb{E}[T] < \infty$. Then

$$\mathbb{E}\left[\sum_{i=1}^{T} X_i\right] = \mathbb{E}[T]\mathbb{E}[X_1]. \tag{15}$$

Proof of Nonnegative Case with Martingales. It turns out that Optional Stopping Theorem is only strong enough to prove the cases where either $X_i \geq 0$ or $\mathbb{E}[|X_i|] \leq c$ for all i.

To get intuition for how we will choose our auxiliary variable, we work backwards from what we want to prove. The identity (15) is equivalent to

$$\mathbb{E}\left[\sum_{i=1}^{T} (X_i - \mu)\right] = 0$$

So define the sequence

$$Y_i = \sum_{k=1}^{i} (X_k - \mu).$$

It is a martingale, for

$$\begin{aligned}\mathbb{E}[Y_i|X_{1:i-1}] &= \sum_{k=1}^{i-1}(X_k-\mu)+\mathbb{E}[X_i]-\mu \\ &= \sum_{k=1}^{i-1}(X_k-\mu) \\ &= Y_{i-1}.\end{aligned}$$

Furthermore, we have bounded differences, since

$$\begin{aligned}\mathbb{E}[|Y_i-Y_{i-1}||X_{1:i-1}] &= \mathbb{E}[|X_i-\mu||X_{1:i-1}] \\ &\le 2\mu < \infty,\end{aligned}$$

since $X_i \ge 0$, or if $\mathbb{E}[|X_i|] \le c$ for all i then we have the bound of $c+\mu$.

Thus, we apply Theorem 7.21 to get

$$0 = \mathbb{E}[Y_0] = \mathbb{E}[Y_T] = \mathbb{E}\left[\sum_{i=1}^{T}(X_i-\mu)\right],$$

which gives us what we want. □

The proof of the more general statement (where X_i are not assumed to be nonnegative or absolutely integrable) uses basic principles, writing out the LHS as a summation, swapping the summation outside the expectation with DCT, and then using a tail sum formula to simply into the RHS.

8 The Lovász Local Lemma

Recall that in the probabilistic method, we prove existence of some object by showing that it occurs with positive probability. Yet in many applications of the probabilistic method, this probability may be high, or even tend to 1 as $n \to \infty$, which leads to an efficient randomized algorithm via sampling.

In particular, we consider a set of "bad" events $\{A_1, \ldots, A_n\}$, whose occurence nullifies our desired object. That is, we may wish to compute the probability all these events are avoided,

$$\mathbb{P}[\cap_{i=1}^n \overline{A}_i].$$

If the probabilities are independent, and each probability has the bound $\mathbb{P}[A_i] \leq p$, then the above probability is at least $(1-p)^n$, which tends to 0 for $n \to \infty$.

The Lovász Local Lemma can be thought of as an extension to the case where there are limited dependencies between the events. Furthermore, based on the exponentially decreasing probability in the fully independent case, we'd expect our desired object to occur with low probability, often exponentially small. That is, we are looking for a "needle in a haystack" when it comes to the LLL. Therefore it does not immediately give us an efficient randomized algorithm (but there is one).

Definition 8.1. An event A is said to be mutually independent of a set of events $\{B_i\}$ if for any subset S of events or their complements contained in $\{B_i\}$, we have $\mathbb{P}[A|S] = \mathbb{P}[A]$.

Theorem 8.2 (Lovász Local Lemma)

Let $\{A_1, \ldots, A_n\}$ be a set of "bad" events with $\mathbb{P}[A_i] \leq p < 1$, such that each event A_i is mutually independent of all but at most d of the other A_j. If $e \cdot p(d+1) \leq 1$ or $4pd \leq 1$ (which is slightly stronger for $d \leq 2$, but asymptotically weaker otherwise), then

$$\mathbb{P}\left[\bigcap_{i=1}^n \overline{A}_i\right] > 0.$$

Proof. First we expand with the chain rule to get

$$\mathbb{P}\left[\bigcap_{i=1}^n \overline{A}_i\right] = \prod_{i=1}^n \left(1 - \mathbb{P}\left[A_i \,\middle|\, \bigcap_{j<i} \overline{A}_j\right]\right).$$

It suffices to find a uniform bound over all terms of the form

$$\mathbb{P}\left[A_i \,\middle|\, \bigcap_{j \in S} \overline{A}_j\right] \leq \frac{1}{d+1},$$

for any strict subset $S \subset \{1, \ldots, n\}$ and any $i \in \{1, \ldots, n\}$. We proceed by induction on $m := |S|$. The base case $m = 0$ is true since

$$\mathbb{P}[A_i] \leq p \leq \frac{1}{e(d+1)} < \frac{1}{d+1}.$$

For the inductive step, partition S into two sets $S_1 = S \cap D_i$ and $S_2 = S \setminus S_1$, where D_i is the dependency set of A_i among all the A_j. Note that $|D_i| \leq d$. We write

$$\begin{aligned}
\mathbb{P}\left[A_i \middle| \bigcap_{j\in S} \overline{A}_j\right] &= \frac{\mathbb{P}\left[A_i \cap (\bigcap_{j\in S_1} \overline{A}_j) \,|\bigcap_{k\in S_2} \overline{A}_k\right]}{\mathbb{P}\left[\cap_{j\in S_1} \overline{A}_j \,|\bigcap_{k\in S_2} \overline{A}_k\right]} \\
&\leq \frac{\mathbb{P}\left[A_i \,|\bigcap_{k\in S_2} \overline{A}_k\right]}{\mathbb{P}\left[\cap_{j\in S_1} \overline{A}_j \,|\bigcap_{k\in S_2} \overline{A}_k\right]} \\
&= \frac{\mathbb{P}[A_i]}{\mathbb{P}\left[\cap_{j\in S_1} \overline{A}_j \,|\bigcap_{k\in S_2} \overline{A}_k\right]}.
\end{aligned}$$

To lower bound the denominator, first denote $S_1 = \{j_1, \ldots, j_r\}$, and assume w.l.o.g. that $r > 0$. Then we can expand by chain rule to get

$$\begin{aligned}
\mathbb{P}\left[\cap_{j\in S_1} \overline{A}_j \middle| \bigcap_{k\in S_2} \overline{A}_k\right] &= \prod_{l=1}^{r} \left(1 - \mathbb{P}\left[A_{j_l} \middle| \left(\bigcap_{l'<l} \overline{A}_{j_{l'}}\right) \cap \left(\bigcap_{k\in S_2} \overline{A}_k\right)\right]\right) \\
&\geq \left(1 - \frac{1}{d+1}\right)^d \\
&> \frac{1}{e}.
\end{aligned}$$

Putting this together, we get

$$\mathbb{P}\left[A_i \middle| \bigcap_{j\in S} \overline{A}_j\right] \leq \frac{\mathbb{P}[A_i]}{1/e} \leq e \cdot p \leq \frac{1}{d+1}.$$

Thus, going back to the beginning of the proof, we get

$$\mathbb{P}\left[\bigcap_{i=1}^{n} \overline{A}_i\right] = \left(1 - \frac{1}{d+1}\right)^n > 0$$

□

8.1 Existence of Satisfying K-SAT Assignment

Theorem 8.3

Any k-SAT formula φ in which no variable appears in more than $\frac{2^{k-2}}{k}$ clauses is satisfiable.

Proof. Fix a k-SAT formula φ. Pick a truth assignment to the variables of φ u.a.r. and let A_i denote the event where clause i is not satisfied. Note that exactly one out of 2^k possible assignments fails to satisfy a particular clause, so write

$$\mathbb{P}[A_i] = 2^{-k} =: p \quad \forall i \in \{1, 2, \ldots, n\}.$$

Furthermore, the dependency set has size given by

$$d \leq k \cdot \frac{2^{k-2}}{k} = 2^{k-2}.$$

So, we may use the condition $4pd \leq 1$ since $\frac{1}{2^k} = p \leq \frac{1}{4d} = \frac{1}{4 \cdot 2^{k-2}}$. So LLL implies

$$\mathbb{P}\left[\bigcap_{i=1}^{n} \overline{A}_i\right] > 0,$$

so there must exist a satisfying assignment. □

8.2 Algorithmic Lovász Local Lemma

We propose the following intuitive "local search" algorithm for finding a satisfying assignment:

```
Initialize with independent random assignment for each Z_j.
While there exists violated clauses:
    Choose a violated clause i arbitrarily.
    Resample all variables Z_j of clause i.
Return the assignment {Z_j}.
```

Let A_i be the event that clause i is violated, i.e. its assignment of variables evaluates to false. Define D_i to be the neighborhood of dependencies of clause i. In our k-SAT setting this is just the set of other clauses j that share a variable with clause i. Further define D_i^+ to be the dependency set D_i augmented with clause i itself. Then we have the following result regarding the above algorithm.

Theorem 8.4

In above setting, if there exists real numbers $x_i \in (0, 1)$ such that

$$\mathbb{P}[A_i] \leq x_i \prod_{j \in D_i} (1 - x_j) \quad \text{for all } i,$$

then the algorithm finds a satisfying assignment of the $\{Z_j\}$ in expected time at most $\sum_i \frac{x_i}{1-x_i}$.

Proof. Let E refer to the execution of the algorithm, and $E(1), E(2), \dots, E(t)$ refer to the violated event A_i picked at times $1, 2, \dots, t$. Then define N_i to be the number of times A_i is picked throughout E. We wish to bound the expected running time, which is just

$$\sum_i \mathbb{E}[N_i].$$

The proof can be divided into two parts:

(i) First, we construct combinatorial objects T called "witness trees", and use them to massage our probabilities into an easier to handle form:

$$\mathbb{E}[N_i] = \sum_{T:\text{root}(T)=A_i} \mathbb{P}[T \text{ occurs in } E] \leq \sum_{T:\text{root}(T)=A_i} \prod_{v \in V(T)} \mathbb{P}[A_{[v]}], \tag{16}$$

where $A_{[v]}$ is the violated event A_i corresponding to vertex v of the witness tree.

(ii) Dominate the above probabilities by mapping into a *multi-type Galton-Watson branching process* (which is just a Galton-Watson branching process but the branching probabilities are allowed to be different). In particular, we make use of the following equality:

$$p_T := \mathbb{P}[\text{GW process yields tree } T] = \frac{1-x_i}{x_i} \prod_{v \in V(T)} x'_{[v]}, \tag{17}$$

where $x'_i := x_i \prod_{j \in D_i}(1 - x_j)$. Note that the initial assumption of the theorem where $\mathbb{P}[A_i] \leq x'_i$ will then allow us to map the terms $\mathbb{P}[A_{[v]}]$ of (16) to probabilities in the Galton-Watson space. This will then give us our final bound of $\sum_i \frac{x_i}{1-x_i}$.

First, let's define a witness tree. Given an execution E, define the witness tree $T(t)$ for each time step t of E as follows. Label the root with $E(t)$. Then, iterating from $i = t-1, t-2, \dots$ backward in time, attach a node labeled with the event $E(i)$ as a child of the deepest node with label in $D^+_{E(i)}$, breaking ties arbitrarily. If there is no event that $E(i)$ depends on already existing in the tree, then do not attach the node for $E(i)$ to the tree.

We say that a witness tree T *occurs* in E if $T = T(t)$ for some t. We call a witness tree *proper* if for each node, all children of that node are distinct.

It's easy to see that if T occurs in E, then it must be proper based on construction. Indeed, if some event A_i is already attached for some $E(t_2)$, and we are looking to attach the event A_j for some $E(t_1)$, where $t_1 < t_2$ and A_j shares variables with A_i, then we must attach $E(t_1)$ at least as low as $\text{depth}(E(t_2)) + 1$. In particular, no two nodes at the same level may share any variables, so any two nodes at the same level are distinct.

Now, define an *evaluation* of T as follows. In reverse level order, visit the nodes of T and resample their variables, independently of previous resamplings. We say that the evaluation *succeeds* if all events were violated upon being resampled by the reverse level order sweep. Therefore we have

$$\mathbb{P}[\text{evaluation succeeds}] = \prod_{v \in V(T)} \mathbb{P}[A_{[v]}]. \tag{18}$$

To complete the proof of (16), we just need to show

$$\mathbb{P}[T \text{ occurs in } E] \leq \mathbb{P}[\text{evaluation succeeds}]. \tag{19}$$

We do this via a *coupling* technique. In particular, use the same source of randomness for both the execution of the algorithm, and the evaluation of T. Define, for each variable Z_j, and infinite sequence of random independent boolean values (this is like setting a random seed for a computer). Then when either the execution or the evaluation needs to sample Z_j, it takes the next value in the sequence, so that the algorithm and the evaluation both take the same value for a given variable if it has been sampled the same number of times in both processes.

Now, consider a time when Z_j is being sampled in the evaluation, at some node v of T. Since T is proper, the number of times Z_j has already been sample before is precisely the number of nodes lower than v that depend on Z_j. Call this number $n_{j,v}$.

Next, consider a time when Z_j is being sampled in the execution. Going back to how we constructed the witness tree T, the number of times we have sample Z_j before is precisely $n_{j,v} + 1$, since we sampled Z_j once at the beginning to initialize it.

Therefore, when the evaluation resamples at a node v, it will assign each variable of $A_{[v]}$ to the value it was assigned in the execution of the algorithm immediately before resampling $A_{[v]}$. But if the algorithm chose $A_{[v]}$ to resample, then $A_{[v]}$ must have been violated in the first place. So, the evaluation will violate the event at each node in its reverse level order sweep, proving (19).

Putting (18) together with (19), this gives us (16). So, we move on to the second part of the proof.

We define a multi-type Galton-Watson branching process as follows. We specify a root A_i, and recursively, for each $A_j \in D_i^+$ independently, we add it as a child of A_i with probability x_j. Then if p_T is the probability this process yields a specific tree T, then we may compute (below we define $W_v \subseteq D^+_{A_{[v]}}$ to be those events which do not occur as children of v in T):

$$\begin{aligned}
p_T &= \frac{1}{x_i} \prod_{v \in V(T)} \left(x_{[v]} \prod_{u \in W_v} (1 - x_{[u]}) \right) \\
&= \frac{1 - x_i}{x_i} \prod_{v \in V(T)} \left(\frac{x_{[v]}}{1 - x_{[v]}} \prod_{u \in D^+_{A_{[v]}}} (1 - x_{[u]}) \right) \\
&= \frac{1 - x_i}{x_i} \prod_{v \in V(T)} x_{[v]} \left(\prod_{u \in D_{A_{[v]}}} (1 - x_{[u]}) \right) \\
&= \frac{1 - x_i}{x_i} \prod_{v \in V(T)} x'_{[v]},
\end{aligned}$$

proving (17).

Finally, we may bound the expectation as follows:

$$\begin{aligned}
\mathbb{E}[N_i] &= \sum_{T:\mathrm{root}(T)=A_i} \mathbb{P}[T \text{ occurs in } E] \\
&\le \sum_{T:\mathrm{root}(T)=A_i} \prod_{v\in V(T)} \mathbb{P}[A_{[v]}] \\
&\le \sum_{T:\mathrm{root}(T)=A_i} \prod_{v\in V(T)} x'_{[v]} \\
&= \frac{x_i}{1-x_i} \sum_{T:\mathrm{root}(T)=A_i} p_T \\
&\le \frac{x_i}{1-x_i}.
\end{aligned}$$

The second line comes from (16), the third line comes from the assumption in the theorem, the fourth line comes from (17), and the final line comes from the fact that the trees are distinct throughout the algorithm (each $T(t)$ contains a different number of A_i's), so that $\sum_T p_T \le 1$.

We can think of the whole second half of the proof as a sort of stochastic dominance between models. Specifically, by (16) we have mapped our probabilities into a sort of "independent" model. Then, we consider the Galton-Watson process as a generative model for our witness trees, and map the separate probability terms $\prod_{v\in V(T)} \mathbb{P}[A_{[v]}]$ to trees generated by the branching process, with a dominating fudge factor of $\frac{x_i}{1-x_i}$. Then since distinct events in our generative model can have sum of probabilities at most 1, this then gives us our desired result. $\square$