The Probabilistic Method in Combinatorics

Lecturer: Professor Yufei Zhao

Notes by: Andrew Lin

Spring 2019

This is an edited transcript of the lectures of MIT's Spring 2019 class **18.218: The Probabilistic Method in Combinatorics**, taught by Professor Yufei Zhao.

Each section focuses on a different technique, along with examples of applications. Additional course material, including problem sets, can be found on the course website.

The main reference for the material is the excellent textbook

> N. Alon and J. H. Spencer, *The probabilistic method*, Wiley, 4ed.

Most of the course will follow the textbook, though some parts will differ.

Please contact Yufei Zhao (yufeiz@mit.edu) and Andrew Lin (lindrew@mit.edu) for any questions or comments regarding these notes. Special thanks to Abhijit Mudigonda, Mihir Singhal, Andrew Gu, and others for their help in proofreading.

Contents

Definition 0.1 (Asymptotic notation)

Given functions or sequences $f, g > 0$ (usually of some parameter $n \to \infty$), the notation in each bullet point below are considered equivalent:

- $f \lesssim g$, $f = O(g)$, $g = \Omega(f)$, $f \leq Cg$ (for some constant C);
- $f \ll g$, $f = o(g)$, $\frac{f}{g} \to 0$, $g = \omega(f)$.
- $f \asymp g$, $f = \Theta(g)$, $g \lesssim f \lesssim g$.
- $f \sim g$, $\frac{f}{g} \to 1$, $f = (1 + o(1))g$.

Some event holds *with high probability* if its probability is $1 - o(1)$.

Warning: analytic number theorists like to use the Vinogradov notation, where $f \ll g$ means $f = O(g)$ instead of $f = o(g)$ as we do. In particular, $100 \ll 1$ is correct in Vinogradov notation.

1 Introduction to the probabilistic method

In combinatorics and other fields of math, we often wish to show existence of some mathematical object. One clever way to do this is to try to construct this object randomly and then show that we succeed with positive probability.

Proposition 1.1

Every edge $G = (V, E)$ with vertices V and edges E contains a bipartite subgraph with at least $\frac{|E|}{2}$ edges.

Proof. We can form a bipartite graph by partitioning the vertices into two groups. Randomly color each vertex either white or black (making the white and black sets the two groups), and include only the edges between a white and a black edge in a new graph G'. Since all vertices are colored independently at random, each edge is included in G' with probability $\frac{1}{2}$. Thus, we have an average of $\frac{|E|}{2}$ edges in our graph by linearity of expectation, and this means that at least one coloring will work. □

This class will introduce a variety of methods to solve these types of problems, and we'll start with a survey of those techniques.

1.1 The Ramsey numbers

Definition 1.2

Let the **Ramsey number** $R(k, \ell)$ be the smallest n such that if we color the edges of K_n (the complete graph on n vertices) red or blue, we always have a K_k that is all red or a K_ℓ that is all blue.

Theorem 1.3 (Ramsey, 1929)

For any integers k, ℓ, $R(k, \ell)$ is finite.

One way to do this is to use the recurrence inequality

$$R(r, s) \leq R(r - 1, s) + R(r, s - 1)$$

by picking an arbitrary vertex v and partitioning the remaining vertices by the color of their edge to v.

Theorem 1.4 (Erdős, 1947)
We have $R(k,k) > n$ for all
$$\binom{n}{k} 2^{1-\binom{k}{2}} < 1.$$

In other words, for any n that satisfies this inequality, we can color K_n with no monochromatic K_k.

Proof. Color the edges of K_n randomly. Given any set R of k vertices, let A_R be the event where R is monochromatic (all $\binom{k}{2}$ edges are the same color). The probability A_R occurs for any given R is $2^{1-\binom{k}{2}}$, since there are only 2 ways to color R, and thus the total probability that K_n is monochromatic is
$$\Pr\left[\bigcup_{R\in\binom{[n]}{k}} A_R\right]$$
and we can "union bound" this: the total probability is at most the sum of the probabilities of the independent events, so
$$\Pr(\text{monochromatic}) \le \sum_R \Pr(A_R) = \binom{n}{k} 2^{1-\binom{k}{2}},$$
and as long as this is less than 1, there is a positive probability that no monochromatic coloring exists, and thus $R(k,k) > n$. □

Fact 1.5
We can optimize Theorem 1.4 with Stirling's formula to find that
$$R(k,k) > \left(\frac{1}{e\sqrt{2}+o(1)}\right) k2^{k/2},$$
where the $o(1)$ term goes to 0 as $k \to \infty$.

This is a lower bound on the Ramsey numbers. It turns out we can also get an upper bound
$$R(s,s) \le \left(\frac{1}{4\sqrt{\pi}} + o(1)\right)\frac{4^s}{\sqrt{s}}.$$

Currently, this is basically the best we can do: it is still an open problem to make the bases of the exponents tighter than $\sqrt{2}$ and 4.

Remark. *Because the name is Hungarian, the "s" in Erdős is pronounced as "sh," while "sz" is actually pronounced "s."*

1.2 Alterations

We can almost immediately improve our previous bound by a bit.

Proposition 1.6
For all k, n, we have
$$R(k,k) > n - \binom{n}{k} 2^{1-\binom{k}{2}}.$$

Proof. As before, color the edges of K_n randomly. This time, let A_R be the **indicator variable** for a set R of k vertices. (This means that A_R is equal to 1 if R is monochromatic and 0 otherwise.) The expected value of each A_R is just the probability that R is monochromatic, which is $2^{1-\binom{k}{2}}$, so the expected number of monochromatic K_ks is the sum of all A_Rs, which is

$$\mathbb{E}[X] = \binom{n}{k} 2^{1-\binom{k}{2}}.$$

Now delete one vertex from each monochromatic k-clique: we delete X vertices at most (possibly with repeats), so now we have an expected

$$n - \binom{n}{k} 2^{1-\binom{k}{2}}$$

vertices. But this graph has all monochromatic k-cliques removed, and thus there exists a graph with at least this many vertices and no monochromatic k-clique. □

Fact 1.7

Using the same optimization with Stirling's formula on Proposition 1.6,

$$R(k,k) > \left(\frac{1}{e} + o(1)\right) k2^{k/2},$$

which is better than the result above by a factor of 2.

Both of these proofs are interesting, because although we now know a graph exists, we can't actually construct such an example easily!

1.3 Lovász Local Lemma

We're going to discuss some methods in this class beyond just picking things randomly: here's one of them. Let's say that we are trying to avoid a bunch of bad events $E_1, E_2, \cdots, E_n$ simultaneously. There's two main ways we know how to avoid them:

- All the probabilities are small, and there aren't too many of them. In particular, if the total sum of probabilities is at most 1, we always have a positive chance of success.
- If all the events are independent, then the probability of success is just the product of individual avoidances.

Theorem 1.8 (Lovász Local Lemma)

Let $E_1, \cdots, E_n$ be events each with probability at most p, where each event E_i is mutually independent of all other E_js except at most d of them. If $ep(d+1) \le 1$, then there is a positive probability that no E_i occurs.

Corollary 1.9 (Spencer, 1975)

We have $R(k,k) > n$ if

$$e\left(\binom{k}{2}\binom{n}{k-2} + 1\right) 2^{1-\binom{k}{2}} \le 1.$$

Proof. Randomly color all the edges, and again let A_R be the indicator variable for a subset R of k vertices forming a monochromatic clique. Note that all A_R and A_S are mutually independent unless they share an edge, meaning $|R \cap S| \ge 2$. For each given R, there are at most $\binom{k}{2}\binom{n}{k-2}$ choices for S, since we pick 2 vertices to share with R

and then pick the rest however we'd like. Now by Lovász Local Lemma, we have a positive probability no A_R occurs as long as

$$ep(d+1) = e\left(\binom{k}{2}\binom{n}{k-2}+1\right)2^{1-\binom{k}{2}} \leq 1.$$

□

Fact 1.10

This time, optimizing n in Corollary 1.9 yields

$$R(k,k) > \left(\frac{\sqrt{2}}{e}+o(1)\right)k2^{k/2}.$$

1.4 Set systems

Let $\mathcal{F}$ be a collection of subsets of $[n] = \{1, 2, \cdots, n\}$ (there are a total of 2^n subsets to put in $\mathcal{F}$). We call this an **antichain** if there is no set in $\mathcal{F}$ that is contained in another one.

Our question: what is the largest possible antichain? One thing we can do is to only use subsets of a fixed size k, since no set can be contained in another. This means we can at least get $\binom{n}{\lfloor n/2 \rfloor}$, the largest binomial coefficient. It turns out that this is the best bound:

Theorem 1.11 (Sperner, 1928)

If $\mathcal{F}$ is an antichain of subsets of $[n]$, then it has size at most $\binom{n}{\lfloor n/2 \rfloor}$.

To show this, we'll prove a more slightly general result:

Theorem 1.12

For any antichain $\mathcal{F}$ of the subsets of $[n]$,

$$\sum_{A\in\mathcal{F}}\binom{n}{|A|}^{-1} \leq 1.$$

This implies the result above, because it is a weighted sum where each weight $\binom{n}{|A|}$ is at most $\binom{n}{\lfloor n/2 \rfloor}$ (and the central binomial coefficients are largest).

Proof. Fix a random permutation σ of $[n]$. Associated with this permutation, we have a chain

$$\varnothing \subseteq \{\sigma(1)\} \subseteq \{\sigma(1), \sigma(2) \subseteq \cdots \subseteq \{\sigma(1), \cdots, \sigma(n)\} = [n].$$

Each subset A has probability $P_A = \binom{n}{|A|}^{-1}$ of appearing in such a chain, since each $|A|$-element subset has the same chance of appearing. However, no two subsets can appear in the same chain, so the events are disjoint. Thus, the sum of probabilities that A appears in the chain must be at most 1, and thus

$$\sum_{A\in\mathcal{F}} P_A = \sum_{A\in\mathcal{F}}\binom{n}{|A|}^{-1} \leq 1.$$

□

Theorem 1.13 (Bollobás' Two Families Theorem)
Given r-element sets $A_1, \cdots, A_m$ and s-element sets $B_1, \cdots, B_m$, if we know that

$$A_i \cap B_j = \varnothing \quad \text{if and only if} \quad i = j$$

(all A_i and B_j intersect except for $i = j$), then $m \le \binom{r+s}{r}$.

Where's the motivation for this coming from?

Definition 1.14
Given a family of sets $\mathcal{F}$, let a **transversal** T be a set that intersects all $S \in \mathcal{F}$, and let the **transversal number** $\tau(\mathcal{F})$ denote the size of the smallest transversal of $\mathcal{F}$. $\mathcal{F}$ is **τ-critical** if we have $\tau(\mathcal{F} \setminus S) < \tau(\mathcal{F})$ for all $S \in \mathcal{F}$.

Corollary 1.15 (of Theorem 1.13)
An r-uniform τ-critical family of sets $\mathcal{F}$ with $\tau(F) = s + 1$ has size at most $\binom{r+s}{r}$.

Proof. Let our family of sets be $A_1, \cdots, A_m$. $\mathcal{F}$ being τ-critical implies that for any i, we can find a transversal of size s for $\mathcal{F} \setminus A_i$. Letting this be B_i, notice that $A_i \cap B_j = \varnothing \iff i = j$, and thus by Bollobás' Theorem we can find the upper bound stated. □

Here's a slightly more general version of Bollobás' Theorem, which we'll prove now:

Theorem 1.16
Let $A_1, \cdots, A_m, B_1, \cdots, B_m$ be finite sets, such that $A_i \cap B_j = \varnothing$ if and only if $i = j$. Then

$$\sum_{i=1}^{m} \binom{|A_i| + |B_i|}{|A_i|}^{-1} \le 1.$$

Notice that if we make $B_i = [n] \setminus A_i$ for all i, we get Sperner's theorem. Meanwhile, if all A_is have size r and all B_js have size s, we get Bollobás' Two Families Theorem.

Proof. Like in Sperner's theorem, randomly order all elements in the union of all A_i and B_js. For any i, the probability that all of A_i occurs before all of B_i is $\binom{|A_i|+|B_i|}{|A_i|}^{-1}$, and we can't have this happen with two different is in any given ordering, because this would mean that either A_i and B_j are disjoint or A_j and B_i are disjoint. Thus all events of this form are disjoint, and we must have $\sum_{i=1}^{m} \binom{|A_i|+|B_i|}{|A_i|}^{-1} \le 1$, as desired. □

Definition 1.17
A family of sets $\mathcal{F}$ is **intersecting** if $A \cap B \neq \varnothing$ for all $A, B \in \mathcal{F}$.

Note that this does not mean they must all have a common element: for example, $\{\{1,2\},\{1,3\},\{2,3\}\}$ is intersecting.

Theorem 1.18 (Erdős-Ko-Rado 1961)
If $n \geq 2k$, then all intersecting families of k-element subsets of $[n] = \{1, 2, \cdots, n\}$ have size at most $\binom{n-1}{k-1}$.

(This can be constructed by having all sets share the element 1, for example.)

Proof. Order the integers $1, 2, \cdots, n$ around a circle randomly. Let a subset $A \subseteq [n]$ be **contiguous** if all elements lie in a contiguous block around the circle. For any subset A with $|A| = k$; the probability it is contiguous is

$$\binom{n}{\binom{n}{k}},$$

(think of picking k of the spots around the circle). So the expected number of contiguous subsets is $|\mathcal{F}|\binom{n}{\binom{n}{k}}$, but if all subsets are intersecting, we can only have k contiguous subsets (here, as long as $n \geq 2k$, all contiguous subsets must pass through a common point, which is why we set up the problem this way). Thus, $|\mathcal{F}|\binom{n}{\binom{n}{k}} \leq k$, and rearranging yields

$$|F| \leq \frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1},$$

as desired. □

1.5 Hypergraph colorings

This is a topic we'll be discussing quite a bit in this class, but the idea is very similar to that of set systems.

Definition 1.19
A ***k*-uniform hypergraph** $H(V, E)$ has a (finite) set of vertices V and a set of edges E, each of which is a k-element subset of V. H is ***r*-colorable** if we can color V with r colors such that no edge is monochromatic (that is, not all the vertices in an edge have the same color).

(Regular graphs are 2-uniform hypergraphs.) Let $m(k)$ to be the minimum number of edges in a k-uniform hypergraph that isn't 2-colorable.

Example 1.20
A triangle is not 2-colorable, so $m(2) = 3$. The Fano plane is not 2-colorable if we interpret lines as edges, so $m(3) = 7$ (any smaller example can be checked).

These quickly become hard to calculate, though: $m(4) = 23$, but $m(5)$ is actually currently unknown.

Theorem 1.21
A k-uniform hypergraph with fewer than 2^{k-1} edges is 2-colorable.

Proof. Color each vertex randomly; each edge has probability 2^{1-k} of being monochromatic, since all k vertices need to be one color or the other. Thus, if we have less than 2^{k-1} edges, the expected number of monochromatic edges is less than 1, so there is a way to 2-color the hypergraph successfully. □

To date, we have the bounds (which are reasonably close to each other)

$$m(k) \geq 2^k \sqrt{\frac{k}{\log k}} \quad \text{and} \quad m(k) = O(k^2 2^k).$$

How do we show the upper bound? We can restate it as follows:

Problem 1.22

Construct a k-uniform hypergraph with $O(k^2 2^k)$ edges that is not k-colorable.

Solution. Start with a set of vertices V where $|V| = n$, and let H be the hypergraph constructed by choosing m edges $S_1, S_2, \cdots, S_m$ at random. For any coloring of the vertices $\chi : V \to$ red, blue, the event $A(\chi)$ refers to H containing no monochromatic edges. Then our goal is to pick m, n so that

$$\sum_{\chi} \Pr(A_i) < 1,$$

because this means there is a graph H that cannot be properly colored regardless of which χ we pick.

A coloring χ that colors a vertices red and b vertices blue has a given S_i monochromatic with probability

$$\frac{\binom{a}{k} + \binom{b}{k}}{\binom{n}{k}} \geq \frac{2\binom{n/2}{k}}{\binom{n}{k}}$$

(since there are $\binom{n}{k}$ total sets of vertices and $\binom{a}{k} + \binom{b}{k}$ of them are monochromatic). Further bounding, this is

$$\geq 2\left(\frac{n/2 - k + 1}{n - k + 1}\right)^k = 2^{-k+1}\left(1 - \frac{k-1}{n-k+1}\right)^k \geq c2^{-k}$$

where we pick $n = k^2$ so that we can have

$$2\left(1 - \frac{k-1}{n-k+1}\right)^k \geq c,$$

a constant. So now the probability that we have a proper coloring (which means no S_i is monochromatic) is at most (looking at all S_is now)

$$(1 - c2^{-k})^m \leq e^{-c2^{-k}m},$$

since we chose our S_is randomly (possibly with replacement), and then $1 + x \leq e^x$ for all x. Therefore, if we sum over all χ, we have

$$\sum_{\chi} e^{-c2^{-k}m} = 2^n e^{-c2^{-k}m} < 1$$

for some value of $m = O(k^2 2^k)$, as desired. □

Now that we have a sampling of some preliminary techniques, we'll begin examining them in more detail in the next few chapters!

2 Linearity of expectation

2.1 Setup and basic examples

Often, a random variable X can be written as

$$X = c_1X_1 + c_2X_2 + \cdots + c_nX_n,$$

where c_i are constants and X_i are random variables, not necessarily independent. In these cases, we know that

$$\mathbb{E}[X] = c_1\mathbb{E}[X_1] + \cdots + c_n\mathbb{E}[X_n].$$

However, it is not necessarily true that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Example 2.1

Given a random permutation of $[n]$, how many fixed points do we expect it to have?

Solution. Let A_i be the indicator variable for i being a fixed point: $\sigma(i) = i$. Since i is a fixed point with probability $\frac{1}{n}$, the expected value of A_i is $\frac{1}{n}$, so the expected number of overall fixed points is just $n \cdot \frac{1}{n} = 1$. □

Let's take a look at a basic graph theory problem:

Definition 2.2

A **tournament** is a complete graph with each edge directed (from one endpoint to the other). A **Hamiltonian path** is a directed path through a graph which passes through all vertices.

Theorem 2.3 (Szele, 1943)

For all n, there exists a tournament on n vertices with at least $n!2^{-n+1}$ Hamiltonian paths.

Proof. Start with K_n and randomly orient each edge. Then for each permutation of the edges, the probability that the edges are all directed correctly to form a Hamiltonian cycle in that order is 2^{-n+1} (since there are only two orientations). Thus, by linearity of expectation, the expected number of Hamiltonian paths is $n!2^{-n+1}$, and thus there exists a tournament with at least that many Hamiltonian paths. □

Alon proved in 1990 that the maximum number is asympotically of that magnitude: we can have at most $\frac{n!}{(2-o(1))^n}$ Hamiltonian paths.

Let's now start to look at some more complicated applications.

2.2 Sum-free sets

Definition 2.4

A subset A of an abelian group is **sum-free** if there are no elements $a, b, c \in A$ with $a + b = c$.

An interesting abelian group to consider is the integers:

Theorem 2.5

Every set of n nonzero integers contains a sum-free subset of size at least $\frac{n}{3}$.

Proof. Let A be a set of nonzero integers with $|A| = n$. Pick a real nmber $\theta \in [0, 1]$, and let

$$A_\theta = \left\{ a \in A \mid \{a\theta\} \in \left(\frac{1}{3}, \frac{2}{3}\right) \right\}$$

(in other words, A_θ contains all points with fractional part of $a\theta$ in the middle third). Note that A_θ is always sum-free, since no two elements with fractional part in the middle third can add to a third. Now uniformly pick θ from 0 to 1: since the probability any a is in A_θ is always $\frac{1}{3}$ (since $a\theta$ ranges from 0 to a), the expected number of points in A_θ is $\frac{n}{3}$, and therefore there is some sum-free subset A_θ with size at least $\frac{n}{3}$, as desired. □

The best we can do currently is $\frac{n+2}{3}$, and it's been shown that $\left(\frac{1}{3} + c\right) n$ is not possible asymptotically for any $c > 0$. However, the constant c' in $\frac{1}{3}n + c'$ is still open!

2.3 Cliques

Theorem 2.6 (Ramsey multiplicity)

There exists a 2-coloring of the edges of K_n with a "relatively small number" of t-cliques: there are at most $2^{1-\binom{t}{2}}\binom{n}{t}$ monochromatic copies of K_t.

Proof. Color all the edges randomly. The expected number of monochromatic K_ts is, by linearity of expectation,

$$\binom{n}{t} 2^{1-\binom{t}{2}}$$

since each t vertices we pick has $\binom{t}{2}$ edges and there are only 2 ways to color them to form a monochromatic K_t. Thus, there is a positive probability that the number of monochromatic K_t is at most this number. □

Definition 2.7

Let c_t be the maximum constant such that every 2-edge coloring of K_n has at least $(c_t + o(1))\binom{n}{t}$ monochromatic t-cliques.

In other words, c_t is the best fractional bound on the number of t-cliques, and we've just found that $c_t \leq 2^{1-\binom{t}{2}}$. Can we do better and find a smaller c_t?

It is known that this is tight for $t = 3$: Goodman's theorem implies that we indeed have $c_3 = \frac{1}{4}$. (Proving this is a good exercise in double counting.) We'd initially suspect that equality can also be achieved for $t = 4$, but it was found by Thomason in 1989 that $c_4 < \frac{1}{33} < \frac{1}{2^5}$. Likewise, the bound has been shown to be not tight for all $t > 4$. In fact, the exact value of c_4 is still an open problem.

But can we prove any kind of lower bound for c_t? Specifically, what techniques do we have to proving positive lower bounds? In other words, we're trying to show that there's a lot of t-cliques, and that sounds vaguely like Ramsey's theorem. One thing we could do is find a copy, delete a vertex, and repeat, but this gives a linear number of t-cliques for n^2 edges, which isn't enough for a positive constant. Instead, we'll use the **sampling trick**!

Theorem 2.8

Every 2-coloring of K_n with $n \geq R(t,t)$ contains $\geq \binom{R(t,t)}{t}^{-1} \cdot \binom{n}{t}$ monochromatic K_ts.

Proof. Suppose there are M monochromatic K_ts in our coloring. Let X be any t-clique: then it has a probability of $\frac{M}{\binom{n}{t}}$ of being monochromatic.

But instead, let's pick the same X in a different way. First, pick a random $R(t,t)$ clique, where $R(t,t)$ is the Ramsey number, and then pick a t-vertex subclique of that. (For this trick to work, we need to be able to pick a random $R(t,t)$ clique.) This second procedure has two random steps, but by Ramsey's theorem, there is at least one monochromatic t-clique in this second step! So X is monochromatic with probability at least $\binom{R(t,t)}{t}^{-1}$.

So putting these together,

$$\frac{M}{\binom{n}{t}} \geq \binom{R(t,t)}{t}^{-1}.$$

□

This is likely far from optimal, but at least it gives us a nonzero lower bound on c_t:

Corollary 2.9

For all positive integers t,

$$c_t \geq \binom{R(t,t)}{t}^{-1}.$$

2.4 Independent sets

Let's turn to a new question: what is the maximum number of edges in an n-vertex K_t-free graph? Note that cliques in a graph G are the same as independent sets in $\overline{G}$ (the graph's complement), so this is a very similar idea to what we've been already been discussing.

Theorem 2.10 (Caro-Wei)

Every graph G contains an independent set I of size

$$|I| \geq \sum_{v \in G} \frac{1}{1+d(v)}.$$

In particular, we should expect large independent sets out of graphs with low degrees, which is convenient for us.

Proof by Alon and Spencer. Consider a random ordering of V, and let I be the set of vertices that appear before all of its neighbors in the graph.

I is an independent set, since no edge can connect two vertices in I (one comes before another). How big is I? By linearity of expectation,

$$\mathbb{E}[|I|] = \sum_{v \in V} \mathbb{P}(v \in I).$$

The probability that a vertex v is in I is $\frac{1}{1+d(v)}$, since there are $d(v)+1$ total vertices to consider here, v and all of its neighbors, and v must be the one in front. So there's a nonzero probability that an independent set of size at least $\sum_v \frac{1}{1+d(v)}$ exists. □

Now, let's take the complement of Caro-Wei. Independent sets become cliques and vice versa, which yields the following:

Corollary 2.11

Every graph G contains a clique of size

$$S \geq \sum_{v \in G} \frac{1}{(n-1-d(v))+1} = \sum_{v \in G} \frac{1}{n-d(v)}.$$

Note that if we hold the number of degrees fixed, so $\sum d(v) = 2|E|$, the sum is minimized when the $d(v)$s are as close as possible.

So where's the equality case of Caro-Wei (and the corollary after it)? To have maximal independent set size and largest multiplicity, we want something like the following:

Definition 2.12

A **Turán graph** $T_{n,r}$ has n vertices and is an r-partite complete graph, such that each part has either $\left\lfloor \frac{n}{r} \right\rfloor$ or $\left\lfloor \frac{n}{r} \right\rfloor + 1$ vertices.

Note that this graph is K_{r+1}-free, and it turns out this is the extreme example:

Theorem 2.13 (Turán's theorem)

Given a graph G with n vertices that is K_{r+1} free,

$$|E(G)| \leq |E(T_{n,r})| \leq \left(1 - \frac{1}{r}\right)\frac{n^2}{2},$$

where the inequalities are tight if $r|n$.

For simplicity, we'll show a slightly weaker result where we skip the middle part of the inequality.

Proof. Since G is K_{r+1} free, by the complement of Caro-Wei,

$$r \geq \sum_{v \in V} \frac{1}{n - d(v)} \geq \frac{n}{n - \overline{d}}$$

by convexity, where $\overline{d}$ is the average degree of the vertices. Since the average degree is $\frac{2|E|}{n}$, rearranging gives the result. □

We just have to be a bit more careful in the case where r doesn't divide n, but it's not too much more difficult.

2.5 Crossing numbers

The next example may seem a bit less familiar in terms of the techniques it uses. Given a graph G, we can draw it on the plane; it may or may not be planar. A graph is **planar** if we can draw it in a way such that all edges are continuous curves and only intersect at vertices.

Fact 2.14 ("Common folklore knowledge" and Kuratowski's theorem)

K_4 is planar, but K_5 and $K_{3,3}$ are not. It turns out these are the only two minimal examples of nonplanar graphs: any nonplanar graph contains a subgraph that is topologically equivalent to K_5 or $K_{3,3}$.

The idea is that if we see a graph with a lot of edges, it should have a lot of crossings. How many such crossing must K_n or $K_{n,n}$ have? In fact, what's the bound for any G with some large number of edges?

The exact answers to K_n and $K_{n,n}$ are famous open questions, but there are conjectures: they're called Hill's conjecture and the Zarankiewicz conjecture, respectively.

Remark (Historical note)**.** *The problem of drawing the complete bipartite graph with the minimum number of crossings is also called Turán's brick factory problem. During World War II, Turán was forced to work in a brick factory pushing wagons of bricks along rail tracks. The wagons are harder to push when the rail tracks cross. This experience inspired Turán to think about how to design the layout of the tracks in order to minimize the number of crossings.*

The conjecture for $K_{n,n}$ is to either place points antipodal on a sphere and connect geodesics, or put one set on the x-axis and the other on the y-axis. That makes this problem hard: two very different constructions do equally well.

Definition 2.15

The **crossing number** $\text{cr}(G)$ is the minimum number of crossings in a planar drawing of G.

Are there any bounds we can place on this? It seems like we should expect $O(n^4)$ crossings, since any 4 points potentially create a crossing. Is that at least correct up to a constant factor?

We'll start by considering some facts in graph theory:

Proposition 2.16 (Euler's formula)

Given a connected planar graph with V vertices, E edges, and F faces,

$$V - E + F = 2.$$

The next few sentences are easy to get wrong, so we're going to be careful.

Proposition 2.17

Every connected planar graph with at least one cycle (not just a tree) has $3|F| \le 2|E|$.

This is true because every face is surrounded by at least 3 edges, and every edge touches exactly 2 faces.

Plugging this into Euler's formula, we also find that $|E| \le 3|V| - 6$ for all connected planar graphs with at least one cycle. There are some graphs that do not satisfy the conditions above, but that's okay - from similar arguments, we can still deduce that all planar graphs satisfy $|E| \le 3|V|$.

So if there are too many edges, we want to be able to say that there are lots of crossings. Basically, every edge beyond the threshold of $3|V|$ could add a crossing, so if we delete one edge per crossing, we get a planar graph. Thus $|E| - \text{cr}(G) \le 3|V|$, or

$$\text{cr}(G) \ge |E| - 3|V|.$$

But this gives $O(n^2)$ crossings for an n-vertex graph, and we're trying to show that $O(n^4)$ crossings exist. Here's where the probabilistic method comes in: we're going to sample like we did with the Ramsey number to get a better answer.

Theorem 2.18 (Crossing number inequality)

Given a graph G with $|E| \ge 4|V|$,

$$\text{cr}(G) \gtrsim |E|^3/|V|^2.$$

Proof. Let $p \in [0,1]$ be a number that we will decide later, and let G' be obtained from G by randomly picking each vertex with probability p. In other words, randomly delete each vertex (and the edges connected to it) with probability $1-p$.

Our graph G' should satisfy

$$\mathrm{cr}(G') \geq |E'| - 3|V'|,$$

and now take expectations of both sides:

$$\mathbb{E}[\mathrm{cr}(G')] \geq \mathbb{E}[|E'|] - 3\mathbb{E}[|V'|]$$

If we start with a drawing of G, each crossing has 4 vertices that contribute to it. This crossing remains with probability p^4, but note that after we delete some vertices and edges, we can potentially redraw the diagram to have less crossings. So the left hand side has an inequality of the form

$$\mathbb{E}[\mathrm{cr}(G')] \leq p^4 \mathrm{cr}(G).$$

The right hand side is easier:

$$\mathbb{E}[|E'|] = p^2|E|, \mathbb{E}[|V'|] = p|V|.$$

Moving the p^4 to the other side now, we have a new bound:

$$\mathrm{cr}(G) \geq p^{-2}|E| - 3p^{-3}|V|$$

From here, we set p so that we have $4p^{-3}|V| \leq p^{-2}|E|$, but note that this only works if $|E| \geq 4|V|$, since our probability needs to be between 0 and 1. This gives the result that we want: □

Notably, if $|V| = n$ and $|E| \gtrsim n^2$ (is quadratic in n), then $\mathrm{cr}(G) \gtrsim n^4$: the crossing number is quartic in n, as desired!

2.6 Application to incidence geometry

Problem 2.19

Given n points and n lines, what's the maximum number of incidences between them?

Let's formulate this more rigorously:

Definition 2.20

Let $\mathcal{P}$ be a set of points and $\mathcal{L}$ be a set of lines. Define

$$I(\mathcal{P}, \mathcal{L}) = \{(p, \ell) \in \mathcal{P} \times \mathcal{L} : p \in \ell\}$$

to be the set of intersections between a point in $\mathcal{P}$ and a line in $\mathcal{L}$.

We wish to maximize $|I(\mathcal{P}, \mathcal{L})|$.

Example 2.21

Let $\mathcal{P}$ be the lattice grid $[k] \times [2k^2]$, and let $\mathcal{L}$ be the lines with small integer slope: $\mathcal{L} = \{y = mx + b, m \in [k], b \in [k^2]$. Then every line in $\mathcal{L}$ contains k points, so

$$|I(\mathcal{P}, \mathcal{L})| = k^4,$$

which gives $O(n^{4/3})$ incidences.

The natural question to ask is whether this is optimal, and the answer is yes. To prove this, let's start trying to find some upper bounds. Assume temporarily that every line has at least two incidences: clearly, there is a bound

$$I(\mathcal{P}, \mathcal{L}) \le |\mathcal{P}||\mathcal{L}|,$$

which is weak if there are at least 2 points or 2 lines. But let's use the fact that there is at most one line through each pair of points: to do this, we'll double count the number of triples $(p, p', \ell) \in \mathcal{P} \times \mathcal{P} \times \mathcal{L}$ with $p \ne p'$ and $p, p' \in \ell$. On one hand, given two points, we've determined the line, so there are at most $|\mathcal{P}|^2$ such triples. On the other hand, if we count the incidences in terms of lines, the number of triples is

$$\sum_{\ell \in \mathcal{L}} |P \cap \ell|(|P \cap \ell| - 1) \ge \frac{I(\mathcal{P}, \mathcal{L})^2}{|\mathcal{L}|} - I(P, \mathcal{L})$$

where we've done bounding by Cauchy-Schwarz. Putting these together,

$$I(\mathcal{P}, \mathcal{L}) \lesssim |\mathcal{P}||\mathcal{L}|^{1/2} + |\mathcal{L}|.$$

By point-line duality, we can also find an analogous statement if we flip L and P. Either way, for n lines and n points, we're getting $O(n^{3/2})$, which is not as strong as $O(n^{4/3})$.

Remark. *We can make this bound that we found tight in some situations, though: it turns out this is the right number of incidences over a finite field $\mathbb{F}_q^2$ if we take all $\Theta(q^2)$ lines and all q^2 points.*

Back to the Euclidean plane. To make the bound tight, we invoke the topology of Euclidean space and the crossing number theorem. Assume, again, that every line has at least 2 incidences. Draw a graph based on the point-line configuration, where the points are vertices and **consecutive** points on a line form an edge. So each line gets chopped up into some number of segments.

How many edges and vertices are there? The points are vertices, so $|V| = |\mathcal{P}|$. A line with k incidences (and $k \ge 2$) has $k - 1 \ge \frac{k}{2}$ edges, so the number of edges is at least

$$|E| \ge \frac{I(\mathcal{P}, \mathcal{L})}{2}.$$

Two lines can cross at most once, so

$$\text{cr}(G) \le |\mathcal{L}|^2.$$

Provided that the number of incidences is at least 8 times the number of points, we can invoke the crossing number inequality:

$$|\mathcal{L}^2| \ge \text{cr}(G) \gtrsim \frac{|E|^3}{|V|^2} \gtrsim \frac{|I(\mathcal{P}, \mathcal{L})|^3}{|\mathcal{P}|^2}.$$

Rearranging, this gives us

$$I(\mathcal{P}, \mathcal{L}) \lesssim |\mathcal{P}|^{2/3}|\mathcal{L}|^{2/3},$$

but this only works if we have a sufficiently large number of incidences, so we need to add a linear $|\mathcal{P}|$ term. We also need to correct for the fact that we're assuming that there are at least 2 incidences per line, which adds a linear $|\mathcal{L}|$ term:

Theorem 2.22 (Szemerédi-Trotter theorem)
For any set of points and lines,

$$I(\mathcal{P}, \mathcal{L}) \lesssim |\mathcal{P}|^{2/3}|\mathcal{L}|^{2/3} + |\mathcal{P}| + |\mathcal{L}|.$$

This is sharp up to constant factors! As a corollary, n points and n lines always have $O(n^{4/3})$ incidences.

2.7 Derandomization: balancing vectors

We'll start by solving a problem with familiar techniques:

Theorem 2.23
Given $v_1, \cdots, v_n \in \mathbb{R}^n$ unit vectors, there exists $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \in \{-1, 1\}$ such that

$$|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n| \leq \sqrt{n}.$$

This is motivated by considering $v_1, \cdots, v_n$ to be a standard basis: our choices can't get the length of the vector any smaller than $\sqrt{n}$. As a sidenote, we can also show that we can pick the ε_is to make the length at least $\sqrt{n}$.

We want to use linearity of expectation, but we have a small problem: we have an expectation of an absolute value. The easiest way to get around this is to square both sides of our equation!

Proof. Let

$$X = |\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n|^2,$$

and pick each ε_i independently and randomly between $\{-1, 1\}$. X expands out to the sum

$$X = \sum_{i,j=1}^{n} \varepsilon_i \varepsilon_j (v_i \cdot v_j)$$

and now that the absolute values are gone, we can just use linearity of expectation: for $i \neq j$, the expectation is 0, and for $i = j$, we get a contribution of $1 \cdot |v_i|^2 = 1$ from each term. So the expected value of X is n, so with some positive probability $X \leq n$ (and also $X \geq n$). □

We can also do this all deterministically: in this case, we don't actually have to use the probabilistic method.

Finding the ε_is algorithmically. We're going to pick our ε_is sequentially and greedily. At each step, we pick the ε_i that minimizes the expected value conditional on the previous choices.

For example, if we pick $\varepsilon_1, \cdots, \varepsilon_{k-1}$, let $w = \varepsilon_1 v_1 + \cdots + \varepsilon_{k-1} v_{k-1}$. Then our conditional probability

$$\mathbb{E}[X \mid \varepsilon_1, \cdots, \varepsilon_k] = \mathbb{E}\left[|w + \varepsilon_k v_k + \varepsilon_n v_n|^2 \mid \varepsilon_1, \cdots, \varepsilon_k\right],$$

and expanding out the square again, this becomes the expected value of

$$|w|^2 + 2\varepsilon_k (w \cdot v_k) + (n - k - 1).$$

To minimize this value, we pick $\varepsilon_k = 1$ if and only if $w \cdot v_k \leq 0$. □

Why couldn't we do something like this for the Ramsey number proof, too? The idea is that we can't compute the number of cliques of other subsets easily! (It is "expensive" to do so.) This idea of turning probabilistic proofs into deterministic ones is called **derandomization**.

2.8 Unbalancing lights

Problem 2.24

Consider a grid of $n \times n$ lights, where we only have light switches for each row and column. How can we maximize the number of lightbulbs turned on given some starting configuration?

Represent this as an array of ± 1 numbers. Let $a_{ij} \in \{-1, 1\}$ for all $1 \le i, j \le n$, and let's say that our light switches are labeled $x_1, \cdots, x_n, y_1, \cdots, y_n \in \{-1, 1\}$. Our goal is then to maximize the quantity

$$\sum_{i,j=1}^{n} a_{ij} x_i y_j,$$

since only the parity of how many times we flip each switch matters (not even the order).

Well, there are n^2 variables, so if we do our probabilistic method naively at random, we can guarantee a linear answer in n, since $\sqrt{n^2} = n$. But we can do better than that:

Theorem 2.25

Given fixed $a_{ij} \in \{-1, 1\}$, we can pick $x_1, \cdots, x_n, y_1, \cdots, y_n \in \{-1, 1\}$, such that

$$\sum_{i,j=1}^{n} a_{ij} x_i y_j \ge \left(\sqrt{\frac{2}{\pi}} + o(1) \right) n^{3/2}.$$

Proof. Choose $y_1, \cdots, y_n \in \{-1, 1\}$ randomly: this means that we pick a random way to flip our columns. Now, for each row, we can choose x_i such that the ith row sum is nonnegative (in other words, flip a row if the sum is negative). Each row sum is

$$R_i = \sum_{j=1}^{n} a_{ij} y_j,$$

and our final sum is just $R = \sum_{i=1}^{n} |R_i|$. Here we use linearity of expectation: the expected value of each R_i is the same, and each R_i is a sum of ± 1s. This gives a binomial distribution: we can use the Central Limit Theorem, since our quantity

$$\mathbb{E}\left(\frac{|R_1|}{\sqrt{n}} \right) \to \mathbb{E}|X| = \sqrt{\frac{2}{\pi}}.$$

(Alternatively, we can directly compute

$$\mathbb{E}[|R_1|] = n 2^{1-n} \binom{n-1}{\lfloor \frac{n-1}{2} \rfloor}$$

and use Stirling's formula.) Regardless, each row has expected value $\left(\sqrt{\frac{2}{\pi}} + o(1) \right) \sqrt{n}$, which is what we want. □

2.9 2-colorings of a hypergraph

Theorem 2.26

Let a k-uniform hypergraph have a vertex set V partitioned as

$$V = V_1 \cup \cdots \cup V_k,$$

where $|V_i| = n$ for all i. Suppose the edges of the complete k-uniform hypergraph on V are colored red and blue such that every edge that intersects all of $V_1, \cdots, V_k$ is colored blue. Then there exists a subset of the vertices $S \subset V$ such that

$$|\# \text{ blue edges} - \# \text{ red edges}| \geq c_k n^k$$

for some constant k.

For example, if $k = 2$, we're looking at a 2-coloring of a complete graph where all of the cross-edges between two halves are blue: our goal is to get a large difference in the number of red and blue edges. Similarly, if $k = 3$, we partition $3n$ vertices into three parts and draw triangles. All the triangles that intersect all three parts are blue, but everything else can be red or blue.

Proof. The idea here is to choose S by including each vertex in a given V_i with probability p_i. We'll leave $p_1, p_2, \cdots, p_k$ undetermined for now.

Let's do the proof for $k = 3$ for illustration, but this generalizes to any k. Let a_{xyz} be the difference in the number of blue and red edges in $V_x \times V_y \times V_z$. When we randomly pick our vertices, by linearity of expectation, the expected number of blue minus red edges is

$$n^3 p_1 p_2 p_3 + \sum_{\substack{x \leq y \leq z \\ \text{not all different}}} a_{xyz} p_x p_y p_z.$$

The first term here comes from the forced blue triangles between all V_is. Our goal is to show this absolute value of this expression is (at least) cubic, and then we'll be done by linearity of expectation.

We haven't chosen our p_is yet, and for each specific choice, we might end up with expected values that are pretty close to 0. So there is always a graph that beats a specific set of p_i, but we just want to find p_1, p_2, p_3 that work given a graph. This is now just an analysis problem:

Lemma 2.27

Let P_k denote the set of polynomials of the form $g(p_1, \cdots, p_k)$ with degree at least k and coefficients having absolute value at most 1, where the coefficient of $p_1 p_2 \cdots p_k$ is exactly 1. Then there exists a constant c_k such that for all polynomials in P_k, there exists $p_1, \cdots, p_k \in [0, 1]^k$ such that

$$g(p_1, p_2, \cdots, p_k) \geq c_k.$$

The proof of this is short: let $M(g)$ be the supremum

$$\sup_{p_1, \cdots, p_k \in [0,1]^k} |g(p_1, \cdots, p_k)|$$

By continuity and compactness, this is actually an achieved maximum, and it is always positive, since all polynomials are nonzero. Furthermore, this map $M : P_k \to \mathbb{R}$ is continuous on a compact domain, so it must achieve its minimum, which is nonzero.

This doesn't give a concrete value of c_k, but it tells us that one exists! And now we're done with the linearity of expectation argument, since all $a_{ijk} < n^3$. □

The main take-away here is that we decide probabilities for our random process in the last step, since no probabilities will work for every configuration.

2.10 High-dimensional sphere packings

Problem 2.28

What is the densest possible packing of unit balls in $\mathbb{R}^n$?

This has been solved for $n = 1$ (trivial), $n = 2$ (a rigorous proof wasn't found until the middle of the 20th century), and $n = 3$ (Kepler's conjecture; proved with computer assistance in the 1990s, and a formal computer proof was recently completed).

Recently, there was a breakthrough that found the answer for $n = 8$ and $n = 24$ as well; those answers come from the E_8 and Leech lattices respectively. However, the problem is open in all other dimensions.

The definition of "density" can be thought of pretty intuitively:

Definition 2.29

Let Δ_n be the maximum fraction of space occupied by non-overlapping unit balls in a large box in $\mathbb{R}^n$ as the volume of the box goes to infinity.

We wish to understand bounds on Δ_n. What are examples of good sphere-packings with high density?

Example 2.30

Consider a packing where we pack greedily: we keep throwing balls in wherever there is space. Alternatively, take any **maximal** packing: basically, find one where we can't fit any additional balls in $\mathbb{R}^n$ anymore without overlap.

What can we say about the density of such a maximal sphere packing? Well, double the radii of every ball, and suppose there is a spot not covered. Then we could just put a unit ball centered at that spot which doesn't intersect any of our initial balls, contradicting maximality of our packing. Thus, we must be able to cover all of $\mathbb{R}^n$ with doubled radii, and thus

$$2^n \Delta_n \geq 1, \text{ so } \Delta_n \geq 2^{-n}.$$

For comparison, what's the packing for $\mathbb{Z}^n$? We can put a ball with radius $\frac{1}{2}$ at every lattice point, and the density is just the volume of a ball of radius $\frac{1}{2}$. This is a pretty standard formula: it's

$$V = \frac{2^{-n}\pi^{n/2}}{(n/2)!} < n^{-cn},$$

so the integer lattice does very poorly compared to the "random" lattice. Are there better ways to construct lattices in higher dimensions? Here's the best bound we know at the moment:

Theorem 2.31 (Kabatiansky–Levenshtein, 1978)

The sphere-packing density in $\mathbb{R}^n$ is at most $2^{-(0.599\cdots+o(1))n}$.

Where does the probabilistic method come into our picture? Although we can't prove the above fact, we want to at least get a better bound than 2^{-n}.

Definition 2.32

A **lattice** is the $\mathbb{Z}$-span of a basis in $\mathbb{R}^n$: given $v_1, v_2, \cdots, v_n$, we can write a matrix with basis vectors as columns. A lattice is **unimodular** if the covolume (volume of the fundamental domain) is 1, which means the matrix has determinant ± 1.

Let's consider matrices A such that $\det A = 1$, so $A \in SL_n(\mathbb{R})$. On the other hand, given a lattice, there's different ways to represent it with a basis: we could always pick $(v_1 + v_2, v_2, \cdots, v_n)$ instead of $(v_1, v_2, \cdots, v_n)$. Any such transformation is matrix multiplication of $B \in SL_n(\mathbb{Z})$.

So the whole point is that lattices are matrices in $SL_n(\mathbb{R})/SL_n(\mathbb{Z})$ through row reduction. Our question: is there a way to pick a random lattice here?

Fact 2.33

The space has a finite Haar measure, so there exists a (normalized) probability Haar measure on $SL_n(\mathbb{R})/SL_n(\mathbb{Z})$, which allows us to choose a random point in the space. That random point will be our random lattice.

Theorem 2.34 (Siegel mean value theorem)

If L is a random unimodular lattice in $\mathbb{R}^n$ (chosen as above according to the Haar probability measure), and if S is any measurable subset of $\mathbb{R}^n$, then

$$\mathbb{E}\left(|L \cap (S \setminus \{0\})|\right) = \text{vol}(S).$$

The idea is that the average point density is 1, so the number of nonzero lattice points is the volume. We exclude 0 because it's always in the lattice.

Proof sketch. Observe that the function $S \to \mathbb{E}\left(|L \cap (S \setminus \{0\})|\right)$ is additive, so it is a measure. Because of how we chose our lattice, it is $SL_n(\mathbb{R})$-invariant, so the measure is also $SL_n(\mathbb{R})$ invariant. Therefore, the only measures that work are constant multiples of the Lebesgue measure.

Now imagine we take a very large ball, much larger than the size of our lattice: then the expected value is the volume minus some boundary errors. So $|S \cap L| \sim \text{vol } S$ and the normalizing constant must be 1. □

How do we use this to find dense lattices?

Proposition 2.35

There exist lattices with sphere packing density greater than 2^{-n}.

Proof. Let S be a ball of volume 1 centered at the origin, and pick a random lattice. By the Siegel mean value theorem, the expected number of nonzero lattice points of L that are in S is 1 (think of this as $1 - \varepsilon$). We can show, then, that there must exist L such that L has no nonzero lattice points in S, since there is a positive probability that there is more than 1 lattice point.

So now put $\frac{1}{2}S$ around every point of L; this gives us a packing with density 2^{-n}. But notice that the nonzero lattice points come in pairs $\{x, -x\}$! In other words, we can take S to be a ball of volume 2. Then we can guarantee the expected number of nonzero lattice points is 2, and we can't have exactly 1 lattice point, so we have the same conclusion as before. This yields a sphere packing with density 2^{1-n}, and this improvement is due to Minkowski. □

Can we do better? There's a lot of connections to the geometry of numbers here. There was a long sequence of improvements made, all of the form $\Delta n \geq cn2^{-n}$, over a few decades. c went from $\frac{1}{2}$ to about 2, but then Venkatesh realized that we can gain factors of k if we have additional symmetry in our lattices: number theory gives such lattices with k-fold symmetry!

For example, consider the lattice corresponding to a cyclotomic field: that is, look at the lattice spanned by a kth root of unity ω. This has a k-fold action, which is multiplication by ω. The end result is that a "random lattice" can be extended to a random unimodular lattice in dimensions $n = 2\phi(k)$, with k-fold symmetry, also satisfying the Siegel mean value theorem conditions. So now k-fold symmetry gives density

$$\Delta_n \geq k \cdot 2^{-n},$$

and this turns out to maximize the gain when $k = p_1 p_2 \cdots p_n$, where p_i is the ith prime. Number theoretic calculations give the following result:

Theorem 2.36 (Venkatesh, 2012)

There exists a lattice packing of unit balls of density

$$\Delta_n \geq cn \log\log n \cdot 2^{-n}$$

for infinitely many values of n and some $c > 0$.

These values of n are very sparse, but this is the state-of-the-art bound. Venkatesh also used a different method to show that (for all sufficiently large n)

$$\Delta_n \geq 60000n \cdot 2^{-n}.$$

It's an open problem whether or not we can get sphere packings of exponentially better density than this, though!

3 Alterations

Recall the naive probabilistic method: we found some lower bounds for Ramsey numbers in Section 1.1, primarily for the diagonal numbers. We did this with a basic method: color randomly, so that we color each edge red with probability p and blue with probability $1-p$. Then the probability that we don't see any red s-cliques or blue t-cliques (with a union bound) is at most

$$\binom{n}{s}p^{\binom{s}{2}} + \binom{n}{t}(1-p)^{\binom{t}{2}},$$

and if this is less than 1 for some p, then there exists some graph on n vertices for which there is no red K_s and blue K_t. So we union bounded the bad events there.

Well, the alteration method does a little more than that - here's a proof that mirrors that of Proposition 1.6. We again color randomly, but the idea now is to delete a vertex in every bad clique (red K_s and blue K_t). How many edges have we deleted? We can estimate by using linearity of expectation:

Theorem 3.1

For all $p \in (0,1)$, $n \in \mathbb{N}$,

$$R(s,t) > n - \binom{n}{s}p^{\binom{s}{2}} - \binom{n}{t}(1-p)^{\binom{t}{2}}.$$

This right hand side begins by taking the starting number of vertices and then we deleting one vertex for each clique. We're going to explore this idea of "fixing the blemishes" a little more.

3.1 Dominating sets

Definition 3.2

Given a graph G, a **dominating set** U is a set of vertices such that every vertex not in U has a neighbor in U.

Basically, we want a subset of vertices such that every vertex is either picked or adjacent to something we picked. Clearly the whole set of vertices is dominating, but our goal is to find small dominating sets relative to the number of vertices.

Theorem 3.3

If our graph G has n vertices and minimum degree δ among all vertices ($\delta > 1$), then G has a dominating set of size at most $\frac{\log(\delta+1)+1}{\delta+1}n$.

Proof. We will do a two-step process. First, pick a random subset X by including every vertex with probability p. Then, add all vertices that are neither in X or the neighbors of X (since those are the ones we haven't covered with our set yet); call this Y. By this point, we have a dominating set $X \cup Y$ by construction.

Now, how many vertices do we have in our dominating set? Any vertex v is in Y if neither v nor any of its neighbors are in X. So v has probability $(1-p)^{\deg(v)+1} \le (1-p)^{1+\delta}$ of being included in Y, meaning that the expected size of $X \cup Y$ is

$$\mathbb{E}[X] + \mathbb{E}[Y] = np + n(1-p)^{1+\delta}.$$

Now we just optimize for p. The important computational trick is that we can bound this pretty well if p is small:

$$\leq np + ne^{-p(1+\delta)}.$$

Turns out the optimal value is $p = \frac{\log(\delta+1)}{\delta+1}$, and this gives the result we want. □

3.2 A problem from discrete geometry

Problem 3.4 (Heilbronn triangle problem)

Place n points in the unit square. How large can we make the smallest area of any triangle formed by our points?

This is related to the ideas of **discrepancy theory**. There are applications when we want to evenly distribute points, and this is one way of quantifying that randomness.

Definition 3.5

Let $\Delta(n)$ be the minimum real number such that for any n points in the unit square, there are three points with triangle area at most $\Delta(n)$.

For example, it's bad to have a square grid of points, since we get a minimal area of 0. If we put the n points on a circle, we get an area on the order of $\frac{1}{n^3}$, which is at least nonzero. The whole point is that we don't want collinearity, so it's hard to think about an efficient picture that is "irregular."

Heilbronn conjectured that $\Delta(n) \lesssim n^{-2}$, but this was disproved in 1982 by KPS: they showed $\Delta(n) \gtrsim \frac{\log n}{n^2}$. On the other hand, the best known upper bound is $\lesssim n^{-\frac{8}{7}+o(1)}$.

Below, we use a randomized construction to show that $\Delta(n) \gtrsim n^{-2}$:

Proposition 3.6

There exist n points in a unit square such that every three form a triangle with area at least cn^{-2} for some constant $c > 0$.

Proof. Choose $2n$ points at random (uniformly in the unit square). How can we find the probability that the area of a triangle pqr is at most ε?

Pick p first. The probability that the distance between p and q is in the range $[x, x + \Delta x]$ is the intersection of the square and the annulus with bounds x and $x + \Delta x$, which is always at most $\Theta(x\Delta x)$ (by taking Δx to be small).

So now, if we fix p and q, what's the probability that our area is less than ε; that is, the height from r to line pq is small? This means we want the distance between line pq and point r to be at most $\frac{2\varepsilon}{\text{dist}(p,q)}$, which is bounded by a constant times $\frac{\varepsilon}{x}$ (because the allowed region is bounded by a rectangle with height $\frac{4\varepsilon}{x}$ and length $\sqrt{2}$).

Putting these together, the probability that the area is at most ε can be bounded by a factor proportional to

$$\int_0^{\sqrt{2}} x \cdot \frac{\varepsilon}{x} dx \lesssim \varepsilon.$$

So now we apply the idea of the alteration method: let X be the number of triangles with area ε, and delete 1 point from each triangle: let's say we delete x triangles. What's the expected number of points that are removed? We remove $\mathbb{E}[X] \propto \varepsilon n^3$ points, and we'll pick $\varepsilon = \frac{c}{n^2}$ for some constant c such that the expected value of x is $\leq n$. Now with positive probability, our process deleted fewer than n points, so we have at least n points with no small triangles of area less than $\frac{c}{n^2}$, and we're done. □

Actually, we can also do a direct algebraic construction. Let's say we want to find n points in a square grid with no three points collinear. Note that a lattice polygon has area at least $\frac{1}{2}$, so take $n = p$ to be a prime number, and let our points be $\{(x, x^2) : x \in \mathbb{F}_{p^2}\}$ in $\mathbb{F}_p^2$. Parabolas have no three points collinear, and thus we've constructed configurations with smallest area proportional to n^{-2} explicitly.

So the idea is that although algebra solutions are pretty, it's often hard to modify algebraic constructions, while combinatorial proofs let us use heavier hammers.

3.3 Hard-to-color graphs

There are many problems in combinatorics for which probabilistic constructions are the only ones we know. Here's an example that Erdős studied.

Definition 3.7

The **chromatic number** $\chi(G)$ of a graph is the minimum number of colors needed to properly color G.

If we look at a very large graph and look at it locally, we can set some lower bounds on the chromatic number. For example, a K_4 means that $\chi(G) \geq 4$. Our question: is it possible to use local information to find that $\chi(G)$ is upper-bounded? Turns out the answer is no!

Definition 3.8

The **girth** of a graph G is the length of the shortest cycle in G.

Theorem 3.9 (Erdős)

For all positive integers k and ℓ, there exists a graph of girth more than ℓ and chromatic number more than k.

The idea is that for graphs with large girth, we only see trees locally, and that won't tell us anything. So the chromatic number is (in some sense) a global statistic!

Theorem 3.10 (Markov's inequality)

Given a random variable X that only takes on nonnegative values, for all $a > 0$,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof.

$$\mathbb{E}[X] \geq \mathbb{E}\left[X \cdot 1_{X \geq a}\right] \geq \mathbb{E}\left[a 1_{X \geq a}\right] = a \Pr(X \geq a).$$

□

This is used with the mindset that if the expected value of X is small, then X is small with high probability.

Proof of Theorem 3.9. Construct an **Erdős-Renyi random graph** $G(n, p)$ with n vertices and each edge appearing with probability p. Here, let's let

$$p = n^{\theta - 1}, 0 < \theta < \frac{1}{\ell}.$$

Let X be the number of cycles of length at most ℓ. By expected value calculations, the number of such cycles is

$$\mathbb{E}[X] = \sum_{i=3}^{\ell} \binom{n}{i} \frac{(i-1)!}{2} p^i$$

since given any i vertices, there are $\frac{(i-1)!}{2}$ different cycles through them. This can be upper bounded by

$$\leq \sum_{i=3}^{\ell} n^i p^i \leq \ell n^\ell p^\ell.$$

Plugging in our choice of p, this evaluates to

$$\ell n^{\theta \ell} = o(n)$$

by our choice of θ. Now, what's the probability we have lots of short cycles? By Markov's inequality,

$$\Pr\left(X \geq \frac{n}{2}\right) \leq \frac{\mathbb{E}[X]}{n/2} = o(1),$$

so this allows us to find a graph with no cycles of length at most ℓ by the alteration method.

Meanwhile, what about the chromatic number? The easiest way to lower bound the chromatic number is to upper bound the independence number $\alpha(G)$, which is the size of the largest independent set. Note that every color class is an independent set (since no two vertices with the same color share an edge), so

$$|V(G)| \leq \chi(G)\alpha(G),$$

which is good for us as it gives a lower bound on the chromatic number. Well, the probability that we can have an independent set of size at least x is

$$\Pr(\alpha(G) \geq x) \leq \binom{n}{x} (1-p)^{\binom{x}{2}},$$

and if this quantity is small, we're good to lower bound the chromatic number. With more bounding,

$$\Pr(\alpha(G) \geq x) < n^x e^{-px(x-1)/2} = (ne^{-p(x-1)/2})^x$$

and by setting $x = \frac{3}{p} \log n$, this quantity becomes $o(1)$ as well.

We're almost done. Let n be large enough so that we have few cycles and large independent set size with high probability: $X \leq \frac{n}{2}$ and $\alpha \geq x$, each with probability greater than $\frac{1}{2}$. There now exists G with at least $\frac{n}{2}$ cycles of length ℓ and $\alpha(G) \leq \frac{3}{p} \log n$, and now remove a vertex from each short cycle (of length ℓ) to get a graph G'. The number of vertices of G' is now at least $\frac{n}{2}$, since we only removed at most $\frac{n}{2}$ cycles worth of vertices, and

$$\alpha(G') \leq \alpha(G) \leq \frac{3}{p} \log n,$$

so

$$\chi(G') \geq \frac{|V(G')|}{\alpha(G')} \geq \frac{np}{6 \log n} = \frac{n^\theta}{6 \log n} > k$$

for some sufficiently large n, and therefore G' is the graph we're looking for. □

3.4 Coloring edges

Recall that we defined $m(k)$ in Section 1.5 to be the minimum number of edges in a k-uniform hypergraph that is not 2-colorable. (Basically, we want to color the vertex sets red and blue so that no edge is monochromatic.) We found an upper and lower bound earlier: a randomized construction gives $m(k) \lesssim k^2 2^k$ using k^2 vertices, and $m(k) \geq 2^{k-1}$, just by randomly coloring the vertices, since each edge fails with some probability. Let's improve this lower bound now:

Theorem 3.11

$$m(k) \gtrsim \sqrt{\frac{k}{\log k}} 2^k.$$

Proof. Let's say a hypergraph H has m edges. Consider a random greedy coloring: choose a random mapping of the vertices to $[0, 1]$, and go from left to right, always coloring blue unless we would create a blue edge (in which case we color red).

What's the probability this gives a proper coloring? The only possible failures are red edges: call two edges e and f **conflicting** if they share exactly one vertex, and that vertex is the final vertex of e and first vertex of f. The idea here is that any failure must give a pair of conflicting edges.

So what's the probability that such a pair exists? Let's bound it: given two edges e and f that share exactly one vertex, the probability that they conflict is

$$P(e, f) = \frac{(k-1)!^2}{(2k-1)!} = \frac{1}{(2k-1)\binom{2k-2}{k-1}}.$$

Asymptotically, $\binom{n}{n/2}$ is $\frac{2^n}{\sqrt{n}}$ up to a constant factor, so the probability that these two edges conflict is $\Theta\left(\frac{1}{2^{2k}\sqrt{k}}\right)$. Now if $P(e, f)$ is less than $\frac{1}{m^2}$, we're happy, because there's less than m^2 edges and we can union bound the bad events. Doing some algebra, this gives

$$m(k) \gtrsim k^{1/4} 2^k.$$

Now let's be more clever. Split the interval $[0, 1]$ into $L = \left[0, \frac{1-p}{2}\right]$, $M = \left[\frac{1-p}{2}, \frac{1+p}{2}\right]$, $R = \left[\frac{1+p}{2}, 1\right]$. A pair of edges that conflict must have $e \subseteq L$, $e \subseteq R$, $f \subseteq L$, or $f \subseteq R$, or they both intersect in the middle.

The probability that e lies in L is just $\left(\frac{1-p}{2}\right)^k$ (each of the k vertices must be in L), and we can say similar things about the cases $e \subseteq R, f \subseteq L, f \subseteq R$. To deal with the middle intersection, if the common vertex between e and f is v, the probability that the second scenario happens is the probability that there are $(k-1)$ vertices to the left of v in M for e and $(k-1)$ vertices to the right of v in M for f. This is bounded by

$$\int_{(1-p)/2}^{(1+p)/2} x^k (1-x)^{k-1} dx \leq p \left(\frac{1}{4}\right)^{k-1}.$$

Putting all of this together, the probability of any pair of conflicting edges is bounded by

$$\leq 2m \left(\frac{1-p}{2}\right)^k + m^2 p \left(\frac{1}{4}\right)^{k-1}$$

and this is less than 1 if $m = c2^k \sqrt{\frac{k}{\log k}}$ and $p = \left(\log \frac{4m}{2^k}\right)/k$, and we've found a bound on m as desired. □

4 The Second Moment Method

Starting in this section, we shift the focus to that of **concentration**: essentially, can we say that the value of our random variable X is realtively close to the mean?

4.1 Refresher on statistics and concentration

We've been discussing expectations of the form $\mathbb{E}[X]$ so far, and let's say that we find $\mathbb{E}[X]$ to be large. Can we generally conclude that X is large or positive with high probability? No, because outliers can increase the mean dramatically.

So let's consider a sum of variables

$$X = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p).$$

If the X_is are independent, we know a lot by the central limit theorem: a lot of random variables will converge to a Gaussian or other known distribution in the large limit. But most of the time, we only have that our variables are "mostly independent" or not independent at all. Is there any way for us to still understand the concentration of the sum?

Definition 4.1
The **variance** of a random variable X is defined to be

$$\text{var}(X) = \mathbb{E}[X - \mathbb{E}[X]]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

We will often let μ denote the mean of a variable, σ^2 denote the variance, and define σ to be the (positive) **standard deviation** of X.

Proposition 4.2 (Chebyshev's inequality)
Given a random variable X with mean μ and variance σ^2, then for all λ,

$$\Pr(|x - \mu| \ge \lambda\sigma) \le \frac{1}{\lambda^2}.$$

Proof. The left hand side is equivalent to

$$\Pr((x - \mu)^2 \ge \lambda^2\sigma^2)$$

which, by Markov's inequality, is

$$\le \frac{\mathbb{E}[|x - \mu|^2]}{\lambda^2\sigma^2} = \frac{\sigma^2}{\lambda^2\sigma^2} = \frac{1}{\lambda^2}.$$

□

Why do we care about these results? The central idea is that if our standard deviation $\sigma \ll \mu$, then we have "concentration" of polynomial decay by Chebyshev.

Corollary 4.3 (of Chebyshev)

The probability that X deviates from its mean by more than ε times its mean is bounded as

$$\Pr(|X - \mathbb{E}[X]| \ge \varepsilon \mathbb{E}[X]) \le \frac{\text{var}(X)}{\varepsilon^2 \mathbb{E}[X]^2}.$$

In particular, if $\text{var}(X) = o(\mathbb{E}[X]^2)$, then $X \sim \mathbb{E}[X]$ with high probability.

Usually, variance is easy to calculate. This is because

$$\text{var}(X) = \text{cov}[X, X],$$

where $\text{cov}[X, Y]$ is the **covariance**

$$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Since this expression is bilinear, if $X = X_1 + \cdots + X_n$, we can expand this out as

$$\sum_{i,j} \text{cov}[X_i, X_j] = \sum_i \text{var}(X_i) + 2\sum_{i<j} \text{cov}[X_i, X_j]$$

Often the second term here is small, because each X_i is independent with many other X_js or there is low covariance between them.

Example 4.4 (Binomial distribution)

If $X = X_1 + X_2 + \cdots + X_n$, where each X_i is independently distributed via the Bernoulli distribution $\text{Bernoulli}(p)$, the mean is $\mathbb{E}[X] = np$, and $\sigma^2 = np(1-p)$. As long as $np \gg 1$, we have $\sigma \ll \mu$, so $X \sim \mu$ with high probability.

Later on, we'll get much better bounds. (By the way, we want $np \gg 1$ so our distribution doesn't approach a Poisson distribution instead)

Example 4.5

Let X be the number of triangles in a random graph $G(n, p)$, where each edge of K_n is formed with probability p.

Is this variable concentrated around its mean? It's pretty easy to compute that mean: X is the sum over all triangles

$$X = \sum_{\substack{i,j,k \in [n] \\ \text{distinct}}} X_{ijk}$$

where X_{ijk} is 1 if they form a triangle and 0 otherwise. Each X_{ijk} can be expanded out in terms of the indicator variables for edges:

$$X = \sum_{\substack{i,j,k \in [n] \\ \text{distinct}}} X_{ij} X_{jk} X_{ik}.$$

By linearity of expectation, each term is p^3, so $\mathbb{E}[X] = \binom{n}{3} p^3$. The variance is a bit harder, and we're mostly worried about the covariance term: when do those cross-terms come up?

Well, given a pair of triples T_1, T_2 of vertices, we can find the covariance for those triangles. If there is at most one vertex of overlap, no edges overlap, so there is no covariance. The others are a bit harder, but we use

$\mathrm{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$:

$$\mathrm{cov}[X_{T_1}, X_{T_2}] = \begin{cases} 0 & |T_1 \cap T_2| \le 1 \\ p^5 - p^6 & |T_1 \cap T_2| = 2 \\ p^3 - p^6 & T_1 = T_2 \end{cases}$$

So we can now finish the computation:

$$\mathrm{var}(X) = \binom{n}{3}(p^3 - p^6) + \binom{n}{2}(n-2)(n-3)(p^5 - p^6) \lesssim n^3 p^3 + n^4 p^5,$$

and we have $\sigma \ll \mu$ if and only if $p \gg \frac{1}{n}$. So this means that the number of triangles is concentrated around its mean with high probability if p is large enough! Later in the course, we will use other methods to prove better concentration.

Fact 4.6

It turns out that X satisfies an asymptotic central limit theorem:

$$\frac{X - \mu}{\sigma} \to N(0, 1).$$

This fact was initially proved by taking moments of the form $\mathbb{E}[X^n]$, and the idea is that if the moments agree with the Gaussian moments, we have a Gaussian distribution. But there's a newer method that can be used called the method of projections.

4.2 Threshold functions for subgraphs

We're going to try to look for small subgraphs in a large random graph $G(n, p)$. Here's an example:

Problem 4.7

For which $p = p_n$ (a sequence in terms of n) does $G(n, p)$ have a K_4 subgraph with high probability $1 - o(1)$?

Lemma 4.8

For any random variable X that takes on nonnegative values,

$$\Pr(X = 0) \le \frac{\mathrm{var}(X)}{\mathbb{E}[X]^2}.$$

Proof. The probability that $X = 0$ is at most the probability $|x - \mu| \ge \mu$, which is at most $\frac{\mathrm{var}(x)}{\mu^2}$ by Chebyshev's inequality. □

Corollary 4.9

Let X take on only nonnegative values. If the variance of X is much smaller than μ^2, then $X > 0$ with high probability.

Definition 4.10

$r(n)$ is a **threshold function** for a property P if $p = p_n \ll r(n)$ means that $G(n, p)$ satisfies P with low probability, while $p = p_n \gg r(n)$ means that $G(n, p)$ satisfies P with high probability.

Proposition 4.11

The threshold for a random graph to contain K_3 (triangles) is $\frac{1}{n}$, so the probability a graph contains a K_3 is 0 if $pn \to 0$ and 1 if $pn \to \infty$.

Proof. Let X be the number of triangles in $G(n,p)$. Recall that

$$\mu = \binom{n}{3}p^3 \sim \frac{n^3p^3}{6}, \sigma^2 = \text{var}(X).$$

If $p \ll \frac{1}{n}$, the mean $\mu = o(1)$, so by Markov's inequality, the probability X has at least one triangle vanishes:

$$\Pr(X \geq 1) \leq \mathbb{E}[X] = o(1).$$

On the other hand, if $p \gg \frac{1}{n}$, $\mu \to \infty$, while $\sigma \ll \mu$. So X is concentrated around its mean with high probability, making it positive with high probability. □

Problem 4.12

Given a subgraph H, what's the threshold for containing H?

Let $X = X_1 + \cdots + X_m$, where each X_i is an indicator variable for A_i. We let $i \sim j$ for $i \neq j$ to mean that A_i and A_j are not independent. So if $i \not\sim j$, then $\text{cov}[X_i, X_j] = 0$, but if $i \sim j$,

$$\text{cov}[X_i, X_j] = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] \leq \mathbb{E}[X_iX_j] = \Pr(A_i \cap A_j).$$

So expanding out the expression for variance,

$$\text{var}(X) = \sum_{i,j} \text{cov}[X_i, X_j] \leq \mathbb{E}[X] + \Delta,$$

where Δ is defined as (the bounded covariance term)

$$\sum_{i<j, i\sim j} \Pr(A_i \cap A_j).$$

So we approximate covariances by probabilities, but if there are very few dependent pairs, we really just care about the number of them. It's possible that all the X_is are all correlated, or that X_is are all nearly independent, but that's not the case here.

Corollary 4.13

If $\mathbb{E}[X] \to \infty$ and $\Delta = o(\mathbb{E}[X]^2)$, then with high probability, X is positive and concentrated around its mean.

Simplifying Δ,

$$\Delta = \sum_{i<j, i\sim j} \Pr(A_i \cap A_j) = \sum_i \Pr(A_i) \sum_{j:j\sim i} \Pr(A_j | A_i)$$

and usually the inner sum doesn't depend on i by symmetry. In such cases, we can define

$$\Delta^* = \sum_{j:j\sim i} \Pr(A_j | A_i).$$

We then have

$$\Delta = \sum_i \Pr(A_i)\Delta^* = \Delta^* \cdot \mathbb{E}[X],$$

and this means that if $\mathbb{E}[X] \to \infty$ and $\Delta^* \ll \mu$, X is positive and concentrated around its mean with high probability.

Proposition 4.14

The threshold for having K_4 as a subgraph is $n^{-2/3}$.

Proof. Let X be the random variable which is the number of K_4 graphs in $G(n,p)$. The expected value of X is

$$\mathbb{E}[X] = \binom{n}{4} p^6 \sim \frac{n^4 p^6}{24},$$

and if $p \ll n^{-2/3}$, then $\mu = o(1)$, so again by Markov, X is 0 with high probability.

On the other hand, if $p \gg n^{-2/3}$, the mean goes to infinity, and we'll look at the second moment by letting A_S be the event that we induce a K_4 on any set S of four vertices. then

$$\Delta^* \lesssim n^2 p^5 + np^3,$$

where n^2p^5 comes from sets sharing two vertices (which means we need to find two more and have 5 edges chosen with probability p), and np^3 comes from sets sharing three vertices (meaning we find one more and have 3 more edges chosen). Provided that $p \gg n^{-2/3}$, both terms here are small: $\Delta^* = o(\mathbb{E}[X])$, and we are done by Corollary 4.13. □

So it seems we should be able to do this with any graph H. But the idea with K_3 and K_4 was that any p with $\mu \to \infty$ gave $X > 0$ with high probability. In general, the answer isn't quite so simple.

Question 4.15. *Consider a K_4 with an extra edge attached to a vertex as the subgraph that we're looking for. What is its threshold density?*

The expected number of copies of this is $\mathbb{E}[X_H] \asymp n^5p^7$, so we might predict that the threshold is $p = n^{-5/7}$. Indeed, if $p \ll n^{-5/7}$, $\mathbb{E}[X]$ is very small, and we have zero copies with small probability. But now let's say $p \gg n^{-5/7}$ but $p \ll n^{-2/3}$. There are no K_4s, so there's no way we can have this graph at all. Finally, when $p \gg n^{-2/3}$, we have a bunch of K_4s: it can be shown that we can easily find another edge to connect to our K_4. Therefore, the threshold density is $n^{-2/3}$, and that threshold is not just dependent on the number of edges and vertices of our subgraph H!

In a way, this is saying that K_4s are the "hard part" of the graph to hit, and the next definition helps us quantify that.

Definition 4.16

Define $\rho(H) = \frac{e_H}{v_H}$, sometimes called the **density of H**, to be the ratio of edges to vertices in our graph H. H is **balanced** if every subgraph H' has $\rho(H') \le \rho(H)$. If H is not balanced, define the **maximum subgraph density** $m(H)$ to be the maximum of $\rho(H')$ across all subgraphs H'.

Example 4.17

Cliques are balanced: the initial density is $\frac{k-1}{2}$, and we can't do better. On the other hand, the K_4 plus an edge is not balanced, since $\rho = \frac{7}{5}$ but the ρ of K_4 is $\frac{3}{2}$.

In fact, $m(H)$ is actually what designates the threshold density:

Theorem 4.18

If we pick each edge of K_n with probability p, the threshold for having H as a subgraph is $p = n^{-\frac{1}{m(H)}}$.

The proof is very similar to what we've been doing.

Proof. Let H' be the subgraph with maximum density $\rho(H') = m(H)$. If p is below the threshold, the expected number of copies of H'

$$\mathbb{E}[X'_H] \asymp n^{V_{H'}} p^{e_{H'}} = o(1),$$

so with high probability $G(n, p)$ has no copies of H' and therefore no H.

Now if $p \gg n^{-1/m(H)}$, we want to compute the number of copies of H. For sets S of vertices with $|S| = v_H$,

$$\Delta^* = \sum_{T:|T|=v_H, |T\cap S|\geq 2} \Pr(A_T|A_S)$$

where T is the event that T contains a copy of H.

Doing cases based on the size of $T \cap S$ (like we did before), let's say T intersects S in k spots. Here's the key step where we use the maximum subgraph density: overlaps in the covariance terms are subgraphs of H. If H' is the overlap between S and T, the contribution to Δ^* is

$$\lesssim n^{v'_H} p^{e'_H} \ll n^{v_H} p^{e_H}$$

for all H', so if we keep track of all the overlaps, we find that $\Delta^* = o(1)$, meaning all overlaps don't contribute much. This finishes the proof by Corollary 4.13. □

4.3 Clique number

Question 4.19. *What can we say about $\omega(G)$, the number of vertices in the maximum size clique of G, if each edge in K_n is included with probability $\frac{1}{2}$?*

We can't quote any of the results from last time, since we're not sticking to fixed-size subgraphs. But this is still not too hard to calculate from first principles.

Let $f(k)$ be the expected number of k-cliques: this is just $\binom{n}{k}2^{-\binom{k}{2}}$ by linearity of expectation. We can have a naive guess: perhaps we have a clique whenever this quantity goes to infinity and not when the quantity goes to 0.

Theorem 4.20

Let $k = k(n)$ be a function such that $f(k) = \binom{n}{k}2^{-\binom{k}{2}}$ goes to infinity. Then

$$\omega\left(G\left(n, \frac{1}{2}\right)\right) \geq k$$

with high probability.

Proof. For all subsets S of the vertices of size k, let A_S be the event that S is a clique, and let χ_S be the indicator variable for A_S. Then the number of k-cliques

$$X = \sum_S \chi_S$$

has expectation $f(k)$, and we want to show that the variance is much smaller than the mean squared. This is very similar to the earlier proof: fixing S, we can find Δ^* by summing over all T that intersect S in at least two vertices (those are the only ones that can be dependent on S):

$$\Delta^* = \sum_{T:|T\cap S|\geq 2} \Pr(A_T|A_S).$$

We can write this down explicitly, since the expression $\Pr(A_T|A_S)$ just depends on the size of the intersection:

$$= \sum_{i=2}^{k} \left(\binom{k}{i}\binom{n-k}{k-i}\right) 2^{\binom{i}{2}-\binom{k}{2}}$$

where the first term is the number of ways to choose T with an overlap of i vertices, and the power of 2 is the probability that T is a clique given that the i vertices in S are all connected. This does indeed turn out to be small enough: omitting the detailed calculations,

$$\Delta^* \ll \binom{n}{k} 2^{-\binom{n}{k}} = \mathbb{E}[X],$$

so we're done. □

We also know by Markov's inequality that if the expected value goes to 0, the probability of having a k-clique is $o(1)$. The idea is that if there's some value k such that $f(k+1) \gg 1$ and $f(k) \ll 1$, then we have a distinctive threshold. But it might be that one of the fs is constant order, and then the theorem doesn't actually let us know what happens for that specific value of k.

Theorem 4.21

There exists a $k_0 = k_0(n)$ such that with high probability,

$$\omega\left(G\left(n, \frac{1}{2}\right)\right) \in \{k_0, k_0+1\}$$

and $k_0 \sim 2\log_2 n$.

This is known as **two-point concentration**. Rephrasing this, if we create this graph at random, we expect one of two values for the clique number.

Proof sketch. We can check that for $k \sim 2\log_2 n$,

$$\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1} 2^{-k} = n^{-1+o(1)} = o(1).$$

(In particular, the gap between two adjacent ks is too large to allow a bunch of ks to give constant order $f(k)$s.) Then let $k_0 = k_0(n)$ be the value such that

$$f(k_0) \geq 1 > f(k_0+1);$$

then $f(k_0-1) \gg 1$ and $f(k_0+2) \ll 1$. □

It turns out for most but not all values of n, there is only one k_0 that ω takes on with high probability! Later in this class, we'll be able to say something more specific.

4.4 Chromatic number

Question 4.22. *What is the expected chromatic number (maximum number of colors needed for a proper coloring) in a random graph $G\left(n, \frac{1}{2}\right)$?*

Remember that we have the result $\chi(G)\alpha(G) \geq n$, because each color class is an independent set (and therefore one of them has size at least $\frac{n}{\chi(G)}$).

Corollary 4.23

The expected independence number of G is also $\sim 2\log_2 n$, since

$$\alpha(G) = \omega(\overline{G}),$$

since including an edge in G with probability $\frac{1}{2}$ is equivalent to including it in $\overline{G}$ with probability $\frac{1}{2}$.

So this means we can guarantee

$$\chi(G) \geq \frac{n}{\alpha(G)} \sim \frac{n}{2\log_2 n}.$$

Do we also have an upper bound? Can we show that we can color $G\left(n, \frac{1}{2}\right)$ with that many colors?

Theorem 4.24 (Bollobás, 1987)

The chromatic number

$$\chi\left(G\left(n, \frac{1}{2}\right)\right) \sim \frac{n}{2\log_2 n}.$$

We'll see how to prove this later on using martingale convergence.

4.5 Number theory

This class was advertised as using probability to solve problems that don't involved probability. The next few examples have no randomness inherently, but we'll still use the second moment method to solve them.

Let $\nu(n)$ denote the number of prime divisors of n, not counting multiplicity. Can we figure out the typical size of $\nu(n)$ just given n?

Theorem 4.25 (Hardy - Ramanujan 1920)

For all ε, there exist a constant c such that all but ε fraction of the numbers $[1, n]$ satisfy

$$|\nu(x) - \log\log n| \leq c\sqrt{\log\log n}.$$

Remark. log *refers to natural log in number theory contexts.*

Proof by Turán, 1934. We're going to use a basic intuition about a "random model of the primes." Statistically, they have many properties that make them seem random, even if the primes themselves are not.

Pick a random $x \in [n]$. For each prime p, let X_p be the indicator variable

$$X_p = \begin{cases} 1 & p|x \\ 0 & \text{otherwise.} \end{cases}$$

Then the number of prime divisors of x less than or equal to M is approximately

$$X = \sum_{p \leq M} X_p,$$

where we pick $M = n^{1/10}$, a constant power of n. Then there are at most 10 prime factors of x larger than M, so

$$\nu(x) - 10 \le X \le \nu(x).$$

Since we're dealing with asymptotics, that constant is okay for our purposes here. We're treating X as a random variable: we want to show that it is concentrated and that its mean is around $\log\log n$. Each X_p is also a random variable, so this is a good use of the second moment method: we have

$$\mathbb{E}[X_p] = \frac{\lfloor n/p \rfloor}{n} = \frac{1}{p} + O\left(\frac{1}{n}\right)$$

for each prime p, so the mean of the random variable is

$$\mathbb{E}[X] = \sum_{p \le M} \left(\frac{1}{p} + O\left(\frac{1}{n}\right)\right).$$

We'll now use a basic result from analytic number theory:

Theorem 4.26 (Merten's theorem)

Adding over all primes up to N,

$$\sum_{p \le N} \frac{1}{p} = \log\log N + O(1).$$

To find the expected value of X^2, we need to understand the covariance between different X_ps. For any primes $p \ne q$,

$$\text{cov}[X_p, X_q] = \mathbb{E}[X_pX_q] - \mathbb{E}[X_p]\mathbb{E}[X_q] = \frac{\lfloor n/(pq) \rfloor}{n} - \frac{\lfloor n/p \rfloor}{n}\frac{\lfloor n/q \rfloor}{n} \le \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right) \le \frac{1}{n}\left(\frac{1}{p} + \frac{1}{q}\right).$$

The idea is that these variables are basically independent by Chinese Remainder Theorem, except for the "edge cases" near n. So the total sum of the covariances is

$$\sum_{p \ne q, p,q \le M} \text{cov}[X_p, X_q] \le \frac{1}{n} \sum_{p \ne q, p,q \le M} \left(\frac{1}{p} + \frac{1}{q}\right) \le \frac{2M}{n} \sum_{p \le M} \frac{1}{p} \lesssim n^{-9/10} \log\log n = o(1),$$

since $M = n^{1/10}$. Now the variance of X is

$$\text{var}(X) = \sum_p \text{var}(X_p) + o(1) = \log\log n + O(1)$$

(which is not very large), and therefore the standard deviation is on the order of $\sqrt{\log\log n}$. Now by Chebyshev's inequality,

$$\Pr\left(|x - \log\log n| \ge \lambda\sqrt{\log\log n}\right) \le \frac{1}{\lambda^2} + o(1),$$

and since X is within 10 of $\nu(x)$, we've shown concentration with high probability (just pick λ to be whatever constant we need in terms of ε). □

What's the distribution, though? Is $\sqrt{\log\log n}$ the right order of magnitude? If we really believe the X_ps are independent, we should believe in the central limit theorem.

Theorem 4.27 (Erdős-Kac theorem)

Picking a random $x \in [n]$, $\nu(x)$ is asymptotically normal:

$$\Pr_{x\in[n]}\left(\frac{\nu(n) - \log\log n}{\sqrt{\log\log n}} \geq \lambda\right) = \frac{1}{\sqrt{2\pi}}\int_\lambda^\infty e^{-t^2/2}dt$$

for all $\lambda \in \mathbb{R}$.

We briefly mentioned the method of moments earlier: instead of looking at second moments, look at higher moments as well. There's a theorem in probability that if all the moments of our function are the same as certain distributions (including the normal distribution), then convergence happens.

We can do this explicitly if we want, but it gets a bit tedious. Here's a trick that simplifies the calculation: let's compare $\mathbb{E}[X^k]$ with that of an "idealized" random variable Y.

Proof. This time, set $M = n^{1/s(n)}$ where $s(n) \to \infty$ slowly. Choosing $s(n) = \log\log\log n$ is fine, but $s(n)$ can't grow too quickly because we have that

$$\nu(x) - s(n) \leq X \leq \nu(x).$$

(Joke: What's the sound a drowning number theorist makes?...) So now let

$$Y = \sum_{p\leq M} Y_p,$$

where Y_p is now idealized to Bernoulli$\left(\frac{1}{p}\right)$, independent of the other variables. This is supposed to model X_p. So now let

$$\mu = \mathbb{E}[Y] \sim \mathbb{E}[X],$$

and

$$\sigma^2 = \text{var}(Y) \sim \text{var}(X).$$

Set

$$\tilde{X} = \frac{X-\mu}{\sigma}, \tilde{Y} = \frac{Y-\mu}{\sigma}.$$

By the central limit theorem, we know that $\tilde{Y}$ converges to $N(0,1)$. Now let's compare $\tilde{Y}$ and $\tilde{X}$, showing that for all k,

$$\mathbb{E}[\tilde{X}^k] = \mathbb{E}[\tilde{Y}^k],$$

which are (by the central limit theorem) also equal to $\mathbb{E}[Z^k]$ for the standard normal distribution.

When we expand out the factors of $\mathbb{E}[X^k - Y^k]$ for distinct primes $p_1, \cdots, p_r \leq M$, they look like

$$\mathbb{E}[X_{p_1}X_{p_2}\cdots X_{p_r} - Y_{p_1}\cdots Y_{p_r}] = \frac{1}{n}\left\lfloor\frac{n}{p_1\cdots p_r}\right\rfloor - \frac{1}{p_1\cdots p_r} = O\left(\frac{1}{n}\right).$$

So if we compare the expansions of $\tilde{X}^k$ in terms of the X_ps, there's $M^k = n^{o(1)}$ terms. Since each term contributes $O\left(\frac{1}{n}\right)$, the moments are essentially the same:

$$\mathbb{E}[\tilde{X}^k - \tilde{Y}^k] = n^{-1+o(1)} = o(1).$$

Since all moments converge, $\tilde{X}$ converges to the normal distribution asymptotically. □

4.6 Distinct sums

Question 4.28. *What's the size of the largest subset $S \subseteq [n]$ such that all $2^{|S|}$ subset sums of S are distinct?*

Example 4.29

We can take $S = \{1, 2, 4, \cdots, 2^k\}$, where $k = \lfloor \log_2 n \rfloor$. All sums are distinct by base-2 expansion.

This set has size $\log_2(n)$. Is there any way we can do much better?

Problem 4.30 (Open; Erdős offered $300 for this one)

Prove or disprove: $|S| \le \log_2 n + O(1)$.

One thing we know is that all subset sums have size at most $n|S|$, since there are only $|S|$ things we can add. There are $2^{|S|}$ sums, so if they're all distinct, by Pigeonhole, we must have $2^{|S|} \le n|S|$, which rearranges to

$$|S| \le \log_2 n + \log_2 \log_2 n + O(1).$$

Can we formulate a better argument than this? Let's try the second moment method! The idea is that if we pick a random subset sum, we should expect some concentration around the mean.

Theorem 4.31

Every subset $S \subseteq [n]$ with distinct subset sums has

$$|S| \le \log_2 n + \frac{1}{2} \log_2 \log_2 n + O(1).$$

Proof. Given our set $S = \{x_1, \cdots, x_k\}$, define a random variable $X = \varepsilon_1 x_1 + \cdots + \varepsilon_k x_k$ where $\varepsilon_i \in \{0, 1\}$ uniformly and independently. The mean is (by linearity of expectation) just $\frac{1}{2}(x_1 + \cdots + x_k)$, and the variance is

$$\sigma^2 = \frac{1}{4}(x_1^2 + \cdots + x_k^2) \le \frac{n^2 k}{4},$$

since all $x_i \le n$. By Chebyshev's inequality, for all $\lambda > 1$,

$$\Pr\left[|X - \mu| < \frac{\lambda n \sqrt{k}}{2}\right] \ge 1 - \frac{1}{\lambda^2}.$$

But X must take distinct values for all different instantiations, so the probability that $X = x$ is at most 2^{-k} for each x. This means that in the probability expression above, X must lie in the range $\left[\mu - \frac{\lambda n \sqrt{k}}{2}, \mu + \frac{\lambda n \sqrt{k}}{2}\right]$, which has a probability

$$\Pr\left[|X - \mu| < \frac{\lambda n \sqrt{k}}{2}\right] \le 2^{-k} \cdot (\lambda n \sqrt{k} + 1).$$

Putting these inequalities together,

$$1 - \frac{1}{\lambda^2} \le 2^{-k}(\lambda n \sqrt{k} + 1),$$

which rearranges to

$$n \ge \frac{2^k(1 - \lambda^{-2}) - 1}{\sqrt{k}\lambda}.$$

We can choose λ to optimize this expression: in this case, $\lambda = \sqrt{3}$ yields the desired result. □

4.7 An application to analysis

We're going to prove the following result using the second moment method:

Theorem 4.32 (Weierstrass approximation theorem)
Let $f : [0,1] \to \mathbb{R}$ be a continuous function on a bounded interval. Given $\varepsilon > 0$, it is possible to approximate f by a polynomial $p(x)$ such that

$$|p(x) - f(x)| \le \varepsilon \ \forall x \in [0,1].$$

Proof. First of all, since $[0,1]$ is compact, f is uniformly continuous and is therefore bounded. In other words, there exists a δ such that

$$|f(x) - f(y)| \le \frac{\varepsilon}{2} \text{ for all } x, y \text{ with } |x-y| \le \delta.$$

Rescale f so that it is bounded by 1, so now $|f(x)| \le 1$ for all x. Let X be a random variable $X \sim \text{Binomial}(n,x)$: then

$$\Pr(X = j) = \binom{n}{j} x^j (1-x)^{n-j} \text{ for all } 0 \le j \le n.$$

We know the statistics $\mathbb{E}[X] = nx, \text{var}(X) = nx(1-x) \le n$. So by Chebyshev,

$$\Pr\left[|X - nx| > n^{2/3}\right] \le n^{-1/3}.$$

In particular, if we take n fixed but large enough – let $n > \max(64\varepsilon^{-3}, \delta^{-3})$ – we can bound this in terms of ε:

$$\Pr\left[|X - nx| > n^{2/3}\right] < \frac{\varepsilon}{4}.$$

We can now write down our approximating polynomial explicitly:

$$\boxed{P_n(x) = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} f\left(\frac{i}{n}\right)}.$$

Basically, chop up $[0,1]$ into n intervals and sample the value at each one. We claim that this works: we do have $|P_n(x) - f(x)| \le \varepsilon$ for all $x \in [0,1]$. To show this, note that by the triangle inequality,

$$|P_n(x) - f(x)| \le \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \left| f\left(\frac{i}{n}\right) - f(x) \right|$$

implicitly using that the sum of $\binom{n}{i} x^i (1-x)^{n-i} = 1$. The idea is that this absolute value is small if $x \approx \frac{i}{n}$; otherwise, Chebyshev bounds the contribution! We'll split this up into two terms - those close and far away from our given x:

$$= \sum_{i : \left|\frac{i}{n} - x\right| \le n^{-1/3}} \binom{n}{i} x^i (1-x)^{n-i} \left| f\left(\frac{i}{n}\right) - f(x) \right| + 2 \left(\sum_{i : \left|\frac{i}{n} - x\right| > n^{-1/3}} \binom{n}{i} x^i (1-x)^{n-i} \right),$$

where the 2 comes from the fact that $|f(x)| \le 1$. But now note that the absolute value in the first term deals with those x within δ of $\frac{i}{n}$, and the second term was bounded earlier:

$$\le \sum_{i : \left|\frac{i}{n} - x\right| \le n^{-1/3}} \left(\binom{n}{i} x^i (1-x)^{n-i} \cdot \frac{\varepsilon}{2} \right) + 2 \cdot \frac{\varepsilon}{4},$$

and now both terms are at most $\frac{\varepsilon}{2}$, so this is at most ε, as desired. □

5 The Chernoff bound

5.1 Setup and proof

The second moment essentially compares the values of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ to each other. Why do we not take higher moments? In general, if we have independent random variables

$$X = X_1 + \cdots + X_n,$$

we can look at $p(X)$, which is some polynomial in X, and apply Markov's inequality in the same way that we did for Chebyshev's inequality. It turns out that if we're allowed to look at arbitrarily high-degree polynomials, it's usually better to just look at the following object:

Definition 5.1

The **moment generating function** of a random variable X is a function of t

$$\mathbb{E}[e^{tX}] = 1 + t\mathbb{E}[X] + \frac{t^2\mathbb{E}[X^2]}{2} + \cdots .$$

What are its applications?

Theorem 5.2 (Chernoff bound)

Let $S_n = X_1 + \cdots + X_n$, where $X_i = \pm 1$ uniformly and independently. Then for all $\lambda > 0$,

$$\Pr(S_n \geq \lambda\sqrt{n}) \leq e^{-\lambda^2/2}.$$

This gives better tail decay! While the second moment method only gave us polynomial decay (right hand side of the form $\frac{1}{\lambda^2}$), this is exponential decay instead.

Proof. Let $t \geq 0$ be a real number, and consider the moment generating function

$$\mathbb{E}[e^{tS_n}].$$

Since S_n is a sum of random independent variables, this is

$$\mathbb{E}[e^{tX_1 + \cdots + tX_n}] = \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}] \cdots \mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX_1}]^n = \left(\frac{e^{-t} + e^t}{2}\right)^n.$$

Since $\frac{e^{-t}+e^t}{2} \leq e^{t^2/2}$ by comparing coefficients of the Taylor expansions:

$$\frac{1}{(2n)!} \leq \frac{1}{n!2^n},$$

our moment generating function is $\leq e^{nt^2/2}$, and by Markov's inequality,

$$\Pr(S_n \geq \lambda\sqrt{n}) \leq \frac{\mathbb{E}[e^{tS_n}]}{e^{t\lambda\sqrt{n}}} \leq e^{-t\lambda\sqrt{n} + t^2 n/2}$$

and setting $t = \frac{\lambda}{\sqrt{n}}$ gives the desired result. □

By symmetry, we have a bound for $S_n \leq \lambda\sqrt{n}$ as well, so combining these, we obtain the following:

Corollary 5.3

Using the definition of S_n above,

$$\Pr\left(|S_n| \geq \lambda\sqrt{n}\right) \leq 2e^{-\lambda^2/2}$$

for all $\lambda > 0$.

But notice that S_n converges to a Gaussian distribution for large n, so something similar should be true for Gaussians as well. This is indeed true:

Fact 5.4

For the standard normal distribution $Z \sim N(0,1)$, for all $\lambda \geq 0$,

$$\Pr(Z \geq \lambda) = \Pr(e^{tZ} \geq e^{t\lambda}) \leq e^{-t\lambda}\mathbb{E}[e^{tZ}] = e^{-t\lambda + t^2/2} \leq e^{-\lambda^2/2}$$

by taking $t = \lambda$.

This is pretty tight: it turns out that in general we're only losing a $c\sqrt{\lambda}$, and in reality we actually have

$$\Pr(Z \geq \lambda) \sim \frac{e^{-\lambda^2/2}}{\sqrt{2\pi\lambda}}.$$

See Appendix A of the textbook for different instantiations of the Chernoff bound. Similarly, we can find exponential decay for Bernoulli variables where $p \neq \frac{1}{2}$:

Fact 5.5

If Y is a sum of independent Bernoulli variables (with not necessarily the same probability), then for all $\varepsilon > 0$,

$$\Pr\left(|Y - \mathbb{E}[Y]| \geq \varepsilon\mathbb{E}[Y]\right) \leq 2e^{-C_\varepsilon \mathbb{E}[Y]}$$

for some constant $C_\varepsilon > 0$.

5.2 An application: discrepancy

Theorem 5.6

Let H be a k-uniform hypergraph with m edges. Then we can color the vertices red and blue so that every edge has an $O(\sqrt{k \log m})$ difference in the number of red and blue vertices.

Proof. Color each vertex uniformly at random: put ±1 on every vertex. Then every edge is of the form $S_m = X_1 + \cdots + X_m$ where all $X_m = \pm 1$, so by the Chernoff bound, the probability $|S_m|$ exceeds $\lambda\sqrt{k}$ is at most $2e^{-\lambda^2/2}$. Note that the absolute value of S_n is exactly the difference between the number of red and blue vertices.

In particular, we can now do a union bound: if $2me^{-\lambda^2/2} \leq 1$, then there exists a graph where none of the bad events happen. Inverting this gives the desired result. □

This kind of log term usually comes from the Chernoff bound. If we only used the second moment method, we'd have a much worse result - polynomial instead of exponential?

Well, what's the truth? Suppose $m = k$: this theorem gives us a difference of $\sqrt{k \log k}$ between the red and blue vertices. But we can do much better:

Fact 5.7

Spencer's paper "Six standard deviations suffice" says that when $m = k$, we can get at most $6\sqrt{k}$ difference between the number of red and blue vertices on every edge.

5.3 Chromatic number and graph minors

Let's start with some graph theory results for motivation:

Proposition 5.8 (Kuratowski's theorem)

If G is not planar, then it contains a $K_{3,3}$ or K_5 subdivision.

Here, a **subdivision** of a graph H is H with some of the edges chopped into smaller pieces. Basically, K_5s and $K_{3,3}$s are not allowed, nor are those graphs with extra vertices along the edges. There's another similar theorem that is actually equivalent to Kuratowski's theorem:

Proposition 5.9 (Wagner's theorem)

If G is not planar, then G contains a $K_{3,3}$- or K_5-minor.

Here, H is a **minor** of G if it can be obtained from deleting edges/vertices or **contracting** an edge. (Basically, take the two vertices of an edge and squish them together.) In particular, K_5 is a minor of a K_5-subdivision.

Theorem 5.10 (Four-color theorem)

If $\chi(G) \geq 5$, then G is not planar.

In particular, if $\chi(G) \geq 5$, it must contain a $K_{3,3}$-minor or a K_5-minor. Having a K_5-minor seems pretty relevant, since we need 5 colors to color a K_5. But $K_{3,3}$ doesn't seem like as much of an obstruction, and that's quantified in the statement below:

Fact 5.11

If $\chi(G) \geq 5$, then G contains a K_5-minor.

Well, does this hold if we replace 5 with other numbers?

Conjecture 5.12 (Hadwidger's conjecture)

If $\chi(G) \geq t$, then G contains a K_t-minor.

Many people consider this to be the biggest open problem in graph theory! We do have some small cases resolved: $t = 1, 2$ are trivial. $t = 3$ is not too hard: If G has no K_3-minor, it is a tree, which is 2-colorable. $t = 4$ requires more work but is elementary, and $t = 5$ is equivalent to the four-color theorem (for which we only have a computer-assisted proof). But Robertson, Seymour, and Thomas showed that the four-color theorem actually implies $t = 6$, and all $t \geq 7$ are open.

Are there variations on this conjecture?

Proposition 5.13 (Hajos conjecture: even stronger)
If $\chi(G) \geq t$, then G has a K_t-subdivision.

Unfortunately, this is false. In fact, by the probabilistic method, Erdős and Fajtlowicz showed that $G\left(n, \frac{1}{2}\right)$ fails this condition with high probability:

Theorem 5.14
With high probability, $G\left(n, \frac{1}{2}\right)$ has chromatic number $\chi(G) \geq (1 + o(1))\frac{n}{2\log_2 n}$ and no $K_{\lceil 10\sqrt{n}\rceil}$-subdivision.

So the theorem is very false in the relation between the two parameters, as well as in its likelihood! Note that the Hajos conjecture is still true for small t: it just fails for larger t due to the arguments below.

Proof. We already lower bounded by upper bounding color classes as independent sets:

$$\chi(G) \geq \frac{n}{\alpha(G)} \sim \frac{n}{2\log_2 n}$$

with high probability. Let's work on the second part.

Suppose we have a K_t-subdivision, where $t = \lceil 10\sqrt{n}\rceil$. Out of the $\binom{t}{2}$ edges in K_t, about half of them are not contained in G, so they must use up other vertices to form paths, and we don't have enough of those.

Let's do this more rigorously. Let G have a K_t subdivision $S \subset V$, where $|S| = t = \lceil 10\sqrt{n}\rceil$. At most n edges in the subdivision can be paths of at least 2 edges (rather than just straight lines between vertices), since each path takes up an external vertex, and all paths use distinct vertices by definition. So the number of edges E involved in the subdivision satisfies

$$E \geq \binom{t}{2} - n \geq \frac{3}{4}\binom{t}{2},$$

where the $\geq$ comes from picking a large enough constant in $t = c\sqrt{n}$ (we chose $c = 10$). But this inequality fails with high probability, since we're supposed to have $\frac{1}{2}\binom{t}{2}$ edges only. Indeed, for every fixed t-vertex S, each edge appears with probability $\frac{1}{2}$, so the number of edges in the subgraph induced by S satisfies

$$\Pr\left(E \geq \frac{3}{4}\binom{t}{2}\right) \leq e^{-t^2/10}$$

by the Chernoff bound. Now by a union bound, ranging over all t-element subsets of vertices, the probability of any subdivision being possible is bounded above by

$$\binom{n}{t} e^{-t^2/10} < n^t e^{-t^2/10} = o(1),$$

so there must not be a K_t-subdivision with high probability. $\square$

6 The Lovász local lemma

The local lemma is an extremely important tool, and we saw a basic use of it early on when finding bounds for diagonal Ramsey numbers. Recall that the vanilla way to do this (a union bound) doesn't use the fact that there are few local dependencies between our "bad events."

Most of our proofs in the previous few sections worked with high probability, but now we are shifting gears: this method is getting us a **small but nonzero** probability of success.

Theorem 6.1 (Local lemma, symmetric case)
Given events $A_1, \cdots, A_n$, where all $\Pr(A_i) \le p$, if all A_is are mutually independent from the set of all A_js except at most d of them, and

$$ep(d+1) \le 1,$$

then with positive probability, none of the events occur.

(e is the best constant we can have here.) If the ps are small, for instance if all $A_i \le \frac{1}{n}$, union bounding tells us that we have a scenario where none of the events happen. Alternatively, if the A_is were independent, we can just multiply the non-event probabilities. So this is a sort of happy medium between the two extremes!

Definition 6.2
An event A_0 is **mutually independent** of $\{A_1, \cdots, A_m\}$ if A_0 is independent of every event of the form $B_1 \cap B_2 \cap \cdots \cap B_m$, where all $B_i = A_i$ or $\overline{A_i}$.

In other words, any boolean information tells us nothing about A_0. As a warning, **this is different from saying that A_0 and A_i are independent for all** i, since events can be pairwise independent but not mutually independent. (For example, consider the three variables $X_1, X_2, X_1 + X_2 \bmod 2$, where X_1 and X_2 are uniform among $\{0, 1\}$.)

Almost all applications of the local lemma are of the following form: we have a set of variables Ω, and each A_i depends on a subset $S_i \subset \Omega$, so two events A_i and A_j are independent if $S_i \cap S_j \ne \varnothing$.

Let's start by looking at a few applications before returning to the proof of the theorem.

6.1 Coloring: hypergraphs and real numbers

Let's first consider a hypergraph: we wish to color all vertices red and blue so that no edge is monochromatic.

Theorem 6.3
A k-uniform hypergraph G is 2-colorable if every edge intersects at most $d = \frac{2^{k-1}}{e} - 1$ other edges.

Proof. Color the edges uniformly at random. For each edge f, let A_f be the event that f is monochromatic: each occurs with probability $2^{-k+1} \equiv p$.

Each bad event A_f is mutually independent of all other events $A_{f'}$ if f and f' do not share any vertices. (Note that here our probability space Ω is the vertices of our graph G.) By the Lovász local lemma, since $ep(d+1) \ge 1$, with positive probability, there is a graph whose coloring is proper. □

What are some consequences of this?

Question 6.4. *For which k is every k-uniform, k-regular hypergraph 2-colorable?*

A triangle is not 2-colorable, and the Fano plane is not 3-colorable, so our statement is false for $k = 2, 3$. Well, if we apply the theorem we've just proved, every edge intersects $k(k-1)$ other edges, so if

$$k(k-1) \geq \frac{2^{k-1}}{e} - 1,$$

then the statement is true. It turns out that is good enough for $k \geq 9$ - what about $4 \leq k \leq 8$? It's known that the statement is actually true for all $k \geq 4$, but that is much harder to prove.

Let's ask a related question now: say we're looking at k-colorings of the real numbers. Basically, we have some function that assigns one of the first k positive integers to each real number:

$$c : \mathbb{R} \to [k],$$

or we can think of our domain as $\mathbb{Z}$ instead if the Axiom of Choice is annoying to deal with. Say that a subset $T \subset \mathbb{R}$ is **multicolored** with respect to c if all k colors appear in $c(T)$.

Theorem 6.5

Fix k. Given a subset S of the reals with $|S| = m$, we can color the real numbers with k colors so that every translate of S is multicolored if

$$e(m(m-1)+1)k\left(1 - \frac{1}{k}\right)^m \leq 1.$$

Doing some calculations (omitted), this can be written as $m > (3+\varepsilon)k \log k$ for sufficiently large k. This is a hypergraph coloring problem, but we can still use the Lovász local lemma to solve it.

Proof. First, we'll show that we can color every finite subset $X \subset \mathbb{R}$ such that every $x + S \subset X$ is multicolored.

Color X uniformly at random among the k colors. Our "bad events" correspond to elements $x \in X$ where $x + S$ is not multicolored, so we can think of having a hypergraph where the vertices are elements of X and the edges correspond to translates. Each fixed translate $x + S \subset X$ is not multicolored with probability (union bound, pick a color to not include)

$$p \leq k\left(1 - \frac{1}{k}\right)^m.$$

Furthermore, each translate of S intersects at most $m(m-1)$ other translates (pick a pair $a \in S, a' \in S'$ to overlap). So now by the local lemma, there exists a good coloring for every finite set as long as the $ep(d+1) \leq 1$ condition is satisfied.

But how do we extend this to the reals? We use compactness: Tikhonov's theorem says that if we assign a discrete topology to k points, then $[k]^{\mathbb{R}}$ is compact. A point in $[k]^{\mathbb{R}}$ is a map from the reals to 1 through k, which is basically a coloring.

So for every $x \in \mathbb{R}$, let $C_x \subset [k]^{\mathbb{R}}$ be a subset of colorings so that $x + S$ is multicolored. Our goal is to show that we can make all of the translates multicolored at once. We've shown so far that every finite intersection of C_xs is nonempty by the local lemma. Under the product topology, C_x is closed for all $x \in \mathbb{R}$ (the set C_x is only limited by the finite set of reals in X), so the intersection of all closed sets

$$C = \bigcap_{x \in \mathbb{R}} C_x$$

is nonempty as well, and that's a coloring of all the real numbers. □

The logic here is "we have a bunch of closed sets, any finite intersection is nonempty, so the infinite intersection is nonempty." To elaborate on this, think of $\mathbb{Z}$ instead of $\mathbb{R}$. Then this is a diagonalization argument: if we can color every prefix of the positive integers, then we can write down something for each prefix, and find infinitely many ways of coloring 1, then 2, and so on.

6.2 Coverings of $\mathbb{R}^3$

Here's a motivating fact that we won't prove:

Fact 6.6 (Mani-Leviska, Pach)
For every $k \geq 1$, there exists a nondecomposable k-fold covering of $\mathbb{R}^3$ by open unit balls. Here, a k-fold covering means that every point in $\mathbb{R}^3$ is covered at least k times. A covering is **decomposable** if it can be partitioned into two coverings.

(This generalizes beyond 3 dimensions, by the way.) Maybe all the points are covered a uniform number of times: can we say that there aren't outliers in our covering? The answer is no:

Theorem 6.7
There exists an absolute constant $c > 0$ such that every nondecomposable k-fold covering of $\mathbb{R}^3$ by open unit balls must cover some point at least $c2^{k/3}$ times.

Proof. If we try to decompose a covering, we're coloring the unit balls red and blue so that each color forms a covering of $\mathbb{R}^3$ on its own. Our "bad events" here are where $x \in \mathbb{R}^3$ is only colored by one color - if no bad events occur, then we have a successful decomposition.

Construct an infinite hypergraph where the vertices F are the set of balls and the edges are

$$E_x = \text{ the set of balls in F containing } x.$$

So the edges are

$$E(H) = \{E_x : x \in \mathbb{R}^3\}$$

where two points in the same "cell" correspond to the same edge - this means decomposable is the same as 2-colorable in our hypergraph.

For the sake of contradiction, assume every point is covered at most t times, where t is to be determined. By a compactness argument, it suffices to show that every finite subgraph of H is 2-colorable.

We claim that every edge intersects $\lesssim t^3$ other edges.Two edges intersect if they share at least one ball in common: if we fix $E_x \cap E_y$ that are intersecting, since every point is covered at most t times, there are at most $4^3 t$ balls involved in E_y, since anything intersecting with x has to be within a ball of radius 4. It turns out m balls can cut $\mathbb{R}^3$ into at most $m^3 + 1$ regions, and each region corresponds to an edge. So there are at most $4^9 t^3 + 1$ other edges that any edge can intersect.

So now the rest is just applying the Lovász local lemma. Each edge is monochromatic with probability at most 2^{-k+1} (since it is covered at least k times), and in the dependency graph, the number of intersections is at most $\lesssim m^3$. So we can apply local lemma if

$$t^3 2^{-k+1} < c$$

for some sufficiently small constant c. So for $t \lesssim 2^{k/3}$, the graph is decomposable by Lovász, and therefore every nondecomposable k-fold covering must cover some point $\gtrsim 2^{k/3}$ times. □

By the way, to prove the claim that m balls can cut $\mathbb{R}^3$ in at most m^3+1 regions, we use induction. Adding a new ball creates regions if we have intersections with other balls – this is at most the number of regions on the sphere cut by m circles, and we can make arguments from there.

6.3 The general local lemma and proof

As previously stated, we have a bunch of bad events $A_1, \cdots, A_n$ that we are trying to avoid. We occasionally have that A_i is mutually independent of all other A_j except for some set $N(i)$ for each i. (Notice that $N(i)$ does not include i.)

Definition 6.8
The **dependency graph** is constructed by having a vertex for each bad event i and joining i to the set $N(i)$ (of dependencies to i).

This graph is sometimes directed, but in almost all applications, it's sufficient to make it undirected. For example, if we have a hypergraph coloring problem and we're coloring each vertex at random, the dependency connects edges that have nonzero vertex intersection. In such a setup, we have the following:

Theorem 6.9 (Local lemma, symmetric form)
If every node in the dependency graph has degree at most d, and every event has probability at most p, then there is a positive probability that no bad events occur as long as

$$ep(d+1) \le 1.$$

We're going to be proving a more general form of this:

Theorem 6.10 (Local lemma, general form)
If we have real numbers $x_1, \cdots, x_n \in [0,1)$ such that for all i,

$$\Pr(A_i) \le x_i \prod_{j \in N(i)} (1 - x_j),$$

then with probability at least $\prod_{i=1}^n (1-x_i)$, no event A_i occurs.

Here, note that x_i is not the probability of A_i: it's something larger than x_i which may be weighted down by the other terms.

By the way, notice that if we can find values x_i to plug in, we can get the symmetric case from the general case:

Deducing the symmetric case. Set all $x_i = \frac{1}{d+1} < 1$. Then notice that

$$x_i \prod_{j \in N(i)} (1 - x_j) = \frac{1}{d+1}\left(1 - \frac{1}{d+1}\right)^{|N(i)|} \ge \frac{1}{d+1}\left(1 - \frac{1}{d+1}\right)^{d} > \frac{1}{(d+1)e},$$

and if we have the hypothesis of the symmetric case of the lemma, then

$$x_i \prod_{j \in N(i)} (1 - x_j) \ge p \ge \Pr(A_i),$$

and therefore the general case's hypothesis must hold as well. □

Here's another way to specialize the general form:

Corollary 6.11

If all events have probability $\Pr(A_i) < \frac{1}{2}$, and

$$\sum_{j \in N(i)} \Pr(A_j) \le \frac{1}{4}$$

for all i, then there is a positive probability that no A_i holds.

Proof. Set $x_i = 2\Pr(A_i)$; this is always less than 1, and the product

$$x_i \prod_{j \in N(i)} (1 - x_j) = 2\Pr(A_i) \prod_{j \in N(i)} (1 - x_j) \ge \Pr(A_i),$$

so all probabilities are smaller than their corresponding products, and we can use the theorem in its general form. □

We'll now present the original proof of the local lemma from its publication – it uses induction.

Proof. Say we have n events. Let S be a subset of $[n]$ which indexes our bad events: we'll induct on $|S|$. The induction hypothesis is the following:

Proposition 6.12

If we have an event $i \notin S$, then

$$\Pr\left(A_i \mid \bigwedge_{j \in S} \overline{A_j}\right) \le x_i.$$

Basically, this is the probability A_i occurs, conditioned on the fact that none of the events indexed by S occur.

If we prove this, then by Bayes' formula, the probability that none of the events occur is at least

$$\Pr(\overline{A_1}) \cdot \Pr(\overline{A_2} \mid \overline{A_1}) \cdot \cdots \cdot \Pr(\overline{A_n} \mid \overline{A_1 A_2 A_{n-1}}) \ge (1 - x_1)(1 - x_2) \cdots (1 - x_n),$$

where the last inequality comes from the inductive hypothesis. So this would imply the local lemma.

First of all, this proposition is easy to show when S is empty, since the probability that A_i occurs is

$$x_i \prod_{j \in N(i)} (1 - x_j) \le x_i.$$

Now for the inductive step, we know i has some neighbors in S: call this set $S_1 = S \cap N(i)$, and call everything else $S_2 = S \setminus S_1$.

Let's understand the conditional probability $\Pr\left(A_i \mid \bigwedge_{j \in S} \overline{A_j}\right)$. We can separate this into contributions from S_1 and contributions from S_2 using Bayes' rule:

$$= \frac{\Pr\left(A_i \wedge \bigwedge_{j \in S_1} \overline{A_j} \mid \bigwedge_{j \in S_2} \overline{A_j}\right)}{\Pr\left(\bigwedge_{j \in S_1} \overline{A_j} \mid \bigwedge_{j \in S_2} \overline{A_j}\right)}$$

First, we can upper-bound the numerator by forgetting the dependencies that are hard to control: this is at most

$$\Pr\left(A_i \mid \bigwedge_{j \in S_2} \overline{A_j}\right)$$

since we're just removing some conditions on what needs to happen. But now since A_i is mutually independent from all of its neighbors, this is just $\Pr(A_i) \le x_i \prod_{j \in N(i)}(1 - x_j)$ by the assumed conditions.

Meanwhile, we can also lower bound the denominator. Label the elements of $S_1 = \{j_1, \cdots, j_r\}$; as before, we can write the denominator as a product of conditional probabilities

$$\Pr\left(\overline{A_{j_1}} \mid \bigwedge_{j \in S_2} \overline{A_j}\right) \Pr\left(\overline{A_{j_2}} \mid \overline{A_{j_1}} \wedge \bigwedge_{j \in S_2} \overline{A_j}\right) \cdots \Pr\left(\overline{A_{j_r}} \mid \overline{A_{j_1}} \wedge \cdots \wedge \overline{A_{j_{r-1}}} \bigwedge_{j \in S_2} \overline{A_j}\right)$$

S_2 is a smaller set than S (or else there are no dependencies in S at all to A_i, in which case this is easy). So by the induction hypothesis, each event $\overline{A_{j_k}}$ occurs with probability at most x_i, so this is at least

$$(1 - x_{j_1})(1 - x_{j_2}) \cdots (1 - x_{j_r}) \ge \prod_{j \in N(i)} (1 - x_j)$$

(since the LHS product is some subset of the RHS product). Putting the numerator and denominator together, we're done! The probability that A_i occurs given that none of the events in S occurs is at least

$$\frac{x_i \prod_{j \in N(i)}(1 - x_j)}{\prod_{j \in N(i)}(1 - x_j)} = x_i,$$

as desired, completing the inductive step. □

6.4 The Moser-Tardos algorithm

We now know that under certain circumstances, it is possible to avoid all of our bad events. But is there a nice way to algorithmically determine an example of this? It's not even obvious how to do a randomized algorithm, since the chance of success (avoiding all bad events) is generally so small.

Consider a **random variable model** to make our problem less abstract: we have a collection of independent random variables, where each event A_i depends only on some subset of variables. We have a dependency graph, where $A \sim B$ if A and B share common variables. Our goal is to find a way to flip the independent variables so that the A_is all do not occur.

It would seem like this algorithm could be very sophisticated, but here's one that isn't!

Algorithm 6.13

First, initialize all our random variables with some values. While some bad event A_i occurs (pick one arbitrarily), resample all of the variables that A_i depends on, since A_i only depends on some finite set of variables. Maybe this induces some other bad events: we just keep running the while loop.

We might be worried that this algorithm might never terminate, or on average, maybe it terminates after an exponential number of iterations. It turns out this never happens:

Theorem 6.14 (Moser–Tardos)

If the Lovász local lemma conditions hold, then the algorithm is expected to resample each A_i at most $\frac{x_i}{1-x_i}$ times. In particular, the expected number of iterations of the while loop is at most

$$E = \sum_{i=1}^{n} \frac{x_i}{1 - x_i}.$$

Note that this algorithm is agnostic of the x_is: it doesn't change based the parameters that we're using.

There are two key concepts in this proof that are needed. We need an **execution log**, which is a list of the resampled events at each step (so we add on 1 event per run through the while loop).

We also need a **witness tree**, which is a finite rooted tree labeled by events such that the children of i are distinct from each other and are a subset of $N(i) \cup \{i\}$ (the neighbors of i and itself).

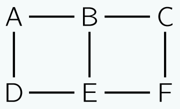
Why is this called a witness tree? We're going to use a prefix of our execution log to make a tree (though the lumberjack may say to go the other way).

Basically, given a prefix of the execution log, read the log **right to left**. The rightmost event is the root of the tree, and for each subsequent event A, if some node in the tree overlaps with A (in other words, it lies in $N(A) \cap \{A\}$, then add A as a child of a deepest such node (arbitrarily) If nothing is dependent with A, just discard it.

The key idea here is that this witness tree tracks some of the possible ways we could have progressed in our algorithm from bottom to top.

Example 6.15

Let's say we have the dependency graph



and we have an execution log with prefix $FBDEADBFDC$.

When we make our rooted tree, C is the root. D is next-to-last, but it couldn't have caused C to be sampled since C and D are not overlapping, so it is discarded. F is adjacent with C, though, so F is now a child of C. This represents the statement "C could have been sampled as a result of F."

Now B must be another child of C for the same reason the next D is discarded, A is a child of B, E is a child of B or F (arbitrarily choose F), and so on.

Events at the same depth are independent, because if they weren't, one would be forced to be a child of the other when it is inserted.

Lemma 6.16

For a given log, all prefixes produce distinct trees.

This is not too hard to show. Given any tree with root A, the number of times A appears in it is just the number of As up to that point in the prefix of the log. So all such trees must have a different number of As, and all trees with root A are different from trees of root B, and so on.

Lemma 6.17

For a given witness tree T, the probability that T appears as a witness tree of some prefix is at most

$$\prod_{v \in T} \Pr(A_{[v]}),$$

where $A_{[v]}$ is the event corresponding to the node v in the witness tree.

To show this, we'll introduce some randomness.

Definition 6.18

Let a **simulation** of T to be the following process: visit all nodes of T in reverse depth (the order doesn't matter for each given level). Then resample all the variables at each visited node. The simulation **succeeds** if all bad events encountered do occur.

Proof of lemma. How can we ensure that there's a level of consistency between the tree and simulation? We're going to do a **coupling** of the two, where we basically have two processes $X, Y \to (X, Y)$ that act in parallel. This way, if we look at just X or Y alone, it looks identical to an initial distribution, but X and Y aren't necessarily independent.

Our goal is to do this in such a way that **the simulation always succeeds if T appears as a witness tree**. Then, the probability that T appears must be at most the probability that the simulation succeeds, which gives us a bound we want.

What's the probability the simulation succeeds? This is easy: each step in the tree resampling is independent, so this is just

$$\prod_{v \in T} \Pr(A_{[v]}).$$

To make this useful, we need to find a common source of randomness. Let's say we have access to an infinite list of realizations of each random variable, and resampling is just taking the next realization of the list. The key idea is to use the same list to run both processes. We claim that if T appears as a witness tree, the simulation must succeed.

Why? During the execution step, there's an initialization of all variables. Since T appears as a witness tree, we know that looking at our execution log, each of the bottom nodes must have been "bad events" that were logged; since this is coupled directly to our witness tree, the same bad event must have occurred there. Now keep propagating up the tree; we'll notice that each event must have occurred because the prior resampling did not work. □

One last tool we're going to use in our proof is the **multitype Galton-Watson branching process.** Generate a tree in the following way: specify a root labeled by some event A_i, and for each possible child $A_j \neq A_i$, keep it with probability x_j (this is not the probablity of A_j - it is the corresponding real number from the statement of Lovász). Repeat the same process for each node; each child survives with some probability x_j or dies with some probability $1 - x_j$.

When this terminates, we get some finite or infinite tree.

Lemma 6.19

Then T is yielded with probability

$$p_T = \frac{1 - x_i}{x_i} \prod_{v \in T} x'_{[v]},$$

where

$$x'_{[v]} = x_{[v]} \prod_{j} (1 - x_{[j]}).$$

Basically, we're just multiplying the probabilities of each vertex working out, and the normalizing factor comes from the fact that our root is already formed for sure. We can think of this as a weighing of our trees in an information-theoretic manner.

It's time for us to prove the Moser-Tardos theorem:

Proof of Theorem 6.14. The number of times an event A_i is resampled is the number of times a witness tree is generated, rooted at A_i. We expect this to be

$$\sum_{T \text{ rooted at } A_i} \Pr(T \text{ appears as a witness tree}).$$

From our construction by coupling, this is bounded by the expression

$$\sum_{T \text{ root at } A_i} \prod_{v \in T} \Pr(A_{[v]}).$$

Now recall that the x_is in the Lovász local lemma are at least their corresponding probabilities:

$$\leq \sum_{T \text{ root at } A_i} \prod_{v \in T} x'_{[v]}.$$

Now by the lemma above, this is just

$$= \frac{x_i}{1 - x_i} \sum_{T \text{ rooted at } A_i} P_T.$$

But summing over P_T gives the probability that some certain trees are our final result, so that sum can be at most 1. Therefore

$$\mathbb{E}[A_i \text{ is sampled}] \leq \frac{x_i}{1 - x_i},$$

and summing over all events A_i gives the desired result. □

6.5 A computationally hard example

Unfortunately, there are instances where we'd like to use the local lemma, but an algorithm may not quickly yield an answer. Here's a simple example of that:

Example 6.20

Let $q = 2^k$ be an integer, and let f be a bijection from $[q] \to [q]$. Let y ben an element of $[q]$; then let A_i be the "bad event" that $f(x)$ and y disagree on the ith bit, where we choose x uniformly on the domain $[q]$. (There are k events $A_1, \cdots, A_k$, since $q = 2^k$ has k bits.)

Notice that all events A_i are mutually independent, since the bits behave independently, by the "local lemma" (a very trivial use of it, aka just independent events), there exists an x such that $f(x) = y$.

But algorithmically, can we always find this efficiently? As a foundation of cryptography, the answer is (believed to be) no. A concrete example is the function $f : \mathbb{F}_q \to \mathbb{F}_q$ sending $f(x) = g^x$ for some generator g (except 0 at $x = 0$). Then we just have the **discrete logarithm** problem, which is believed to be computationally difficult.

6.6 Back to independent sets

We know that in a graph with maximum degree Δ, we have an independent set of size at least $\frac{|V|}{\Delta+1}$. (Take a vertex, include it and throw away all of its neighbors, and repeat.) But we can achieve something better with the local lemma:

Theorem 6.21

Let G be a graph with maximum degree Δ, and let the vertex set $V = V_1 \cup V_2 \cup \cdots \cup V_r$ be partitioned into sets of size $|V_i| \geq 2e\Delta$. Then there exists an independent set with one vertex in each V_i.

Then the size of this set is similar to what we have naively (we lose a constant), but we have much more control over what the set looks like!

Proof. First, we may assume all $|V_i| = k = \lceil 2e\Delta \rceil$ for simplicity: we can always toss away extra vertices that we just decide not to use.

For each V_i, pick a uniformly random v_i independently from all the others: our goal is to show that with positive probability, we get an independent set.

What are our bad events? We don't want our vertices to be adjacent, and there's a few ways we can try to approach this.

Attempt 1: Let's let A_{ij} be the events that v_i and v_j are connected by an edge. Then in the worst case, $\Pr(A_{ij}) \le \frac{\Delta}{k}$, since any given vertex in V_i can only be connected to Δ of the things in V_j.

But how large is our dependency set? The events A_{ij} and A_{rs} are dependent whenever some vertex in $V_i \cup V_j$ is adjacent to some vertex in $V_r \cup V_s$. The maximum degree of the dependency graph is therefore $2\Delta k$: each of the $2k$ vertices in $V_i \cup V_j$ has at most Δ neighbors. So the condition for local lemma requires

$$\frac{\Delta}{k} \cdot 2\Delta k$$

to be approximately constant, but this is too large and this method fails.

Attempt 2: So an alternative way to make bad events is to let A_e be the event that both endpoints of e are chosen, where $e \in E(G)$ is an edge between two different V_i, V_j. We know that

$$\Pr(A_e) \le \frac{1}{k^2},$$

since there are k vertices in each of V_i and V_j, and what do we know about the dependency graph? Again, A_e and A_f can be dependent if and only if one of the edges is in $V_i \cup V_j$: thus the maximum degree is $2\Delta k$. Thus by an application of the local lemma, as long as we have

$$\frac{1}{k^2}(1 + 2\Delta k) < \frac{1}{e},$$

we are good. □

So in many applications, it's important to pick the right events to think about.

6.7 Graphs containing large cycles

Theorem 6.22

For any k, there exists a d such that every d-regular directed graph has a directed cycle of length divisible by k.

Here, d-regular means that each vertex has d outgoing and d incoming edges, and a directed cycle is a cycle with all arrows pointing correctly. Also, note that if we have a (connected) $2d$-regular undirected graph, then we can find an Eulerian tour (since all degrees are even). So we can direct an undirected graph. If we have odd degree, we can just add a vertex (left as an exercise). Either way, this means that we can prove the undirected version of this theorem as well! We'll actually prove something stronger:

Theorem 6.23

Every directed graph with minimum out-degree δ and maximum in-degree Δ contains a cycle of length divisible by a constant k if

$$k \leq \frac{\delta}{1+\log(1+\delta\Delta)}.$$

Proof. We'll simplify the problem: we can delete some edges and assume that all outdegrees are exactly δ. Assign every element a uniform random element of $\mathbb{Z}/k\mathbb{Z}$: only look at those vertices where the label increases by 1 mod k along the arrow. In other words, we can split up the vertices into k groups, corresponding to the residues. Then if there exists a directed cycle in this new graph, it must have length divisible by k.

What can go wrong? It's bad to go along the cycle and get a dead end, so our goal is for every vertex to have at least 1 outgoing edge that remains. Then we can just keep going, and we must eventually terminate.

So let A_v be the event that v doesn't have any outgoing edges. The probability this happens is

$$\left(\frac{k-1}{k}\right)^{\Delta},$$

since all δ of the outgoing edges must not work. But the dependencies are a bit more subtle: when do we put an edge between A_v and A_w? Define $N^+(v)$ to be the out-neighbors of v and $N^+(w)$ be the out-neighbors of w: then we need to make sure

$$(v \cup N^+(v)) \cup (w \cup (N^+(w))) \neq \varnothing,$$

since the event A_v only depends on v and its out-neighbors. But we can do a little bit better:

Lemma 6.24

A_u is independent of all A_w such that

$$N^+(u) \cap (N^+(w) \cup \{w\}) = \varnothing.$$

Proof of lemma. To show this, notice that given a vertex u with some outgoing edges, if w points to u, A_u is independent of A_w because A_u has the same probability even when we fix the value of A_w. Specifically, if w points to u but not any of its out-neighbors, we can just pick whether w has outgoing edges first: this can't affect u. □

So now we can calculate the degrees of the dependency graph: for a fixed u, the number of w that are adjacent to u in our dependency digraph is at most $\delta\Delta$. Thus the conditions of the local lemma are satisfied as long as

$$ep(\delta\Delta+1) = e\left(\frac{k-1}{k}\right)^{\delta}(\delta\Delta+1) \leq e^{1-\delta/k}(\delta\Delta+1) \leq 1,$$

and rearranging gives the result. □

Fact 6.25

We've often been saying "the dependency graph," but the way we should probably think of it as follows: specify a graph, and then say that our collection of events is consistent with the graph we've created. Basically, there are multiple options for our dependency graph for a single setup.

6.8 Bounds on the linear arboricity conjecture

Lemma 6.26

Let $k \leq d^{0.9}$. Then every d-regular directed graph can be vertex-colored (in any way) using k colors so that every vertex has $\frac{d}{k} + O\left(\sqrt{\frac{d \log d}{k}}\right)$ in-neighbors of each color (and also simultaneously the same number of out-neighbors).

So looking at our graph from every vertex, the graph looks fairly equidistributed.

Proof. Color every vertex uniformly at random. Bad events are of the form "a vertex has the wrong number of in-neighbors or out-neighbors of a color": denote $A^+_{v,c}$ to be the event "v doesn't have the correct number of out-neighbors of some color c," and analogously, let $A^-_{v,c}$ be the analogous bad event for in-neighbors.

The probability that each $A_{v,c}$ occurs is the probability we deviate too much from the mean:

$$\Pr\left(\left|\text{Binomial}\left(d, \frac{1}{k}\right) - \frac{d}{k}\right| > C\sqrt{\frac{d \log d}{k}}\right)$$

We do have to be careful about how we use our Chernoff bound, but we can pick C such that

$$\Pr(A_{v,c}) < \frac{1}{100d^3}$$

(to be safe in our calculations later). Now, when can $A^+_{v,c}$ and $A^+_{w,c'}$ be dependent? This only happens if they're reasonably close to each other in the graph:

$$(v \cup N^+(v)) \cap (w \cup N^+(w)) \neq \varnothing,$$

and the maximum degree that can occur here is $(2d)^2k$. (Note that this is saying that if we condition on any event of the non-neighbors of v, the probability is still independent of those conditions.) Now by the Lovász Local Lemma, we can check that $ep(d+1) \leq 1$ and we're done. □

Now let's move on to an open problem:

Definition 6.27

A **linear forest** is a disjoint union of paths.

Conjecture 6.28 (Linear arboricity)

The edge-set of every graph with maximum degree Δ can be decomposed into $\lceil \frac{\Delta+1}{2} \rceil$ linear forests.

We can't really do any better than this for any given graph: any path only contributes 2 to the degree, so we need at least $\frac{\Delta}{2}$ forests. Notice that every graph of maximal degree Δ is a subgraph of a Δ-regular graph, so it suffices to consider Δ-regular graphs.

For a long time, there was a constant factor gap between the best-known bounds, until Alon showed in 1988 that $\frac{\Delta}{2} + o(\Delta)$ linear forests will work. Alon and Spencer found in 1992 that we can have $\frac{\Delta}{2} + \tilde{O}(n^{2/3})$ (which means there are some other log terms that are neglected). Finally, last year, Ferber, Fox, and Jain proved $\frac{\Delta}{2} + O(\Delta^{2/3-\alpha})$.

There's also a directed version of this conjecture:

Conjecture 6.29 (Directed linear arboricity or DLAC)
The edge-set of every Δ-regular directed graph can be decomposed into $\Delta + 1$ directed linear forests.

This implies the undirected version: take G to be a $2d$-regular graph, and now use an Eulerian tour to orient and get a d-regular digraph. If G starts as an odd-degree regular graph, we can just modify it a bit.

Also, we need at least $\Delta + 1$ directed linear forests for the same reason: every vertex can only contribute one to each vertex's indegree and outdegree, but we can't get 1 on everything (because of the endpoints of our paths).

Let's first outline some of the key steps before jumping into a somewhat weaker bound for DLAC. First, we show that the result is true if the girth of the graph is at least $8e\Delta$. The idea there is that (using Hall's theorem) we can decompose this graph into cycles. This gives Δ subgraphs, each of which is a disjoint union of cycles. To turn this into a covering of linear forests, we can just cut out an edge from each cycle, and our goal is to make sure the remaining edges form a directed linear forest as well.

After we prove this version with large girth, we can divide into subgraphs with large girth. That's also something we did: we found how to produce cycles with length divisible by k, and if $k \geq 8e\Delta$, we can get cycles of long length.

In our proof, we will use a few graph theory results:

Lemma 6.30 (A consequence of Hall's theorem)
A d-regular bipartite graph has a perfect matching.

Notice that removing a perfect matching from a d-regular bipartite graph gives a $(d-1)$-regular bipartite graph.

Corollary 6.31 (Konig's theorem)
Every d-regular bipartite graph can be decomposed into d matchings.

We can then use Konig's theorem to prove the following lemma:

Lemma 6.32
The edge-set of every Δ-regular digraph can be decomposed into Δ 1-regular spanning subgraphs.

For a directed graph, a 1-regular spanning subgraph is a collection of directed cycles using all vertices exactly once.

Proof. Construct two copies of the original vertex set. For each edge going from v_i to v_j, draw an edge from v_i in the first copy to v_j in the second copy: this gives a Δ-regular bipartite graph. Applying Konig, we get Δ perfect matchings: collapse each matching back to the original graph, and it becomes a collection of 1-regular directed graphs (think of v_i going to v_j as a permutation of the vertices), as desired. □

Lemma 6.33 (An easy bound on the DLAC)
The edgeset of every directed Δ-regular graph can be decomposed into at most 2Δ directed linear forests.

Proof. Use the lemma above to get Δ 1-regular spanning subgraphs: for each of these, split the cycles into two paths arbitrarily, and each 1-regular spanning subgraph becomes 2 directed linear forests. □

Theorem 6.34 (DLAC for large directed girth)
The directed linear arboricity conjecture is true if the directed girth is at least $8e\Delta$.

Proof. We can decompose the edgeset into Δ 1-regular spanning subgraphs, $F_1, F_2, \cdots, F_\Delta$, and now each F_i is a disjoint union of cycles. Our goal is to show that we can break up the cycles in a nice way.

Consider the **line graph** of G, $L(G)$, where the vertices are edges of G and $e_1 \to e_2$ is drawn if e_1 and e_2 are incident (share a vertex). Our goal is to select an independent set from this line graph that hits every cycle in the F_is above, because this would allow us to get $\Delta + 1$ directed linear forests.

By the first lemma, we know there exists a matching $M \subset E(G)$ that contains an edge from each cycle, since we have partitioned our graph into cycles have length at most $8e\Delta$, and each vertex in our line graph has degree at most 4Δ. Take $F_1 \setminus M, F_2 \setminus M, \cdots$: we've now broken up the cycles, so each of those is a linear forest. Thus M plus these gives the $\Delta + 1$ directed linear forests, as we want. □

Now we're ready to prove the main result of this section:

Theorem 6.35 (A better bound for DLAC)
Every Δ-regular directed graph can be decomposed into at most $\Delta + \tilde{O}(\Delta^{3/4})$ (possibly contains poly-log factors) linear forests.

Our goal is to produce a lot of cycles with long length. One thing we can do, as last time, is to label our vertices mod k, and make sure all cycles go from vertices i to $i+1$.

Proof. Pick a prime k in the range $\left[10\sqrt{\Delta}, 20\sqrt{\Delta}\right]$ (this exists). By the second lemma, there exists a coloring of the vertex set by elements of $\mathbb{Z}/k\mathbb{Z}$, so that every vertex has $\frac{\Delta}{k} + \tilde{O}\left(\sqrt{\frac{\Delta}{k}}\right)$ neighbors of each element.

For each i, let D_i be the subgraph of edges whose color increases by i mod k from the start to end point. We know that Δ_i, the maximum degree of D_i, is at most $\frac{\Delta}{k} + \tilde{O}\left(\sqrt{\frac{\Delta}{k}}\right)$. For all $i \neq 0$, D_i has girth at least $k \geq 8e\Delta_i$ (this is where we need $k \gtrsim \sqrt{\Delta}$), so the DLAC for large girth tells us that we can decompose each D_i for nonzero i using $\Delta_i + 1$ linear forests. For D_0, use the easy bound of twice the degree from Lemma 6.33. This means the total number of linear forests is

$$\leq \left(\frac{\Delta}{k} + \tilde{O}\left(\sqrt{\frac{\Delta}{k}}\right)\right)(k+1) \leq \Delta + \tilde{O}(\Delta^{3/4}),$$

as desired. □

Notice that we needed k to be prime to ensure that all cycles go through all k colors.

How can we improve this? Notice that we can decompose K_4 into Hamiltonian paths: in general, we can decompose any K_{2n}. If we have $2n$ colors in our proof above, consider all the edges between the color groups: decompose into Δ_1 matchings for each "connection" between color groups, and this means we can decompose into Δ_1 paths, minus some edges. Applying the easy bound to the edges within the color groups again, this gives k Hamiltonian paths, and our bound is now

$$\leq \left(\frac{\Delta}{k} + \tilde{O}\left(\sqrt{\frac{\Delta}{k}}\right)\right)k + \frac{2\Delta}{k} + \tilde{O}\left(\sqrt{\frac{\Delta}{k}}\right)$$

and optimizing for the correct value of k, we minimize this at

$$\leq \Delta + \tilde{O}(n^{2/3}).$$

This was nice because we can pick a smaller value of k, and we don't have the girth requirement anymore!

6.9 The lopsided local lemma

Remember in the proof of the local lemma, we made a bound of the form

$$\text{numerator} \le \Pr\left(A_i \mid \bigwedge_{j \in S_2} \overline{A_j}\right),$$

where S_2 is elements outside of $N(i)$ [not connected]. By the definition of the dependency graph, this is just **equal to** $\Pr(A_i) \le x_i \prod_{j \in N(i)} (1 - x_i)$. This was the only place we used the independence hypothesis!

What if we instead just assumed an inequality in that equality? In other words, what if we always just had positive dependence, so bad events actually make each other less likely to occur?

Here's the formal setup. Let's say we have some bad events $A_1, \cdots, A_n$. The **negative dependence digraph** is one where (if $N(i)$ is the outneighbors of i), we have

$$\Pr\left(A_i \mid \bigwedge_{j \in S} \overline{A_j}\right) \le \Pr(A_i) \forall i, S \subset N(i)^C.$$

So this graph records potential negative dependence: $N(i)$ notes the bad events that can potentially be worrying for us (since the events are more likely to occur separately), and everything else either does nothing or actually helps us (because it means the events have positive dependence). This was called the **lopsidependency graph**.

Theorem 6.36 (Lopsided Lovász Local Lemma)
If there exist real numbers $x_1, \cdots, x_n \in [0, 1)$ such that

$$\Pr(A_i) \le x_i \prod_{j \in N(i)} (1 - x_j) \quad \text{for all } i \in [n],$$

then with probability at least $\prod_{i=1}^n (1 - x_i)$, no event A_i occurs.

The proof is exactly the same. Just change the boxes equals to an inequality. In fact, we can replace the $\le \Pr(A_i)$ condition with $\le x_i \prod_{j \in N(i)} (1 - x_j)$.

There's also a symmetric version, which is easier to use: if the negative dependency (di)graph has all neighborhoods size with size at most d, and $ep(d + 1) \le 1$, then with positive probability, no bad event occurs.

We showed using LLL that every k-uniform hypergraph is 2-colorable if every edge intersects at most $\frac{2^{k-1}}{e} - 1$ other edges: color all the vertices uniformly at random. What if we refine that argument a bit?

Proposition 6.37
Every k-uniform hypergraph is 2-colorable if every edge intersects at most $\frac{2^k}{e} - 2$ other edges.

Solution. For each edge f, we defined the bad event A_f if f was monochromatic, but now, let's let $A_{f,c}$ be the event that all vertices in f are colored with color c. It is clear that $\Pr(A_{f,c}) = 2^{-k}$.

Let our colors be 0 and 1. Consider the graph where $A_{f,c}$ and $A_{f',1-c}$ are adjacent if f and f' intersect. Edges of our dependency graph then come from two intersecting edges with opposite colors: we claim this is a negative dependency graph. The intuition is that compared to the vertex overlap graph, this one tells us a little more: if we have an edge f, and we want to know the probability of f being colored all red, it's more likely if we know f' is not colored all blue.

So if the degree of the negative dependency graph is at most $d+1$, then $ep(d+2) \leq 1$ as long as $d \leq \frac{2^k}{e} - 2$, and applying the Lopsided Lovász Local Lemma yields the result. □

6.10 Latin squares

Let's now do an example where the lopsidedness matters:

Definition 6.38

A **Latin square** of order n is an $n \times n$ square filled with n symbols (usually 1 through n), such that every symbol appears exactly once in each row and column.

A **transversal** of an $n \times n$ array is a set of n entries with one per row and one per column. A **Latin transversal** has distinct entries in the Latin square.

Conjecture 6.39

All odd order Latin squares have a Latin transversal.

This is still open, so we're not going to prove it. Instead, we'll show the following weakening of that result:

Theorem 6.40 (Erdős, Spencer)

Every $n \times n$ array where every entry appears at most $\frac{n}{4e}$ times has a Latin transversal.

There's an open problem that every odd-n Latin square has a Latin transversal: this is a looser restriction of that. (It's not necessarily true for even n, but the conjecture is that we can always find $(n-1)$ different entries in that case in a "rook placement.")

Remark (Historical note)**.** *These objects are called Latin squares, because Euler started playing with them and wrote a paper where they had Latin entries instead of numbers.*

It's not really clear what the bad events can be here, but the point is that different permutations seem to involve every row and every column, so there's not that many disjoint supported variables. That makes it hard to apply the vanilla LLL. But some of the dependencies only help us, so we can apply the lopsided version.

Proof. Pick a transversal uniformly at random: this is equivalent to picking one of $n!$ permutations uniformly at random. Then we can have our bad events be of the form A_{ijkl}, where (i,j) and (k,l) are both picked, and they have the same entry written in them. (Note that the second part of this condition is not random: it is already determined by our initial array.) If we avoid all such bad events, we get a Latin transversal.

The probability of $0A_{ijkl}$ is 0 if the two squares (i,j) and (k,l) are in the same row or column, since we can't pick them both in our Latin transversal. For all others, the probability is $\frac{1}{n(n-1)}$, since there are $(n-2)!$ permutations that go through these two points.

If we try to do a dependency graph problem like the Lovász Local Lemma, we run into a problem: most bad events are dependent on each other, but what about the negative dependency graph?

Let G be the graph (V,E), where

- V consists of unordered pairs $\{(i,j),(k,l)\}$ that have the same entry and aren't in the same row or column.
- $\{(i,j),(k,l)\}$ and $\{(i',j'),(k',l')\}$ are connected by an edge if and only if there are any rows or columns that are shared: $\{i,k\} \cap \{i',k'\} \neq \varnothing$ or $\{j,l\} \cap \{j',l'\} \neq \varnothing$.

If we can show this is a negative dependency graph, then we can finish in the following way: given any pair of vertices, another pair is in the negative dependency graph if we pick one of the $(4n-4)$ squares in the same row or column, and then pick one of the $\frac{n}{4e} - 1$ other squares with the same entry. Then the maximum degree is at most $\frac{n(n-1)}{e} - 1$, and therefore $ep(d+1) \leq 1$, meaning we can apply the Lopsided Lovász Local Lemma.

So to show that we do have a negative dependency graph G, fix some i, j, k, l. Let J be a subset of the nonneighbors of the bad event $\{(i,j),(k,l)\}$. Our goal is to show that

$$\Pr\left(A_{ijkl} \mid \bigwedge_{i'j'k'l' \in J} \overline{A_{i'j'k'l'}}\right) \leq \Pr(A_{ijkl}).$$

Without loss of generality, we can assume $\{(i,j),(k,l)\} = (1,1,2,2)$ by permuting rows and columns. We need to show that the probability that the transversal goes through $(1,1)$ and $(2,2)$ is at least as large as it is conditioned on not going through any of the other events $A_{i'j'k'l'}$.

So we're not conditioning on any information through the first two columns or rows. How many transversals are there satisfying these conditions? Set $\mathcal{S}_{r,s}$ to be the number of transversals through $(1,r),(2,s)$ that avoid all bad events in J. Note that the $\mathcal{S}_{r,s}$s over distinct $r, s \in [n]$ partition $\bigwedge_{i'j'k'l'} \overline{A_{i'j'k'l'}}$ into $n(n-1)$ subsets. Our goal is to show that $|\mathcal{S}_{1,2}| \leq |\mathcal{S}_{r,s}|$ for all r, s: if we show this, it would follow that

$$\Pr\left(A_{ijkl} \mid \bigwedge_{i'j'k'l' \in J} \overline{A_{i'j'k'l'}}\right) = \frac{|\mathcal{S}_{1,2}|}{\sum_{r \neq s} |\mathcal{S}_{r,s}|} \leq \frac{1}{n(n-1)}.$$

So we're saying that not having any bad events in rows 3 through n only helps us in the $1,2$ case. To do this, we'll construct an injection: $\mathcal{S}_{1,2} \to \mathcal{S}_{r,s}$ by swapping the 1st and rth row, as well as the 2nd and sth row. This can't cause any bad events to occur, because all potential bad events occur in the bottom $(n-2)$ by $(n-2)$ square! This finishes the proof. □

Notice that the hardest part of the proof is finding the right bad events and dependency graph to work with: the rest is fairly straightforward.

7 Correlation and Janson's inequalities

7.1 The Harris-FKG inequality

We're often interested in understanding probabilities related to a graph $G(n, p)$: for example, we might want the probability of some event occurring, such as the probability of having a triangle (which we don't know how to compute yet).

But we can also ask questions like "does this probability increase or decrease if we condition on some other event?" For example, if we have a Hamiltonian cycle, does this increase or decrease our chance of having a triangle?

These don't seem all that relevant to each other, but let's look more closely. Having a Hamiltonian cycle is indicative of having more edges in the graph, and thus the probability of having a triangle should go up. This isn't a proof, but the basic idea here is correct. Similarly, if we know that our graph is planar, we should expect fewer edges, so we should expect a smaller probability of having a triangle. We can rigorize this!

Let's say we have n independent Bernoulli variables $x_1, \cdots, x_n$ (for most applications, the probabilities are the same). An event A is **increasing** if changing some 0s to 1s never destroys the events. In other words, if $x \le x'$ pointwise, and $x \in A$, then $x' \in A$ as well. Notice that we can view A as a subset of $\{0, 1\}^n$: this is an **up-set**, because it's closed upwards. Similarly, decreasing events are **down-sets**.

Example 7.1

Let's say we have a graph $G(N, p)$, and we have $n = \binom{N}{2}$ random variables for the edges. Having a Hamiltonian cycle and being connected are increasing, while having average degree at least 5, planarity, and being 4-colorable are decreasing.

Proposition 7.2 (Harris inequality)

If A and B are increasing events of independent boolean variables, then

$$\Pr(A \cap B) \ge \Pr(A)\Pr(B).$$

We also have that $\Pr(A|B) \ge \Pr(A)$.

(Both of these imply that A and B are **positively correlated**.) More generally, we can let each Ω_i be a probability space that is linearly ordered (for example, $\{0, 1\}$ in the case above).

Definition 7.3

A function $f(x_1, \cdots, x_n)$ is **monotone increasing** if given two vectors $x \le x'$ (pointwise in every coordinate), $f(x) \le f(x')$.

Theorem 7.4 (More general version of the Harris inequality)

Let f, g be increasing functions of independent random variables $x_1, \cdots, x_n$. Then $\mathbb{E}[fg] \ge \mathbb{E}[f]\mathbb{E}[g]$.

(This implies the first version by picking f and g to be indicator functions.)

Later generalizations were made by "FKG" (Fortuin, Kastelyn, Ginibre), but we won't discuss FKG inequalities in their full generality. The idea is that we can relax the independence condition, use a distributive lattice, and use a few other conditions.

Proof of the Harris inequality. We will use induction on the number of random variables. If $n = 1$, then we're saying that given two functions f, g that are monotone increasing on one variable,

$$\mathbb{E}[fg] \geq \mathbb{E}[f]\mathbb{E}[g].$$

This is due to Chebyshev (the same guy as earlier in the course): the proof is that picking x, y independently,

$$0 \leq \mathbb{E}\left[(f(x) - f(y))(g(x) - g(y))\right]$$

since $f(x) - f(y)$ and $g(x) - g(y)$ have the same sign, and expanding this out,

$$= 2\mathbb{E}[fg] - 2\mathbb{E}[f]\mathbb{E}[g],$$

implying the result for one variable. Now for the inductive step, let $h = fg$. We'll fix x_1, defining a new function

$$f_1(x) = \mathbb{E}[f \mid x_1 = x] :$$

basically, fix one of the variables in our function f. Likewise, let $g_1(x) = \mathbb{E}[g \mid x_1 = x]$ and similar for h_1. We know that f_1, g_1 are monotone increasing functions on the remaining variables. Note that

$$h_1(x_1) \geq f_1(x_1)g_1(h_1)$$

by the induction hypothesis, so

$$\mathbb{E}[fg] = \mathbb{E}[h] = \mathbb{E}[h_1]$$

(letting x_1 be random as well now), and pointwise, this is

$$\geq \mathbb{E}[f_1 g_1] \geq \mathbb{E}[f_1]\mathbb{E}[g_1] = \mathbb{E}[f]\mathbb{E}[g]$$

by the base case, since f_1 and g_1 are one-variable functions. □

Corollary 7.5

Decreasing events are also positively correlated:

$$\Pr(A \cap B) \geq \Pr(A) \cap \Pr(B).$$

(Take the complement of a decreasing event to get an increasing event.) Similarly, if one event is increasing and another is decreasing, they are negatively correlated. Finally, if all A_is are increasing or all decreasing, we can say that

$$\Pr(A_1 \cdots A_k) \geq \Pr(A_1) \cdots \Pr(A_k).$$

7.2 Applications of correlation

Example 7.6

Let's find the probability that $G(n, p)$ is triangle-free.

There are lots of possible appearances of triangles, and lots of dependent probabilities. But we know that these events are all correlated! In particular, let A_{ijk} be the event that (i, j, k) is not a triangle. A_{ijk} is a decreasing event (on the edges), and this means the probability of having no triangles at all is (by Harris' inequality)

$$= \Pr\left(\bigwedge_{ijk} \overline{A_{ijk}}\right) \ge \prod_{ijk} \Pr(\overline{A_{ijk}}) = (1-p^3)^{\binom{n}{3}}.$$

How close is this to the truth? Taking $p = o(1)$, we can approximate this as

$$\ge e^{-(1+o(1))p^3n^3/6}.$$

(By the way, the probability $G(n,p)$ is triangle free is monotone for p by coupling - having a higher chance of including each edge just makes our chances worse.) One way to obtain an upper bound is by Janson's inequality, which is kind of dual to the Lovász Local Lemma, but we'll see that in the next few sections.

Example 7.7

What is the probability that $G\left(n, \frac{1}{2}\right)$ has maximum degree at least $\frac{n}{2}$?

Let A_v be the event that the degree of v is at most $\frac{n}{2}$: each of these has probability at least $\frac{1}{2}$, so by the Harris inequality, the probability is at least the product of the individual vertex probabilities, which is at least 2^{-n}.

Is this close to the truth - what's the actual value? It turns out the probability is indeed of the form $(c+o(1))^n$. Is $c = \frac{1}{2}$, meaning that our correlation inequality is essentially tight, or is $c > \frac{1}{2}$, which means our lower bound is not very good?

Theorem 7.8 (Riordan-Selby, 2000)

The probability that $G\left(n, \frac{1}{2}\right)$ has maximum degree at least $\frac{n}{2}$ is $(c+o(1))^n$, where $c \approx 0.6102$.

This is very technical, but let's do a "physicist proof" to see where the number comes from.

Solution motivation. Use a continuous model instead. Instead of making each variable Bernoulli, put a standard normal distribution on each (undirected) edge of K_n instead. Now the degree is just the sum of the standard normals of the edges connected to each vertex.

Let $W_v = \sum_{u \ne v} Z_{uv}$ be this sum: We know each event $W_v \le 0$ has a $\frac{1}{2}$ chance of being at most 0. What's the probability all W_vs are less than 0? We know the W_vs are a joint normal distribution, entirely dependent on their covariance matrix. Then the variance of W_v is $n-1$, and the covariance between W_u and W_v is 1 because of the shared edge. So now we can directly compute by using a different model with the same covariance matrix: this is identically distributed as

$$\sqrt{n-2}\,(Z_1', \cdots, Z_n') + Z_0'(1, 1, \cdots, 1)$$

(where each Z_i' is a standard normal distribution independent of the others).

Then what is the probability that this vector has all entries less than or equal to 0? That's an explicit calculation: let Φ be the cumulative distribution of the standard normal distribution: conditioning on Z_0', we find that the desired probability is

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \Phi\left(\frac{-z}{\sqrt{n-2}}\right)^n dz$$

where dz refers to us picking Z_0' first. To evaluate this, we can substitute $z = y\sqrt{n}$ for scaling, this integral becomes

$$\sqrt{\frac{n}{2\pi}} \int_{-\infty}^{\infty} e^{-nf(y)} dy,$$

where $f(y) = \frac{y^2}{2} - \log \Phi\left(y\sqrt{\frac{n}{n-2}}\right)$. We can pretend $f(y) = \frac{y^2}{2} - \log \Phi(y)$ and bound the error, and then look at the asymptotic property of this integral as $n \to \infty$. Well, there's a general principle:

Fact 7.9

If f is a "sufficiently nice" function with a unique minimum at y_0, then

$$\int_{-\infty}^{\infty} e^{-nf(y)} dy = \left(e^{-f(y_0)} + o(1)\right)^n.$$

Basically, as $n \to \infty$, we only get contributions from the smallest $f(y)$. So the rest is just finding the right value of y_0: we can just do this by taking the derivative, and that yields $c \approx 0.6102$ as desired. □

The actual proof is very technical, but this is just a general idea to explain where the constant c potentially comes from. A lot of probability theory now is rigorizing physical intuitions!

7.3 The first Janson inequality: probability of non-existence

The Harris-FKG inequality gives us lower bounds on the probabilities of certain events, but those are not necessarily tight bounds. We'll now start to explore some methods of obtaining upper bounds that are hopefully close to the Harris lower bounds.

Setup 7.10

Pick a random subset R of $[N]$, where each element is chosen independently (usually with probability $\frac{1}{2}$). Refer to $[N]$ as the "ground set." Suppose we have some subsets $S_1, S_2, \cdots, S_k \subseteq [N]$, and A_i be the "bad event" that $S_i \subseteq R$ for all i.

Denote $X = \sum_i 1_{A_i}$ to be the number of A_is that occur. Note that

$$\mu = \mathbb{E}[X] = \sum_i \Pr(A_i),$$

and we have a dependency graph $i \sim j$ if $i \neq j$ and $S_i \cap S_j \neq \varnothing$ (the two underlying subsets overlap). Much like in our earlier covariance calculations, let

$$\Delta = \sum_{(i,j): i \sim j} \Pr(A_i \wedge A_j).$$

(Δ is an upper bound on the variance.) For example, in the current problem we're considering, the S_is are the triangles, and $[N]$ is the set of edges.

Back with the second moment method, we found that if the standard deviation was small relative to the mean, then we have concentration, so we want Δ to generally be small. Janson's inequalities are going to give us better control over our concentration!

Theorem 7.11 (First Janson inequality)

With the definitions in Setup 7.10, the probability that no bad events occur is

$$\Pr(X = 0) \leq e^{-\mu + \frac{\Delta}{2}}.$$

So if Δ is small relative to the mean, then we are essentially upper bounded by $e^{-\mu}$. By the way, this is pretty close to the truth: if all bad events occur with some probability $p = o(1)$, and $\Delta = o(\mu)$, then our lower bound from Harris is

$$\Pr(X = 0) \ge \prod_i \Pr(\overline{A_i}) = e^{-(1+o(1))\mu}$$

by using $(1 + x)^n \approx 1 + nx$. The original proof interpolated the derivative of the exponential generating function, but we'll look at a different one.

Proof by Boppana and Spencer, with a modification by Warnke. Let

$$r_i = \Pr(A_i \mid \overline{A}_1\overline{A}_2 \cdots \overline{A}_{i-1})$$

(so this is conditioned on the probability that none of the previous bad events occur). Then the probability that no bad events occur is a chain of conditional probabilities:

$$\Pr(X = 0) = \Pr(\overline{A_1}) \Pr(\overline{A_2} \mid \overline{A_1}) \cdots \Pr(\overline{A_k} \mid \overline{A_1} \cdots \overline{A_{k-1}}) = (1 - r_1)(1 - r_2) \cdots (1 - r_k) \le e^{-r_1-r_2-\dots-r_k}.$$

It now suffices to show that for all $i \in [k]$, we have

$$r_i \ge \Pr(A_i) - \sum_{j<i,j\sim i} \Pr(A_i \wedge A_j).$$

where the sum only accounts for those A_j with $j < i$ that are dependent on A_i. Then we'd be done, since we have

$$\Pr(X = 0) \le e^{-\sum \Pr(a_i)+\sum_{i\sim j} \Pr(A_i\wedge A_j)} = e^{-\mu+\frac{\Delta}{2}},$$

as desired. Well, the proof of that statement is somewhat reminiscent of the Lovász Local Lemma proof.

Fix i, and split up the events into those that depend and don't depend on i: let D_0 be $\bigwedge_{j<i:j\not\sim i} \overline{A_j}$ and D_1 be $\bigwedge_{j<i:j\sim i} \overline{A_j}$. This partitions all events A_j with $j < i$, and now

$$r_i = \Pr(A_i|D_0D_1) = \frac{\Pr(A_iD_0D_1)}{\Pr(D_0D_1)} \ge \frac{\Pr(A_iD_0D_1)}{\Pr(D_0)}$$

by Bayes' formula. We're trying to use the fact that D_0 is independent of A_i here, so by Bayes' rule again, this is equal to

$$= \Pr(A_iD_1|D_0).$$

We can write this as the probability

$$\Pr(A_i|D_0) - \Pr(A_i\overline{D_1}|D_0).$$

By independence of A_i and D_0, this first term is now $\Pr(A_i)$. Now because A_i is increasing, $\overline{D_1}$ is increasing, and D_0 is decreasing, we can use the Harris-FKG inequality: conditioning on a decreasing event must make the probability of an increasing event go down, so

$$\Pr(A_i\overline{D_1}|D_0) \le \Pr(A_i\overline{D_i}) = \Pr(A_i \cap \vee_{j<i,j\sim i}A_j)$$

and by a union bound, this is at most

$$\le \sum_{j<i,j\sim i} \Pr(A_i \cap A_j),$$

as desired.

□

Fact 7.12

Janson's inequality is originally about random subsets including a certain S_i, but we actually only need that the A_is are increasing events. So they don't necessarily have to be of the form $S_i \subseteq R$ for our random subset R.

Previously, our dependency graph had edges wherever the S_is overlapped. In general, there's a difference between pairwise independence and mutual independence, so it seems that we have to be careful. However, we get lucky:

Lemma 7.13

If A, B, C are increasing events, and A is independent of B and C, then A is independent of $B \wedge C$.

So the pairwise dependence graph is the same as the dependency graph. This may not be that intuitive: remember that one counterexample was $R \subset [3]$, where A_1 is the event "$|R \cap \{1,2\}|$ is even", A_2 is the event"$|R \cap \{1,3\}|$ is even," and A_3 is the event "$|R \cap \{2,3\}|$ is even:" we have pairwise independence but not mutual independence.

Proof. Note that $\Pr(A \wedge (B \wedge C)) + \Pr(A \wedge (B \vee C)) = \Pr(A \wedge B) + \Pr(A \wedge C) = \Pr(A)(\Pr(B) + \Pr(C))$. But on the other hand, by Harris-FKG, since the events are increasing,

$$\Pr(A \wedge (B \wedge C)) \geq \Pr(A)\Pr(B \wedge C), \Pr(A \wedge (B \vee C)) \geq \Pr(A)(\Pr(B \vee C)).$$

But these last two inequalities actually add to the first equality! So equality must occur, and A is independent of $B \wedge C$ and $B \vee C$. □

Let's use the first Janson inequality to get an upper bound on the probability that $G(n, p)$ is triangle-free. Recall the Second Moment Method calculations that we made earlier: the expected number of triangles is

$$\mu = \binom{n}{3} p^3 \asymp n^3 p^3.$$

Meanwhile, Δ is counting the number of pairs of triangles with a shared edge: these look like 4-cycles with a diagonal, and that evaluates out to

$$\Delta \asymp n^4 p^5.$$

So $\Delta \ll \mu \iff p \ll n^{-1/2}$, and therefore the probability that $G(n, p)$ is triangle free is $e^{-(1+o(1))\mu}$ if $p \ll n^{-1/2}$. Note that this is exactly the right asymptotic behavior (see the logic after Theorem 7.11).

Well, what if p is larger - does the formula still hold when $p \gtrsim n^{-1/2}$, and is Harris a good approximation? Bipartite graphs are triangle-free, so the probability of a triangle-free $G(n, p)$ is at least the probability that it is bipartite. This is at least the probability that $G(n, p)$ has no edges, which is $(1-p)^{\binom{n}{2}} = e^{-(1+o(1)pn^2/2}$ for $p \ll 1$. This is actually much larger than $e^{-c\mu}$ if $p \gg n^{-1/2}$! So the lower bound of Harris is true, but it's inferior to very stupid bounds.

7.4 The second Janson inequality

Let's try to strengthen our inequality when $\Delta > \mu$

Theorem 7.14 (Second Janson Inequality)

Again using the assumptions of Setup 7.10, if $\Delta \geq \mu$, then

$$\Pr(X = 0) \leq e^{-\frac{\mu^2}{2\Delta}}.$$

Proof. For each subset of the bad events $T \subset [k]$, let

$$X_T = \sum_{i \in T} 1_{A_i}$$

be the number of bad events in T only, and $\mu_T = \sum_{i \in T} \Pr(A_i), \Delta_T = \sum_{(i,j) \in T^2, i \sim j} \Pr(A_i \wedge A_j)$ be defined similarly. Then the probability that none of the bad events occur is always

$$\Pr(X = 0) \le \Pr(X_T = 0) \le e^{-\mu_T + \frac{\Delta_T}{2}}$$

Choose T randomly: include each element independently with some probability q (to be determined). Then μ_T has expectation $q\mu$, and Δ_T has expectation $q^2\mu$ (since both A_i and A_j need to be kept for any term to count). Thus,

$$\mathbb{E}(-\mu_T + \frac{\Delta_T}{2}) = -q\mu + \frac{q^2\Delta}{2}.$$

Minimizing this, pick $q = \frac{\mu}{\Delta}$, which is at most 1 by the theorem statement. This yields $\frac{\mu^2}{2\Delta}$, so there exists some choice of T so that $-\mu_T + \frac{\Delta_T}{2} \le \frac{-\mu^2}{2\Delta}$, as desired. □

These two Janson inequalities work in different regimes, and it's interesting that the proof of the second uses the proof of the first!

Remark. *This "bootstrapping argument," where we start with a weak inequality and make it stronger, is reminiscent of the crossing number inequality. We had*

$$cr(G) \ge |E| - 3|V|,$$

and this was only quadratic in n. To get a stronger result, we sampled our graph G, which gave us a much stronger inequality of the form $cr(G) \gtrsim \frac{|E|^3}{|V|^2}$.

How do these compare to the second moment method calculations? There, we said that if $\Delta \ll \mu^2$, and $\mu^2 \to \infty$, then X is concentrated around its mean, meaning $\Pr(X = 0) = o(1)$. But here, we have an explicit exponential decay, rather than just knowing that the probability goes to zero.

Does this give a better bound than the first Janson inequality for $G(n,p)$ being triangle free (when p is large)? Say that $p \gg n^{-1/2}$, so that we have $\Delta \ge \mu$. By Janson's second inequality, the probability that $G(n,p)$ is triangle free is now

$$\le e^{-\frac{\mu^2}{2\Delta}} = e^{-\Theta(n^2 p)}.$$

The exponent matches the order of "probability $G(n,p)$ has no edges" from above, which means this is essentially tight! So

$$\Pr(G(n,p) \text{ is triangle free}) = \begin{cases} e^{-(1+o(1))n^3p^3/6} & p \ll n^{-1/2} \\ e^{-\Theta(n^2 p)} & p \gtrsim n^{-1/2}. \end{cases}$$

What is the constant here in the Θ? We can do a bit better than "$G(n,p)$ has no edges," since $G(n,p)$ has probability of being bipartite at least

$$(1-p)^{2\binom{n/2}{2}} = e^{-(1+o(1))n^2p/4}.$$

Is this the dominating way of generating graphs with no triangles? The answer turns out to be yes, but we don't yet have the tools to show that. The modern way to think about this is through something called the container method.

Fact 7.15

A lot of probability distributions that are nonnegative-integer-valued are Poisson or Gaussian. In particular, for the $p \ll n^{-1/2}$ case, if the exponent converges to a constant, we get Poisson behavior for the number of triangles in $G(n, p)$.

7.5 Lower tails: the third Janson inequality

One more time, let X be the number of triangles in $G(n, p)$. This time, we want to estimate the probability that X is at most $0.9\mathbb{E}[X]$ (or some other constant times the mean): this is a generalization of estimating the probability that $X = 0$.

This is on a larger order than the standard deviation of X, so Chebyshev-like tools don't help us. It turns out that in these cases, we have exponential decay:

Theorem 7.16 (Third Janson inequality)

Use the assumptions in Setup 7.10 again. For any $0 \le t \le \mu$,

$$\Pr(X \le \mu - t) \le \exp\left(\frac{-t^2}{2(\mu + \Delta)}\right).$$

We'll come back to the proof of this later - interestingly, it also bootstraps the first Janson inequality. First, let's look at a consequence of this by looking some more at triangle counts (here, triangle can also be replaced with any subgraph H). If we still let X be the number of triangles in $G(n, p)$, and we let $t = c\mu \asymp n^3p^3$ for some constant c, then

$$\Pr(X \le (1 - c)\mathbb{E}[X]) \le \exp\left(-\Theta\left(\frac{n^6p^6}{n^3p^3 + n^4p^5}\right)\right).$$

We can clean this up a bit by splitting by dominant term:

$$\Pr(X \le (1 - c)\mathbb{E}[X]) = \begin{cases} \exp(-\Theta(n^3p^3)) & p \lesssim n^{-1/2} \\ \exp(-\Theta(n^2p)) & p \gtrsim n^{-1/2} \end{cases}$$

Are these inequalities tight – that is, do we have a corresponding lower bound? Turns out the answer is yes! The probability of having at most $(1 - c)\mathbb{E}[X]$ includes the probability that there at exactly 0 triangles, and the upper bounds that we just found have exponents on the same order as what we found for the 0-triangle case. This means our bounds are tight up to constant factors in the exponent.

Unfortunately, we don't actually know the value of the constants except for some values of c. We are essentially asking "what is the best way of getting few triangles?", and one good way to do this is to uniformly decrease probability of edges everywhere, which helps for small c. On the other hand, when c is close to 1, we expect that bipartite graphs dominate the few-triangle space. However, it's still a research problem to look at values of c in between and find the dominating graphs.

By the way, there's a reason we're only mentioning the lower tail: the upper tail is completely different. If we use $X \ge (1 + c)\mathbb{E}[X]$, our inequalities are false!

Proof of Theorem 7.16 by Warnke. Define the parameter $q \in [0, 1]$ (value to be determined later). Let $T \subseteq [k]$, where each element is included with probability q independently; let's consider

$$X_T = \sum_{i \in T} 1_{A_i}.$$

We can alternatively write this as a sum over the original bad events:

$$= \sum_{i \in [k]} 1_{A_i} W_i,$$

where each W_i is distributed according to a Bernoulli distribution: 1 with probability q if $i \in T$ and 0 otherwise.

Notice that X, our actual number of bad events, tells us

$$\Pr(X_T = 0 | X) = (1-q)^X,$$

because this is the probability that none of the bad events that occurred are included in T. Taking expectations on both sides,

$$\mathbb{E}[(1-q)^X] = \Pr(X_T = 0) \le e^{-\mu' + \Delta'/2}$$

by the first Janson inequality, where $\mu' = \mathbb{E}[X_T] = q\mu$ and Δ' (similarly) is $= q^2\Delta$. We can think of the left side as a moment generating function if we want to get exponential bounds!

So now by Markov's inequality,

$$\Pr(X \le \mu - t) = \Pr((1-q)^X \ge (1-q)^{\mu - t}) \le (1-q)^{-\mu+t} \mathbb{E}[(1-q)^X]$$

and plugging in our result,

$$\le (1-q)^{-\mu+t} e^{-q\mu' + q^2 \frac{\Delta}{2}}.$$

Now we just optimize for q. One thing we can do is take the derivative and set things equal to 0, but instead, we can give ourselves some slack: let's let $1 - q = e^{-\lambda}$ for some $\lambda \ge 0$. By the Taylor expansion, $\lambda - \frac{\lambda^2}{2} \le q \le \lambda$. Plugging this in,

$$\Pr(X \ge \mu - t) \le \exp\left[\lambda(\mu - t) - \left(\lambda - \frac{\lambda^2}{2}\right) + \frac{\lambda^2 \Delta}{2}\right] = \exp\left[-\lambda t + \frac{\lambda^2}{2}(\mu + \Delta)\right],$$

and now set $\lambda = \frac{1}{\mu + \Delta}$ to get the result. □

Notice that the proof of Janson's third inequality only works for lower tails: in particular, the proof starts by using the probability that $X = 0$ and builds up from there. In fact, upper tails are completely different! Let's show that for $p \gtrsim n^{-1/2}$,

$$\Pr(X \ge (1+c)\mathbb{E}[X]) \ge e^{-Cn^2 p}.$$

To do this, we ask: what's the "cheapest" possible way to generate lots of triangles? If we plant a clique, we get something pretty good:

$$\Pr(X \ge 2\mathbb{E}[X]) \ge \Pr(G(n,p) \text{ has a clique on the first } 10np \text{ vertices}),$$

because this already gives $\binom{10np}{3}$ triangles, which is more than $2\binom{n}{3}p^2$. Then the probability $G(n,p)$ has a clique in the first $10np$ vertices is

$$p^{\binom{10np}{2}} \ge e^{-cn^2p^2 \log(1/p)},$$

and this is exponentially larger than the previous bound for lower tails.

So what's the truth here? There was a paper written about the "infamous upper tail:" it just demonstrated that a wide range of techniques did not work for showing bounds on the upper tail. About 10 years ago, though, the following result (on the same order of the planted clique) was proved:

$$\Pr(X \geq 2\mathbb{E}[X]) = e^{-\Theta(n^2 p^2 \log(1/p))} \text{ if } p \gtrsim \frac{\log n}{n}.$$

The constant in the Θ here is a bit tricky, too. Basically, there are two main constructions for generating lots of triangles: have a hub of size cnp^2 and connect them to everything else, or have a clique of some select size. But this isn't obvious, and there's lots of open research here!

7.6 Revisiting clique numbers

Recall that in our Second Moment Method discussions, we found that

$$\omega\left(G\left(n, \frac{1}{2}\right)\right) \sim 2\log_2(n)$$

with high probability (this is Theorem 4.21). We actually found two-point concentration: let's review the ideas of the proof so that we can look at them more closely.

Proof. Let X_k be the number of k-cliques in $G\left(n, \frac{1}{2}\right)$: the expected value of X_k is then

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

If k is a function of n, and our mean $\mu_k \to 0$, then $X = 0$ with high probability by Markov's inequality. Meanwhile, if $\mu_k \to \infty$, then $\Delta \ll \mu^2$: variance is much smaller than the squared mean, so by the second moment method, there is always a k-clique with high probability due to concentration. □

Where does the quantity μ_k cross a threshold? We defined k_0 to be the largest k such that $\mu_k \geq 1$. Another routine calculation shows us that around that value of $k \sim 2\log_2 n$,

$$\frac{\mu_{k+1}}{\mu_k} = n^{-1+o(1)}$$

and this implies that the clique number of $G\left(n, \frac{1}{2}\right)$ is concentrated around $2\log_2 n$ with high probability: in fact, it's either k_0 or $k_0 - 1$ with high probability.

Problem 7.17

Is the clique number of $G\left(n, \frac{1}{2}\right)$ really concentrated around two different points, or do we just have a weakness in our proof?

Solution. Let's say that we have k as a function of n so that $\mu = \binom{n}{k} 2^{-\binom{k}{2}} \to c$ converges as $n, k \to \infty$. Doing the relevant calculations, we find that $\Delta = o(1)$, so $\Delta \ll \mu$ here. For all S that are k-element subsets of $[n]$, let A_S be the event that S forms a clique: then the probability for any given A_S is $2^{-\binom{k}{2}}$. Then $X = 0$, which is the probability that there are no k-cliques, is

$$\Pr\left(\omega\left(G\left(n, \frac{1}{2}\right)\right) < k\right) \leq e^{-\mu + \Delta/2}$$

by the first Janson inequality, and we also have $e^{-\mu(1+o(1))}$ as a lower bound by Harris. So because $\Delta \ll \mu$, the lower and upper bounds grow closer:

$$\Pr\left(\omega\left(G\left(n, \frac{1}{2}\right)\right) < k\right) = e^{-(1+o(1))\mu} \to e^{-c},$$

which is some specific value between 0 and 1. To be more rigorous about this, we're saying that for $\lambda \in (-\infty, \infty)$, if $n_0(k)$ is the minimum n such that $\binom{n}{k}2^{-\binom{k}{2}} \ge 1$, then $n = n_0(k)\left(1 + \frac{\lambda + o(1)}{k}\right)$, and thus the mean

$$\mu = \binom{n}{k}2^{-\binom{k}{2}} = e^{\lambda} + o(1).$$

So that means that

$$\Omega\left(G\left(n, \frac{1}{2}\right)\right) = \begin{cases} k-1 & \text{with probability } 1 - e^{-e^{\lambda}} + o(1) \\ k & \text{with probability } e^{-e^{\lambda}} + o(1) \end{cases}.$$

and we can get sequences with two-point concentration of any probability we want. □

Fact 7.18

On the other hand, most n do have one-point concentration, so both cases (one- and two-point concentration) do occur. One way to view this problem in general is that these situations look sort of Poisson in their distribution, and that's the right kind of case to use Janson's inequalities.

7.7 Revisiting chromatic numbers

We're ready to go back to thinking about this result from earlier in the class now. Remember that Corollary 4.23, we found that the independence number

$$\alpha\left(G\left(n, \frac{1}{2}\right)\right) \sim 2\log_2 n$$

with high probability, and because each color class is an independent set,

$$\chi(G) \ge \frac{n}{\alpha(G)} \ge (1 + o(1))\frac{n}{2\log_2(n)}$$

with high probability. This is a lower bound: the methods we had before didn't allow us to get an upper bound, but now we have the Janson inequalities.

Theorem 7.19

We have with high probability that

$$\chi\left(G\left(n, \frac{1}{2}\right)\right) \sim \frac{n}{2\log_2(n)}.$$

Proof. The idea is that we will use a "greedy coloring," since our goal is to show that we can color all of G using $(1 + o(1))\frac{n}{2\log_2 n}$ colors with high probability.

At each step of our strategy, we take out an independent set of size $(1 + o(1))2\log_2(n)$. Each time we do that, color all of those with one of the colors, and remove the independent set from our graph. We'll stop when we have $o\left(\frac{n}{\log n}\right)$ vertices left: at that point, we finish by just coloring with a different color for every vertex.

Why can we perform that step of removing an independent set of size $(1 + o(1))2\log_2(n)$? It is sufficient to show that with "very" high probability, every "not-too-small" subset of G has an independent set of size $(1 + o(1))2\log_2 n$: here "very high" means exponentially small.

Lemma 7.20

There exists $k \sim 2\log_2 n$ such that

$$\Pr\left(\omega\left(G\left(n, \frac{1}{2}\right)\right) < k\right) < e^{-n^{2-o(1)}}.$$

By extension, this is also true for the independence number of G.

Proof of lemma. Use the notation from the clique number calculations (specifically k_0). Let $k = k_0 - 3$ (so that we have a significantly smaller number of cliques): what's the probability that we don't have any k-cliques? This should be very small if there's high probability of k_0-clique! In particular, the mean of the number of k-cliques is (because we change by a factor of n each time we change k by one)

$$\mu_k > n^{3-o(1)}.$$

We can also calculate

$$\Delta \sim \mu^2 \frac{k^2}{n^2} > n^{4+o(1)} > \mu,$$

so by the second Janson inequality,

$$\Pr\left(\omega\left(G\left(n, \frac{1}{2}\right)\right) < k\right) \le e^{-\frac{\mu^2}{2\Delta}} = e^{-n^{2+o(1)}}$$

as desired. □

This decays very quickly. So now given a graph $G \sim G\left(n, \frac{1}{2}\right)$, let's take $m = \left\lfloor \frac{n}{\log^2 n} \right\rfloor$: for every set of m vertices, let $G[S]$ be the graph induced by the vertices S. The probability the independence number of $G[S]$ is less than k is

$$e^{-m^{2-o(1)}} = e^{-n^{2-o(1)}},$$

because the $G[s]$ looks like $G\left(m, \frac{1}{2}\right)$ for some $k \sim 2\log_2 m \sim 2\log_2 n$. Summing over all S and doing a union bound, the probability that

$$\Pr(\alpha(G[S]) < k \text{ for some } S) < 2^n e^{-n^{2-o(1)}} = o(1),$$

so with high probability, every m-element subset of G contains a k-element independent set. Thus we can carry out our greedy coloring, and the total number of colors we use (including the last part where we use one color for each vertex) is

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

as desired. □

Note that this proof only works because we can get an exponential bound from the Janson inequalities! Bollobás' theorem guarantees some kind of concentration, but the window of deviation is still basically an open problem.

8 Martingale convergence and Azuma's inequality

8.1 Setup: what is a martingale?

Definition 8.1

A **martingale** is a sequence of random variables $Z_0, Z_1, \cdots$, such that for every n, $\mathbb{E}|Z_n| < \infty$ (this is a technical assumption), and

$$\mathbb{E}[Z_{n+1}|Z_0, \cdots, Z_n] = Z_n.$$

This comes up in a lot of different ways:

Example 8.2

Consider a random walk $X_1, X_2, \cdots$ of independent steps, each with mean 0. Then we can define the martingale

$$Z_n = \sum_{i \le n} X_i,$$

which fits the definition because we always expect our average position after step $n+1$ to be the same as where we just were after step n.

Example 8.3 (Betting strategy)

Let's say we go to a casino, and all bets are "fair" (have expectation 0). For example, we may bet on fair odds against a coin flip. Our strategy can adapt over time based on the outcomes: let Z_n be our balance after n rounds. This is still a martingale!

This is more general than just a random walk, because now we don't need the steps to be independent.

Example 8.4

Let's say my goal is to win 1 dollar. I adapt the following strategy:

- Bet a dollar; if I win, stop.
- Otherwise, double the wager and repeat.

This is a martingale, because all betting strategies are martingales. With probability 1, we must always win at some point, so we end up with 1 dollar at the end! This sounds like free money, but we have a finite amount of money (so this would never occur in real life).

Remark. *This is called the "martingale betting strategy," and it's where the name comes from!*

Definition 8.5 (Doob or exposure martingale)

Suppose we have some (not necessarily independent) random variables $X_1, \cdots, X_n$, and we have a function $f(x_1, \cdots, x_n)$. Then let

$$Z_i = \mathbb{E}[f(x_1, \cdots, x_n)|X_1, \cdots, X_i].$$

Basically, we "expose" the first i outputs to create Z_i. It's good to check that this is actually a martingale: show that

$$\mathbb{E}[Z_i | Z_0, \cdots, Z_{i-1}] = Z_{i-1}.$$

Note that f may also be some random variable: for example, it could be the chromatic number of the graph, and X_i are indicator variables of the edges. Then Z_0 is $\mathbb{E}[f]$, Z_1 is a revised mean after we learn about the status of an edge, and so on. This is called an **edge-exposure martingale**.

Let's discuss that more explicitly: we reveal the edges of $G(n, p)$ one at a time. For example, let's say we want $\chi(G(3, \frac{1}{2}))$. There are eight possible graphs, with equal probabilities, and six of them have chromatic number 2, one has chromatic number 3, and one has chromatic number 1. The average is $Z_0 = 2$.

Now, the chromatic number is either 2.25 or 1.75, depending on on whether the bottom edge is present or not. This average is 2, and then we can keep going: Z_2 is either 2.5 or 2 if $Z_1 = 2.25$, and 2 or 1.5 if $Z_1 = 1.75$. The idea is that each Z_{n+1}'s expected value is dependent on the previous mean.

Alternatively, we can have a **vertex-exposure martingale**: at the ith step, expose all edges (j, i) with $j < i$. So there are different ways of constructing this martingale, and which one to use depends on the application!

8.2 Azuma's inequality

Why are martingales useful? Here's an important inequality that's actually not too hard to prove:

Theorem 8.6 (Azuma's inequality)

Given a martingale $Z_0, \cdots, Z_n$ with bounded differences

$$|Z_i - Z_{i-1}| \le 1 \forall i \in [n],$$

we have a tail bound for all λ:

$$\Pr(Z_n - Z_0 \ge \lambda\sqrt{n}) \le e^{-\lambda^2/2}.$$

More generally, though, if we have $|Z_i - Z_{i-1}| \le c_i$ for all $i \in [n]$, then for all $a > 0$,

$$\Pr(Z_n - Z_0 \ge a) \le \exp\left(\frac{-a^2}{2\sum_{i=1}^n c_i^2}\right).$$

(This is sometimes also known as Azuma-Hoeffding.) We've seen this before from our discussion of Chernoff bounds, which is a special case by making the martingale a sum of independent random variables! This is not a coincidence - we'll notice similarities in the proof.

This theorem is useful when none of the martingale steps have big differences. It is generally more difficult to prove any sort of concentration when we don't have bounded differences in our martingale, though.

Proof. We can shift the martingale so that $Z_0 = 0$. Let $X_i = Z_i - Z_{i-1}$ be the martingale differences: these X_is do not need to be independent, but they must always have mean 0.

Lemma 8.7

If X is a random variable with $\mathbb{E}[X] = 0$ and $|x| \le c$, then

$$\mathbb{E}[e^X] \le \frac{e^c + e^{-c}}{2} \le e^{c^2/2}$$

by looking at Taylor expansion and comparing coefficients.

Proof of lemma. Basically, we maximize e^X by having a $\pm c$ Bernoulli variable - this is true because of convexity! Specifically, we can upper-bound e^X by the line connecting the endpoints:

$$e^x \le \frac{e^c + e^{-c}}{2} + \frac{e^c - e^{-c}}{2c}x,$$

and now take expectations of the statement when $x = X$. □

So now let $t \ge 0$, and consider the moment generating function $\mathbb{E}[e^{tZ_n}]$. We'll split this up as

$$\mathbb{E}[e^{tZ_n}] = \mathbb{E}[e^{t(X_n + Z_{n-1})}] = \mathbb{E}\left[\mathbb{E}\left[e^{tX_n} \mid Z_{n-1}\right] e^{tZ_{n-1}}\right]$$

By definition, the inner expectation is the moment generating function of a mean-zero random variable bounded by tc_n, and thus

$$\mathbb{E}[e^{tZ_n}] \le e^{t^2 c_n^2/2}\mathbb{E}[e^{tZ_{n-1}}].$$

Repeating this calculation or using induction, we find that the expectation of e^{tZ_n} is bounded by

$$\le \exp\left[\frac{t^2}{2}(c_n^2 + c_{n-1}^2 + \cdots + c_1^2)\right].$$

To finish the proof, we repeat the logic of the Chernoff bound proof: by Markov's inequality on the moment generating function,

$$\Pr(Z_n \ge a) \le e^{-ta}\mathbb{E}[e^{tZ_n}] \le e^{-ta + t^2(c_1^2 + \cdots + c_n^2)/2}.$$

We can now set t to be whatever want: taking $t = \frac{a}{\sum_i c_i^2}$ yields the result. □

The main difference from Chernoff is that we do one step at a time, and this crucially requires that we have bounded differences. We can also get a lower tail for Z_n in the exact same way, and putting these together yields the following:

Corollary 8.8

Let Z_n be a martingale where $|Z_i - Z_{i-1}| \le c_i$ for all $i \in [n]$, as in Theorem 8.6. Then for all $a > 0$,

$$\Pr(|Z_n - Z_0| \ge a) \le 2\exp\left(\frac{-a^2}{2\sum_i c_i^2}\right).$$

Basically, we can't walk very far in either direction in a martingale with an interval of $\sqrt{n}$, even when our choices can depend on previous events.

8.3 Basic applications of this inequality

The most common way Azuma is used is to show concentration for Lipschitz functions (on a domain of many variables).

Theorem 8.9

Consider a function

$$f : \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n \to \mathbb{R}$$

such that $|f(x) - f(y)| \le 1$ whenever x and y are vectors that differ in exactly 1 coordinate. (This is known as being **1-Lipschitz** with respect to Hamming distance.) Then if $Z = f(X_1, \cdots, X_n)$ is a function of **independent random variables** $X_i \in \Omega_i$, we have high concentration:

$$\Pr(Z - \mathbb{E}[Z] \ge \lambda\sqrt{n}) \le e^{-\lambda^2/2}.$$

Proof. Consider the Doob martingale

$$Z_i = \mathbb{E}[Z|X_1, \cdots, X_i].$$

Note that $|Z_i - Z_{i-1}| \le 1$, because revealing 1 coordinate cannot change the value of our function by more than 1. But now Z_0 is the expected value of our original function Z, since we have no additional information: thus, $Z_0 = \mathbb{E}[Z]$. Meanwhile, Z_n means we have all information about our function, so this is just Z. Plugging these into the Azuma inequality yields the result. □

It's important to note that the $|Z_i - Z_{i-1}| \le 1$ step only works if we have independence between our random variables - that step is a bit subtle.

Example 8.10 (Coupon collecting)

Let's say we want to collect an entire stack of coupons: we sample $s_1, \cdots, s_n \in [n]$. Can we describe X, the number of missed coupons?

Explicitly, we can write out

$$X = |[n] \setminus \{s_1, \cdots, s_n\}|.$$

It's not hard to calculate the expected value of X: by linearity of expectation, each coupon is missed with probability $\left(1 - \frac{1}{n}\right)^n$, so

$$\mathbb{E}[X] = n\left(1 - \frac{1}{n}\right)^n.$$

This value is between $\frac{n-1}{e}$ and $\frac{n}{e}$. Typically, how close are we to this number? Changing one of the s_is can only change X by at most 1 (we can only gain or lose up to one coupon). So by the concentration inequality,

$$\Pr\left(\left|X - \frac{n}{e}\right| \ge \lambda\sqrt{n} + 1\right) \le \Pr(|X - \mathbb{E}[X]| \ge \lambda\sqrt{n}) \le 2e^{-\lambda^2/2},$$

where the $+1$ is for the approximation of $\frac{1}{e}$ we made. So the number of coupons we miss is pretty concentrated! This would have been more difficult to solve without Azuma's inequality, because whether or not two different coupons are collected are dependent variables.

Theorem 8.11

Let $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \in \{-1, 1\}^n$ be uniformly and independently chosen. Fix vectors $v_1, \cdots, v_n$ in some norm space (Euclidean if we'd like) such that all vectors $|v_i| \le 1$. Then $X = ||\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n||$ is pretty concentrated around its mean:

$$\Pr(|X - \mathbb{E}[X]| \ge \lambda\sqrt{n}) \le 2e^{-\lambda^2/2}.$$

Even if we can't compute the mean of this variable, we can still get concentration! Note that if the v_is all point along the same axis, then we essentially end up with the Chernoff bound.

Proof. Our Ω_is are $\{-1, 1\}$, and we have a function defined as

$$f(\varepsilon_1, \cdots, \varepsilon_n) = ||\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n||.$$

If we change a coordinate of f, the norm can change by at most 2 by the triangle inequality, because each v_i has norm at least 1. Plugging this into Azuma, we find that

$$\Pr(|X - \mathbb{E}[X]| \geq \lambda\sqrt{n}) \leq 2e^{-\lambda^2/8}.$$

This is usually good enough (the exponent is of the right order), but with a little more care, we can change the constant in our exponent from $\frac{1}{8}$ to $\frac{1}{2}$. Let's go back to the exposure martingale, and let Y_i be the expected value of our function f after having $\varepsilon_1, \ldots, \varepsilon_i$ revealed: we claim that $|Y_i - Y_{i-1}| \leq 1$.

Why is this the case? If we've revealed the first $i-1$ coordinates, let $\vec{\varepsilon}$ and $\vec{\varepsilon'}$ be two vectors in $\{-1, 1\}^\varepsilon$ differing only in the ith coordinate. Then

$$Y_{i-1}(\vec{\varepsilon}) = \frac{Y_i(\vec{\varepsilon}) + Y_i(\vec{\varepsilon'})}{2}:$$

we should average over what happens in the ith coordinate if we already know what happens in the ith coordinate. So now plugging in,

$$|Y_i(\varepsilon) - Y_{i-1}(\varepsilon)| = \frac{1}{2}|Y_i(\vec{\varepsilon}) - Y_i(\vec{\varepsilon'})| \leq \frac{1}{2} \cdot 2||v_i|| \leq 1,$$

and now Azuma gives us the desired constant! □

8.4 Concentration of the chromatic number

Last time, we derived (using Janson's inequality) an estimate for the chromatic number of a random graph, and this took some work. But it turns out that we can prove concentration of the chromatic number without knowing the mean:

Theorem 8.12

Let $G = G(n, p)$ be a random graph. Then

$$\Pr\left(|\chi(G) - \mathbb{E}[\chi(G)]| \leq \lambda\sqrt{n-1}\right) \leq 2e^{-\lambda^2/2}.$$

To prove this, let's think about the process of finding the chromatic number as a martingale. This does not even require us knowing that $\chi(G)$ is about $\frac{n}{2\log_2 n}$: proving concentration is somehow easier than finding the mean here!

Proof. There are many ways to expose the edges of a graph: sometimes we need to choose between edge and vertex exposure. Here, we'll do the latter.

Consider the vertex-exposure martingale. Basically, we're given the status of all edges connected to one of the first i vertices, and then we try to figure out the estimate from there. Note that $|Z_i - Z_{i-1}| \leq 1$: we could always just give the ith vertex a new color to preserve proper coloring, so the expected chromatic numbers can't differ by more than 1. But now we're done by applying Azuma's inequality. □

Let's try seeing what happens if we used the edge-exposure martingale instead. We have more steps: there are $\binom{n}{2}$ edges to reveal, so we should get about $\Theta(n)$-size deviation. That's already not so good, since we're trying to find chromatic number (which is size n). We can't even bound $|Z_i - Z_{i-1}|$ any better than before!

Remark. *It's important to note that in general, it's not always better to use the vertex or edge exposure martingale. Instead, our method really depends on what the maximum differences are between subsequent steps.*

As a final note, can we also set up a Lipschitz function to rephrase the setting of this problem? Our random variable spaces in the vertex-exposure martingale, Ω_i, aren't edges but batches of edges: at step i, we want $\{0,1\}^{i-1}$, the edges going "left" from vertex i. So the vertex-exposure partitions our edge and exposes groups at a time: if we do the batching appropriately, we get a better gain than naively having our probability spaces just be $\{0,1\}$ for each edge. So the Ω_is are fairly general, but they must still be independent.

The idea of a tail bound is the same as a confidence interval: Azuma tells us that we can take an interval with width on the order of $\sqrt{n}$, and at least some constant fraction of our random graphs will give chromatic number in that interval. This might be overly generous, though: it's a major open problem to know the actual fluctuation of the chromatic number of a random graph!

8.5 Four-point concentration?

Interestingly, we can get even better concentration if we have sufficiently small p:

Theorem 8.13

Let $\alpha > \frac{5}{6}$ be a fixed constant. If $p < n^{-\alpha}$, then the chromatic number $\chi(G(n,p))$ is concentrated among four values with high probability. Specifically, there exist a function $u = u(n,p)$ such that as $n \to \infty$,

$$\Pr(u \le \chi(G(n,p)) \le u+3) = 1 - o(1).$$

In other words, sparser graphs are easier to estimate. Because the probability of an edge appearing here is relatively small, we can get more concentration than with our earlier calculations.

Proof. It suffices to show that for any fixed $\varepsilon > 0$, we can find a sequence $u = u(n,p,\varepsilon)$ such that as $n \to \infty$,

$$\Pr(u \le \chi(G(n,p)) \le u+3) > 1 - \varepsilon - o(1).$$

Pick u to be the smallest positive integer such that $\Pr(\chi(G(n,p)) \le u) > \varepsilon$. (This is deterministic, even if we may not know how to evaluate it.) Then the probability that $\chi(G) < u$ is at most ε, and we just want to show that $\chi(G) > u+3$ with probability $o(1)$.

The next step is very clever: let $Y = Y(G)$ be the minimum size of a subset of the vertices $S \subset V(G)$ such that $G - S$ may be properly colored with u colors. Basically, we color as well as we can, and Y tells us how close we are to success.

Now Y is 1-Lipschitz with respect to the vertex-exposure martingale: if we change a vertex in G, then $Y(G)$ changes by at most 1. So by Azuma's inequality,

$$\Pr(Y \le \mathbb{E}[Y] - \lambda\sqrt{n}) \le e^{-\lambda^2/2},$$

$$\Pr(Y \ge \mathbb{E}[Y] + \lambda\sqrt{n}) \le e^{-\lambda^2/2},$$

This trick will come up a lot: we'll use both the upper and lower tail separately. We don't need to know the expectation to find concentration, but we'll use the lower-tail bound to bound $\mathbb{E}[Y]$. With probability at least ε, G is u-colorable. That's equivalent to saying that $Y = 0$, which occurs with probability

$$\varepsilon < \Pr(Y \le \mathbb{E}[Y] - \mathbb{E}[Y]) \le \exp\left(-\frac{\mathbb{E}[Y]^2}{2n}\right).$$

Simplifying this, this already gives us a bound (for some λ, which is a function of ε)

$$\mathbb{E}[Y] \le \sqrt{2\log\left(\frac{1}{\varepsilon}\right)n} = \lambda\sqrt{n},$$

which is what we should expect from a martingale of this form. Similarly, we can do an upper-tail bound to show that Y is rarely too big relative to the mean:

$$\Pr(Y \ge 2\lambda\sqrt{n}) \le \Pr(Y \ge \mathbb{E}[Y] + \lambda\sqrt{n}) \le e^{-\lambda^2/2} = \varepsilon$$

by the definition of λ. So the number of uncolored vertices is not too big: by the definition of Y, we now know that with probability at least $1-\varepsilon$, we can color all but $2\lambda\sqrt{n}$ vertices. Here's the key step: We'll show that with high probability, we can color the remaining vertices with just 3 colors.

Lemma 8.14

Fix $\alpha > \frac{5}{6}$ as before, as well as a constant C, and let $p < n^{-\alpha}$. Then with high probability, every subset of size at most $C\sqrt{n}$ vertices in $G(n,p)$ can be properly 3-colored.

We want to union bound the bad probabilities, but we must be a bit careful here. Suppose the lemma were false for some graph G (that is, we're in one of the bad cases). Choose a minimal size $T \subset V(G)$ that is not 3-colorable. Consider the induced subgraph $G[T]$ (taking only the edges between the vertices in T). This has minimum degree 3, because if there's a vertex x with $\deg_T(x) < 3$, then $T - x$ is also not 3-colorable, which contradicts the minimality of T.

So $G[T]$ has at least $3|T|/2$ edges, and now we can just bound the probability that there exists some T (of size at least 4, since that's the only way for it to be not 3-colorable) with $|T| \le C\sqrt{n}$ that contains at least $3|T|/2$ edges: union bounding, this is at most

$$\le \sum_{t=4}^{C\sqrt{n}} \binom{n}{t}\binom{\binom{t}{2}}{3t/2} p^{3t/2}.$$

Now we just need to show that this quantity is $o(1)$ (as n goes to ∞): this simplifies to

$$\le \sum_{t=4}^{C\sqrt{n}} \left(\frac{ne}{t}\right)^t \left(\frac{te}{3}\right)^{3t/2} n^{-3t\alpha/2} = \sum_{t=4}^{c\sqrt{n}} \left(O(n^{1-3\alpha/2+1/t})\right)^t = o(1)$$

if $\alpha > \frac{5}{6}$.

So in summary, we know that once we have all but $C\sqrt{n}$ of the points colored with u colors with $1-o(1)$ probability, we have $1-o(1)$ probability of coloring the rest in at most 3 colors. Now just take ε arbitrarily small to show the result. □

The hardest part of this proof is finding an informative random variable that is Lipschitz! It turns out that better bounds are known: we actually have two-point concentration for all $\alpha > \frac{1}{2}$, and the proof comes from refinements of this technique.

8.6 Revisiting an earlier chromatic number lemma

Remember that when we discussed Janson's inequality, we considered the following key claim from Bollobás' paper, helpful for taking out large indepndent sets:

Lemma 8.15

Let $\omega(G)$ be the number of vertices in the largest clique of $G(n,p)$, and let k_0 be the minimum positive integer such that $\binom{n}{k_0}2^{-\binom{k_0}{2}} \ge 1$. Then if $k = k_0 - 3$ (here $k \sim 2\log_2 n$),

$$\Pr\left(\omega\left(G\left(n, \frac{1}{2}\right)\right) < k\right) = e^{-n^{2-o(1)}}.$$

(This is also Lemma 7.20.) Here's an alternative proof.

Proof. Let Y be the maximum number of **edge-disjoint** sets of k-cliques in G. Y is not 1-Lipschitz with respect to vertex-exposure: for example, my graph could have a bunch of cliques connected to only one point. However, it *is* 1-Lipschitz with respect to edge-exposure (since each edge can only be part of one k-clique anyway).

So now the probability that $\omega(G) < k$ is the probability $Y = 0$ (there are no cliques). Using the lower tail Azuma's inequality,

$$\Pr(Y = 0) = \Pr(Y \le \mathbb{E}[Y] - \mathbb{E}[Y]) \le \exp\left(-\frac{\mathbb{E}[Y]^2}{2\binom{n}{2}}\right).$$

It now remains to show that $\mathbb{E}[Y]$ is large: if we can show that the expected value is $n^{2-o(1)}$, then lower tail estimates tell us that it is very rare for Y to be 0.

So we have this graph $G\left(n, \frac{1}{2}\right)$, and we're asking how many k-cliques we can pack into it. Remember the problem set: the trick is to create an auxiliary graph H whose vertices are k-cliques of G. Then two cliques S, T are adjacent in H if they overlap in at least two vertices, so they have a common edge.

Now $Y = \alpha(H)$ is the size of the largest independent set in H: by Caro-Wei, it suffices to show that H has lots of vertices and not that many edges, since we get the convexity bound

$$\alpha(H) \ge \sum \frac{1}{1 + d(v)} \ge \frac{|V(H)|}{1 + \overline{d}}.$$

By the second moment method, $|V(H)|$, the number of k-cliques in G, is concentrated with high probability around its mean, which is $\binom{n}{k}2^{-\binom{k}{2}}$ by linearity of expectation. By definition of k_0, this is at least $n^{3-o(1)}$, and on the other hand, the expected number of edges in H is concentrated around $\frac{\mathbb{E}[|V(H)|]^2 k^4}{2n^2} = n^{4+o(1)}$. So by Caro-Wei, the expected value of Y is

$$\mathbb{E}[Y] = \mathbb{E}[\alpha(H)] \ge \mathbb{E}\left[\frac{|V(H)|^2}{|V(H)| + 2|E(H)|}\right] \ge n^{2-o(1)},$$

as desired. But here's another way to reach that conclusion, which we've seen a few times: the sampling technique!

Choose a q-random subset of k-cliques in G (each clique with probability q). We'll just get rid of one clique from each overlapping pair to get a large $\alpha(H)$. We expect to get $\mathbb{E}[q|V(H)|]$ cliques, but $\mathbb{E}[q^2|E(H)|]$ overlapping pairs (since H is random as well). Now pick q to maximize: $q = \frac{|V(H)|}{2|E(H)|}$, and this means that the expected size of our independent set is at least $\frac{\mathbb{E}|V(H)|^2}{2\mathbb{E}|E(H)|}$ and we're done. □

9 Concentration of measure

9.1 The geometric picture

Concentration of measure is an important concept in high-dimensional probability and geometry. We've shown examples of concentration of Lipschitz functions of many variables, and it turns out this concept is integrally connected to isoperimetry. For example, what's the largest area we can fence off with a given perimeter? This can be rephrased:

Problem 9.1

Given some constant volume V, what's the minimum possible surface area of that volume?

An example of a space we can work in is the **Hamming cube**: if we have an n-dimensional cube, and we label some number of points, what's the minimum size of the boundary? We can consider the **Hamming distance**, the number of differing coordinates between two vertices of the cube. It seems that we want to take some portion of the cube which is within some Hamming distance of a fixed point (this is a "ball"), which turns out to be true:

Theorem 9.2 (Harper)

If B is a (Hamming) ball, and the volume of A is equal to the volume of B, then the volume of A_t is at least the volume of B_t, where A_t is the set of all points within a distance t from a fixed point A.

What does this have to do with concentration of measure? We can prove an approximate version of Harper's theorem. For n very large, the distribution of Hamming distances looks like a normal distribution with width $\sqrt{n}$, so starting with a Hamming ball with ε area can be thought of as the set of points below $\frac{-t\sqrt{n}}{2}$ on the normal distribution.

Theorem 9.3

For every $\varepsilon > 0$, there exists a $t > 0$ such that for any subset $A \subset \{0,1\}^n$ of the Hamming cube with $|A| \geq \varepsilon 2^n$, $|A_{t\sqrt{n}}| \geq (1-\varepsilon)2^n$.

Proof. We're looking at a hypercube: pick a random vertex x in $\{0,1\}^n$ uniformly, and let X be the distance between x and the closest point in A. By the triangle inequality, this is 1-Lipschitz, and this is informative because $X = 0$ is the same as saying $x \in A$, which happens with probability at least ε. By the Azuma lower tail inequality, the probability that $X = 0$ is

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

This gives an upper bound on the expectation of X: $\mathbb{E}[X] \leq \sqrt{2\log\left(\frac{1}{\varepsilon}\right)n}$, and now we use the upper tail estimate. That tells us that x shouldn't deviate too much: the probability $x \notin A_{t\sqrt{n}}$, where $t = 2\sqrt{2\log\left(\frac{1}{\varepsilon}\right)}$, is

$$\Pr\left(X > 2\sqrt{2\log\left(\frac{1}{\varepsilon}\right)n}\right) \leq \Pr\left(X > \mathbb{E}[X] + \sqrt{2\log\left(\frac{1}{\varepsilon}\right)n}\right) \leq \varepsilon.$$

(Rephrased, our variable is pretty large in expectation, and it is rarely very large.) So x is in $A_{t\sqrt{n}}$ with probability at least $1-\varepsilon$, as desired. □

This is actually a fairly general result, and we can go back and forth between the geometric and combinatorial interpretations of this statement.

Proposition 9.4

Let $t, \varepsilon > 0$ be real numbers, and let Ω be a probability space on which there exists a metric (such as the Hamming cube with the Hamming metric). Then the following are equivalent:

1. (Approximate isoperimetry) For all subsets $A \subset \Omega$ with $\Pr(A) \ge \frac{1}{2}$, then given a set
$$A_t = \{\omega : \text{dist}(\omega, A) \le t\},$$
we have $\Pr(A_t) \ge 1 - \varepsilon$.
2. (Concentration of Lipschitz functions) For all functions $f : \Omega \to \mathbb{R}$ that are 1-Lipschitz – $|f(x) - f(y)| \le \text{dist}_\Omega(x, y)$ – if we have a **median** $m \in \mathbb{R}$ such that $\Pr(f \le m) \ge \frac{1}{2}$ and $\Pr(f \ge m) \ge \frac{1}{2}$, then
$$\Pr(f > m + t) \le \varepsilon.$$

Note that this is concentration around the median, not the mean. We'll soon see that these aren't that different, though.

Proof. First let's show that (1) implies (2). Take the half of the probability space

$$A = \{\omega \in \Omega : f(\omega) \le m\}.$$

This is at least half of our probability space by the definition of the median, and since f is 1-Lipschitz,

$$f(\omega) \le m + t \ \forall \omega \in A_t.$$

Thus,

$$\Pr(f > m + t) \le \Pr(\overline{A_t}) \le \varepsilon$$

by condition (1).

The reverse implication (2) to (1) is not that hard either. We want to show that given any set A with half the space, its t-neighborhood consumes almost the whole space. The natural choice for our Lipschitz function f is the distance

$$f(\omega) = \text{dist}(\omega, A).$$

We pick $m = 0$, and now

$$\Pr(\overline{A_t}) = \Pr(\text{dist}(\omega, A) > t) \le \varepsilon,$$

by condition (2). Now take the complement, $\Pr(A_t) = \Pr(f \le t) \ge 1 - \varepsilon$, and we've shown condition (1). □

This can be useful, because sometimes it's more natural to think in terms of isoperimetry instead of functions (or vice versa).

9.2 Results about concentration: median versus mean

Let's look at another form of concentration of Lipschitz functions:

Proposition 9.5

If we have a 1-Lipschitz function $f : \{0,1\}^n \to \mathbb{R}$ (with respect to the Hamming metric), pick $\omega \sim \text{Unif}(\{0,1\}^n)$, and let $X = f(\omega)$ be our random variable. Then for all $s \in \mathbb{R}, t > 0$,

$$\Pr(X \le s)\Pr(X \ge s+t) \le e^{-t^2/(4n)}.$$

We should think about as taking "either s or $s+t$ to be a median:" then one of the terms becomes a constant $\frac{1}{2}$, and moving that to the other side gives the Gaussian-like tail for the other term.

Proof. We'll apply Azuma's inequality twice. Let μ be the mean of $X - s$ (we can always just shift the variable X so that $s = 0$). If $\mu < 0$, then by Azuma upper tail,

$$\Pr(X \le s)\Pr(X \ge s+t) \le \Pr(X \ge s+t) = \Pr(X - s - \mu \ge t - \mu) \le \exp\left(-\frac{(t-\mu)^2}{2n}\right) \le e^{-t^2/2n}$$

and we're done. Similarly, if $t - \mu < 0$,

$$\Pr(X \le s)\Pr(X \ge s+t) \le \Pr(X \le s) = \Pr(X - s - \mu \le -\mu) \le e^{-\mu^2/2n} \le e^{-t^2/2n},$$

and we're again done. So we're just left with the case where $\mu > 0$ and $t - \mu \ge 0$.

In this case, by Azuma's inequality (lower tail), we can say that

$$\Pr(X \le s) = \Pr(X - s - \mu \le -\mu)$$

and since $X - s - \mu$ is mean-zero and Lipschitz, this is at most $e^{-\mu^2/(2n)}$. On the other hand, by Azuma (upper tail),

$$\Pr(X \ge s+t) = \Pr(X - s - \mu \ge t - \mu) \le e^{(t-\mu)^2/(2n)}.$$

(Be careful here: we can really use this for $t - \mu \ge 0$, but otherwise we can just repeat the argument the other way around by starting with an upper tail argument instead.) Putting these together,

$$\Pr(X \le s)\Pr(X \ge s+t) \le \exp\left[-\frac{\mu^2 + (t-\mu)^2}{2n}\right] \le \exp\left[-\frac{t^2}{4n}\right]$$

by convexity, which is what we want. □

The next few sections are essentially about how to interpret these ideas. One way is to think about $A \subset \{0,1\}^n$ as a subset of the Boolean cube. If $\Pr(A) = \frac{|A|}{2^n}$ (uniform measure), then we have the following:

Corollary 9.6

Consider a uniform measure on the Boolean cube, and let $t > 0$. Then

$$\Pr(A)\Pr(\overline{A_t}) \le e^{-t^2/(4n)}.$$

In particular, if A is at least half the cube, and we expand it by some $c\sqrt{n}$, we get almost the entire cube:

$$\Pr(A_t) \ge 1 - 2e^{-t^2/(4n)}.$$

Earlier in the class, we were using the mean for concentration and other concepts, but now we have the median instead: what relations are there between the mean and median? Suppose that we have a bound of the form

$$\Pr(\overline{A_t}) \le Ce^{-(t/\sigma)^2}$$

for all A with $\Pr(A) \ge \frac{1}{2}$. (In this case, $\sigma \asymp \sqrt{n}$.) Given any 1-Lipschitz function f, and letting $X = f(\omega)$ (where ω is random in Ω), if we have a median m of X, then the difference between the mean and median is

$$|\mathbb{E}[X] - m| \le \mathbb{E}|X - m| = \int_0^\infty \Pr(|X - m| \ge t)dt.$$

If we have sub-Gaussian tail bounds, this is

$$\le \int_0^\infty 2Ce^{-(t/\sigma)^2} = C\sqrt{\pi}\sigma,$$

so the mean and median don't differ by more than a constant times σ. This is actually the tightest bound we can produce: consider the function $f : \{-1, 1\}^n \to \mathbb{R}$ defined by

$$f(x_1, \cdots, x_n) = |X_1 + \cdots + X_n|.$$

We can evaluate its mean and median by the Central Limit Theorem: $\frac{\mathbb{E}[X]}{\sqrt{n}}$ and $\frac{\text{med } X}{\sqrt{n}}$ converge to $\mathbb{E}|Z|$ and $\text{med}|Z|$, where Z is the standard normal – those are different constants! So the idea is that in general, we have

$$\Pr(|X - \mathbb{E}[X]| \ge t) \le C'e^{-(t/\sigma)^2},$$

and we have concentration around the mean as well - just possibly with a worse constant than the median version.

9.3 High-dimensional spheres

Most of our intuition about high-dimensional geometry is wrong! A good reference is Keith Ball's "An elementary introduction to modern convex geometry."

Theorem 9.7 (Isoperimetric inequality in $\mathbb{R}^n$)
If $A, B \subset \mathbb{R}^n$, B is a ball, and $\lambda(A) = \lambda(B)$ (they have the same measure), then

$$\lambda(A_t) \ge \lambda(B_t)$$

for all $t > 0$.

Basically, this is asking for the "smallest perimeter" among all sets of the same volume. We saw a version of this earlier by Harper: if $A, B \subset \{0, 1\}^n$ are subsets of the Hamming cube, $|A| = |B|$ and B is a Hamming ball, then $|A_t| \ge |B_t|$ for all $t > 0$.

Here's the most conceptual way to think about it: this is called **Steiner symmetrization** (alternatively shifting or compression) in the discrete case. The idea is to transform A to preserve the volume and decrease its perimeter. Cut it in half so we have half the volume on each side: if one side has smaller perimeter, then lose the worse side and reflect over the cut. If we can't do this, then every cut must cut the perimeter in two: then just show that such a shape must be a ball.

This isn't actually a proof though: we might need to do this infinitely many times, so there are some compactness issues with this idea. For the discrete setting, we just keep compressing our shape in some direction.

It turns out there's also a spherical isoperimetric inequality:

Theorem 9.8 (Levy)

On the unit sphere $S^{n-1} \subset \mathbb{R}^n$, use the arc distance (though it doesn't really matter). Then given two subsets $A, B \subset S^{n-1}$, where B is some spherical cap and $\lambda(A) = \lambda(B)$, we have $\lambda(A_t) \geq \lambda(B_t)$ for all $t > 0$.

This isn't easy to show, but remember that approximate isoperimetry is connected to concentration of measure. The counterintuitive thing is that distribution of measure is very different in high dimensions than in our ordinary 3-D space.

Fact 9.9

The volume of a t-neighborhood of a hemisphere C is almost everything:

$$\Pr(C_t) \geq 1 - e^{-ct^2 n},$$

where n is the number of dimensions.

We should not think of an n-dimensional shape as a very ball-like object: distribution of mass looks normal along any axis, with standard deviation $\frac{c}{\sqrt{n}}$. Also, most of the mass is near the surface rather than the middle!

Since there's a spherical isoperimetric inequality, we should also have an analogous statement about Lipschitz functions for the sphere:

Proposition 9.10

There exist absolute constants c, C (not dependent on n) such that given a function $f : S^{n-1} \to \mathbb{R}$ that is 1-Lipschitz,

$$\Pr(|f - \mathbb{E}[f]| > t) \leq Ce^{-ct^2 n}.$$

This can be rephrased as "every Lipschitz function is nearly constant nearly everywhere."

For cultural value, here's one more space that's good to mention: the Gauss space in $\mathbb{R}^n$ has the Euclidean metric, and we have the probability distribution

$$\vec{Z} = (Z_1, \cdots, Z_n) : Z_i \sim N(0, 1)$$

(with all Z_is independently distributed). We again have an isoperimetry theorem:

Theorem 9.11

If $A, B \subset \mathbb{R}^n$, and B is a "ball" with $\Pr(A) = \Pr(B)$, then

$$\Pr(A_t) \geq \Pr(B_t) \quad \forall t > 0.$$

A "ball" in Gauss space is intuitively supposed to look like a sphere of radius $\sqrt{n}$, because the probability density function is

$$f_n(x) = (2\pi)^{-n} e^{-|x|^2/2}.$$

The nice thing here is that Gaussian vectors are rotationally symmetric, and now the length of this vector can be written more simply:

$$|\vec{z}|^2 = z_1^2 + \cdots + z_n^2.$$

The expectation of $|\vec{z}|^2$ is n, and along with the spherical isoperimetry inequality, we now have a way to describe balls in the Gauss space: B should be some half-space. (It can have measure not equal to $\frac{1}{2}$ if we don't have the boundary of that half-space passing through the origin.)

9.4 Projections onto subspaces

The idea with our next section is that we want to represent a bunch of points in a smaller dimension without distorting the distances too much.

Theorem 9.12 (Johnson-Lindenstrauss Lemma)
Let $s_1, \cdots, s_N$ be points in $\mathbb{R}^d$. Then there exist $s_1', s_2', \ldots, s_N' \in \mathbb{R}^m$, where $m = O(\varepsilon^{-2} \log N)$, such that

$$(1-\varepsilon)|s_i - s_j| \le |s_i' - s_j'| \le (1+\varepsilon)|s_i - s_j|.$$

So we can approximately preserve distances up to a small multiplicative error.

Proof. Pick a random (orthogonal) projection onto an m-dimensional subspace (chosen uniformly at random). This projection is actually agnostic to the set of points $s_1, \cdots, s_N$. We claim that with positive probability, the desired outcome occurs. If we do this naively, everything gets smaller, so we'll scale by $\sqrt{\frac{n}{m}}$ to correct for that. Basically, our claim is that all the length ratios are generally preserved.

Lemma 9.13
Let P be a projection from $\mathbb{R}^n$ onto a random m-dimensional subspace, and let $z \in \mathbb{R}^n$ be some fixed vector. If we let $z' = Pz$ be a random variable, then $\mathbb{E}[|z'|^2] = \frac{m}{n}|z|^2$, and we have

$$(1-\varepsilon)\sqrt{\frac{m}{n}}|z| \le |z'| \le (1+\varepsilon)\sqrt{\frac{m}{n}}|z|$$

with probability at least $1 - e^{-c\varepsilon^2 m}$.

Proof of lemma. Note that fixing our vector and picking a random subspace is equivalent to fixing our projection P and choosing a random unit vector in $\mathbb{R}^n$. By rotational symmetry, we can make P the span of the first m basis elements $\{e_1, \cdots, e_m\}$, and thus any $z = (z_1, \cdots, z_n)$ corresponds to $z' = (z_1, \cdots, z_m, 0, \cdots, 0)$.

Note that $\mathbb{E}[|z'|^2] = \mathbb{E}[z_1^2 + \cdots + z_n^2]$: in general, it's easier to look at squared lengths than lengths. By symmetry, because all the z_i^2s have the same expectation, each z_i^2 has expected value $\frac{1}{n}$ (using linearity), so

$$\mathbb{E}|z'|^2 = \mathbb{E}[z_1^2 + \cdots + z_m^2] = \frac{m}{n}.$$

But now note that projection $z \to |z'|$ is a 1-Lipschitz function, so by Levy concentration (the isoperimetric inequality on the sphere), we know that

$$\Pr\left(\left||z'| - \sqrt{\frac{n}{m}}|z|\right| > \varepsilon\sqrt{\frac{n}{m}}\right) \le \exp\left[-cn \cdot \frac{m}{n}\varepsilon^2\right] = \exp[-m\varepsilon^2].$$

(remember that mean and median are reasonably close because of Gaussian tails), which is exactly what the lemma claims. □

So now we can finish with a union bound: since everything happens with high probability, we can just say that the probability some pair (i, j) fails the distance check is at most

$$\leq N^2 e^{-c\varepsilon^2 m} < 1$$

as long as m is chosen to be $O(\varepsilon^{-2} \log N)$, and we have the desired result. □

9.5 What if we need stronger concentration?

Unfortunately, Azuma's inequality is not enough to solve all of our problems. Consider the following:

Problem 9.14

Let V be a fixed d-dimensional subspace (through the origin), and we pick a point $X \sim \text{Unif}(\{-1, 1\}^n)$. How well is the Euclidean distance $\text{dist}(x, V)$ concentrated?

We have n independent Boolean variables, so we have a Lipschitz function of x, which gives $\sqrt{n}$ concentration of our random variable X by Azuma's inequality. In particular, the probability that our variable is within $t\sqrt{n}$ of its mean decays like a Gaussian in t.

But the diameter of the cube itself is proportional to $\sqrt{n}$, so this is a pretty bad estimate!

Note that we can change the problem a bit, and Azuma does become pretty good. Specifically, if we pick V uniformly at random (then x can be either a fixed point or chosen randomly - it doesn't matter), Azuma's gives pretty good concentration. Alternatively, we could pick x uniformly at random from the sphere that goes through the vertices of the Boolean cube as well, and Azuma still yields reasonable concentration. These are the same by rotational symmetry, and in the second case, we're asking for the concentration of a Lipschitz function on a sphere. We know then that if f is Lipschitz on a $\sqrt{n}$-radius n-dimensional sphere, then

$$\Pr(|f - \mathbb{E}[f]| > t) \leq Ce^{-ct^2}.$$

So if we can get $O(1)$ concentration on the sphere, intuitively we should also be able to get it on the Boolean cube as well. We just haven't been able to do this with the methods introduced so far.

9.6 Talagrand's inequality: special case

As often happens with Euclidean distances, it's hard to calculate the mean of X, but analyzing X^2 is much easier. Let P be the projection operation onto the orthogonal complement of V, our d-dimensional subspace: then P is some matrix $\in \mathbb{R}^{n \times m}$, and we have

$$X^2 = \langle X, PX \rangle = \sum_{ij} x_i x_j p_{ij}.$$

Since the x_is are orthonormal, this just leads us to

$$\mathbb{E}[X^2] = \sum_i p_{ii} = \text{tr} P = n - d.$$

Notably, this expectation of X^2 does not depend on the orientation of V, though the distribution of X^2 does. So we should expect X to be concentrated about $\sqrt{n-d}$, and that gives us a center to work with. We're trying to claim that we have $O(1)$-concentration; specifically, we'd like to show that there is exponential decay with a constant deviation. To do this, we finally introduce the inequality we want:

Theorem 9.15 (Talagrand's inequality, simplified)

Let $A \subset \mathbb{R}^n$ be a **convex subset**, and let x be a uniform random point in the Boolean cube

$$x \sim \text{Unif}(\{0,1\}^n).$$

Then for all $t > 0$,

$$\Pr(x \in A) \Pr(\text{dist}(x, A) \geq t) \leq e^{-t^2/4},$$

where we use the Euclidean distance.

Convexity here is extremely important. Talagrand is just not true otherwise - for example, consider A to be just the set of points in $\{0,1\}^n$ with weight (sum of entries) at least $\frac{n}{2}$. Then a random vertex is generally $O(\sqrt{n})$ away: specifically, there is probability at least $\frac{1}{4}$ that the weight of x is at most $\frac{n}{2} - c\sqrt{n}$ for some c.

Then the Euclidean distance is the square root of the Hamming distance on the Boolean cube, so the distance from x to A is on the order of $n^{1/4}$, which is not constant.

So what's really going on with this inequality? Given a convex set - for example, the convex hull of those same points in our Boolean cube - we're now measuring the distance to possibly some convex average of our vertices, and that distance is generally much smaller than if we were only allowed to use the vertices themselves.

Definition 9.16

Define a function f to be **quasi-convex** if all sets $\{f \leq a\}$ for $a \in \mathbb{R}$ are convex. (All convex functions are quasi-convex as well.)

Corollary 9.17

Let's say we have a function $f : \mathbb{R}^n \to \mathbb{R}$ that is quasi-convex and 1-Lipschitz with respect to the Euclidean distance: then for all $r \in \mathbb{R}, t > 0$, for x picked uniformly from the cube $\{0,1\}^n$,

$$\Pr(f(x) \leq r) \Pr(f(x) \geq r + t) \leq e^{-t^2/4}.$$

This is a direct translation of the isoperimetric inequality. The theorem implies the corollary by letting A be the set of values $\{f \leq r\}$, which is convex if f is quasi-convex by definition. Since f is 1-Lipschitz by the triangle inequality, we have

$$f(x) \leq r + t \quad \forall x : \text{dist}(x, A) \leq t.$$

With this, we're now ready to answer our initial problem:

Theorem 9.18

Let V be a fixed d-dimensional subspace, and let $f(x) = \text{dist}(x, V)$. If we pick x uniformly on the cube $\{0,1\}^n$, then there exist constants $C, c > 0$ such that for all $t > 0$,

$$\Pr(|f - \mathbb{E}[f]| > t) \leq Ce^{-ct^2}.$$

Proof sketch. Let m be a median of f. Using Corollary 9.17, set $r = m$ to get the upper tail

$$\Pr(f \geq m + t) \leq 2e^{-t^2/4}.$$

Meanwhile, set $r = m - t$ to get the lower tail

$$\Pr(f \le m - t) \le 2e^{-t^2/4}.$$

We also mentioned that the median and the mean are very close for sub-Gaussian distributions. With some calculations, we can show that the median of X

$$\text{med}(X) = \sqrt{n-d} + O(1)$$

(with an absolute constant), or else we get inconsistency with tail bounds. Since $\mathbb{E}[f^2] = n - d$ and we have constant concentration, $\mathbb{E}[f]$ is also $\sqrt{n-d} + O(1)$, and thus we have constant deviation from the mean, as desired. □

So the whole point is that Talagrand is about **concentration of convex Lipschitz functions when evaluating at a random point of the Boolean cube.** We're not going to prove the inequality in class, because there are some tedious calculations involved. Instead, let's focus on combinatorial applications.

9.7 Random matrices

Let A be a random symmetric matrix with independent entries ±1, where $a_{ij} = a_{ji}$. This can be thought of as being related to the adjacency matrix of a random graph.

It turns out the largest eigenvalue λ_1 is also the operator norm of A:

$$\lambda_1(A) = ||A||_{\text{op}}.$$

How well is this concentrated? We have about $O(n^2)$ variables, so Azuma's inequality gives something like $O(n)$ concentration about the mean. But this is pretty bad, because typically the largest eigenvalue

$$\lambda_1(A) \lesssim \sqrt{n},$$

so linear concentration doesn't really help at all. On the other hand, let's try to use Talagrand's inequality. We need to check a few things: consider the function $f : A \to ||A||_{\text{op}}$.

- Convexity comes from the fact that the operator norm is a norm, so we can use the triangle inequality.
- To show this function is 1-Lipschitz, we need

  $$|f(x) - f(y)| \le ||x - y||_2,$$

  where we're using the L^2 norm. This can be proved using Cauchy-Schwarz.

So now Talagrand's inequality tells us that we have constant-window concentration, independent of n. In other words, we've just showed that

$$\Pr(|\lambda_1(A) - \mathbb{E}(\lambda_1(A))| \ge t) \le Ce^{-ct^2}$$

for some C, c, which decays like a Gaussian.

Fact 9.19

We actually know more about the concentration: it's actually $\Theta(n^{-1/6})$, and it converges to something called a Tracy-Widom distribution when normalized. Also, we know the mean of this distribution: the easiest way is to make the entries Gaussian instead of ±1, but the answer is approximately $\sqrt{2n}$ regardless of the distribution.

As a sidenote, we can't actually use this method to prove the concentration of the second largest eigenvalue yet, since that's not convex as a function of our matrix entries. But the bottom line is that Talagrand's inequality is not just about the Boolean cube.

9.8 Talagrand's inequality in general

If we have a space $\Omega = \Omega_1 \times \cdots \times \Omega_n$, we may want to find the distance between two points.

Definition 9.20

Given a vector $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n_{\geq 0}$, define the **weighted Hamming distance**

$$d_\alpha(x, y) = \sum_{i: x_i \neq y_i} \alpha_i.$$

This kind of distance is defined even if the individual Ω_is don't have metrics! So now if we have some subset $A \subset \Omega$ of our product space, we can define

$$d_\alpha(x, A) = \inf_{y \in A} d_\alpha(x, y).$$

This should still feel fairly familiar: for example, for $\Omega = \{0, 1\}^n$, if we have a fixed α with $|\alpha| = 1$ (under the L^2 norm), we have

$$d_\alpha(x, y) = |\langle \alpha, 1_{x \neq y} \rangle|.$$

Azuma's inequality then tells us that for if we choose x uniformly on $\{0, 1\}^n$, if we have a **fixed** α and subset $A \subset \{0, 1\}^n$,

$$\Pr(|d_\alpha(x, A) - \mathbb{E}(d_\alpha(x, A))| \geq t) \leq 2e^{-t^2/2}.$$

(This is because the weights in Azuma's inequality satisfy $\sum c_i^2 = 1$ from the definition of α.) But Azuma only gives this to us for a fixed α: having this condition be true in all directions is much stronger, and that's what Talagrand's inequality tells us.

Definition 9.21

Define the **convex distance**

$$d_T(x, A) = \sup_{\substack{\alpha \in \mathbb{R}^n \\ |\alpha| = 1}} d_\alpha(x, A).$$

(Basically, choose the "worst" possible α: the one that separates x and A by the most.) This is easier to visualize if we think of $\Omega = \{0, 1\}^n$ and $A \subset \Omega$ being a subset of that Boolean cube: $d_T(x, A)$ is then just the Euclidean distance from x to the convex hull of A. (In general, we do not need each coordinate to be limited to $\{0, 1\}$, though: they can take on any set of values.)

Theorem 9.22 (Talagrand's inequality, general)

For any $A \subset \Omega = \Omega_1 \times \cdots \times \Omega_n$, let x be a random point in Ω with independent coordinates. Then

$$\Pr(x \in A) \Pr(d_T(x, A) \geq t) \leq e^{-t^2/4}.$$

Here's another interpretation of the convex distance: we want to convert this to a "distance to convex hull" type argument even when we don't have a Boolean cube.

Definition 9.23

Define $U_A(x)$ to be the set of $s \in \{0,1\}^n$ such that there is some $y \in A$ so that $s_i = 1$ for all i with $x_i \neq y_i$. In other words, $s \in U_A(x)$ if the support of s contains the support of $x - y$ for some $y \in A$.

We can think of this as the "set of coordinates we need to change to get from x to A," ignoring the actual coordinates: notice that this is a subset of the Boolean cube even when A isn't. Notably, this is an increasing subset of the cube $\{0,1\}^n$.

Lemma 9.24

Letting dist be the Euclidean distance,

$$d_T(x, A) = \text{dist}(\vec{0}, \text{convex hull}(U_A(x))),$$

Proof. The left hand side is (by definition) the supremum over all weight vectors α of norm 1 of the α-distance between x and A, which is equivalent to looking at the closest point in A under this α: this is then

$$\sup_\alpha \inf_{y \in A} d_\alpha(x, y) = \sup_\alpha \inf_{y \in U_A(x)} (\alpha \cdot y).$$

Since A is convex, by von Neumann, we can swap the inf and sup as long as we extend to the convex hull

$$= \inf_{\substack{y \in \text{convex} \\ \text{hull}(U_A(x))}} \sup_\alpha (\alpha \cdot y) = \inf_{\substack{y \in \text{convex} \\ \text{hull}(U_A(x))}} |y|,$$

as desired. □

So how do we apply Talagrand? The idea is that we can adjust our α to favor certain coordinates and give us better bounds. α plays the role of a "certificate" that guarantees the existence of a small or large value. In particular, it follows from Talagrand that (rearranging)

$$\Pr(d_{\alpha(x)}(x, A) \geq t) \leq \frac{1}{\Pr(A)} e^{-t^2/4}.$$

Specifically, we can pick a different certificate α for each x:

Corollary 9.25

Let A, B be subsets of $\Omega = \Omega_1 \times \Omega_2 \cdots \times \Omega_n$. Suppose that for all $y \in B$, there exists an $\alpha = \alpha(y) \in \mathbb{R}^n_{\geq 0}$ so that for all $x \in A$,

$$d_\alpha(x, y) \geq t|\alpha|.$$

(This means the distance between A and B is large in the specific Talagrand sense.) Then

$$\Pr(A)\Pr(B) \leq e^{-t^2/4}.$$

To understand this, let's do another proof of the largest eigenvalue of our random matrix:

Proof. Let X be an $n \times n$ symmetric random matrix with independent entries in the interval $[-1, 1]$ (they can be distributed in any way, as long as they are independent and bounded). If we let $t > 0$, $M \in \mathbb{R}$, and we have the sets

$$A = \{X : \lambda_1(X) \leq M\}, B = \{X : \lambda_1(X) \geq M + t\},$$

we want to verify that for every matrix in B with large eigenvalue, we can certify this somehow: we pick some $\alpha(y)$ such that B is far away from A. Specifically, there exists some $\alpha \in \mathbb{R}^m$, where $m = \frac{n(n+1)}{2}$, such that $d_\alpha(x, y) \geq ct|\alpha|$ for all $x \in A$.

Let $\vec{v} \in \mathbb{R}^n$ be the top eigenvector corresponding to $\lambda_1(y)$. Then let

$$\alpha_{ij} = \begin{cases} v_i^2 & i = j \\ 2|v_i||v_j| & i \neq j; \end{cases}$$

the reason for doing this will become quickly apparent. By the Courant-Fischer characterization of the top eigenvector $\vec{v}$,

$$v^T Y v = \lambda_1(Y) \geq M + t,$$

and because X does not have large eigenvalue, we can set a contrasting bound for X:

$$v^T X v \leq \lambda_1(X) \leq M.$$

In particular, this means that we can use our eigenvector v to "separate" A and B:

$$t \leq v^T (X - Y) v,$$

and expanding out the difference as a bilinear form,

$$t \leq \sum_{i,j} v_i v_j (X_{ij} - Y_{ij}).$$

This is upper bounded by looking at only those where the two matrices differ in their entries:

$$\leq 2 \sum_{ij} |v_i||v_j| 1_{X_{ij} \neq Y_{ij}} \leq 2 d_\alpha(X, Y).$$

(Here, we used that the entries of X and Y are bounded by $[-1, 1]$.) Now note that the length of α is at most 2, and plug this into the corollary to get concentration. □

9.9 Increasing subsequences

Problem 9.26

Pick a uniformly random permutation $\sigma \in S_n$ of the first n integers. How long is the longest increasing subsequence?

Call this length X - it's important to note that we can skip entries (so subsequences don't need to be contiguous). For example, $5, 3, 1, 4, 6, 2, 7$ has longest increasing subsequence of length 4.

Our goal is to show that X is concentrated. Let's try to use the tools we have: first of all, let's try Azuma. We need independence of our underlying variables, so let's try to make our Ω_is independent: let $x_1, \cdots, x_n \sim \text{Unif}[0, 1]$ independently, and get a permutation of $[n]$ from the **relative orderings** of the x_is.

Then the length of the longest increasing subsequence changes by at most 1 if we change 1 coordinate, so it is 1-Lipschitz here. Azuma tells us that we have sub-Gaussian decay with a window size of $O(\sqrt{n})$.

How good is this? Let's do a first moment calculation to see the average size of X:

$$\Pr(X \geq k) \leq \binom{n}{k} \cdot \frac{1}{k!} \leq \frac{n^k}{(k!)^2},$$

since we pick any of the $\binom{n}{k}$ sequences of length k, and they have probability $\frac{1}{k!}$ of working.

So if $k = 100\sqrt{n}$, then this probability is $o(1)$, and thus we should expect the permutation to be no more than $c\sqrt{n}$ long. That means our concentration bound is bad! (In particular, any permutation of length n has either an increasing or decreasing sequence of length $\sqrt{n}$ by Pigeonhole.)

Let's see if Talagrand tells us anything better. The idea is that Talagrand is useful when we can "witness" rare events: showing such a sequence exists (that is, making the length **certificable**) doesn't use that many of the coordinates of $\vec{x}$. So Talagrand will actually tell us that we have fluctuations on the order of $O(\sqrt{x})$.

Here's that idea in more rigor:

Theorem 9.27

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$, and let $f : \Omega \to \mathbb{R}$ be a 1-Lipschitz function with respect to the Hamming distance. Suppose that we can verify

$$\{\omega : f(\omega) \geq r\}$$

by checking at most s coordinates. Then for every t,

$$\Pr(f(\omega) \leq r - t\sqrt{s}) \Pr(f(\omega) \geq r) \leq e^{-t^2/4}.$$

When we say **checking at most** s **coordinates** here, we specifically mean that with any ω with $f(\omega) \geq r$, there exists some subset $I \subset [n]$ with $|I| \leq S$ such that for all other ω' such that ω agrees with ω' on I, $f(\omega') \geq r$. In other words, knowing those s coordinates guarantees that our condition is true.

Proof. Let A, B be the sets

$$A : \{\omega : f(\omega) \leq r - t\sqrt{s}\}, B : \{\omega : f(\omega) \geq r\}.$$

Our goal is to check that for all $y \in B$, there exists $\alpha \in \mathbb{R}^n_{\geq 0}$ such that for all $x \in A$, $d_\alpha(x, y) \geq t|\alpha|$. (Basically, A is far away from B, even if we zoom in on I.)

But by definition of checking coordinates, there exists a set $I \subset [n]$ with $|I| \leq s$ for each $y \in B$. Here's the key: if we fix y and let $\alpha = 1_I$ (1 in the spots of I and 0 in the others), every $x \in A$ disagrees with y on at least $t\sqrt{s}$ coordinates of I (or else we could change x by less than $t\sqrt{s}$ coordinates and get x' to agree with y on I, meaning $f(x') \geq r$). This means that the weighted Hamming distance $d_\alpha(x, y) \geq t\sqrt{s} \geq t|\alpha|$, and now we can apply Talagrand's directly. □

Corollary 9.28

Given a 1-Lipschitz function on $f : \Omega = \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ (with respect to the Hamming distance), if $\{f \geq r\}$ can be verified by checking r coordinates, and m is a median of X, then for all t,

$$\Pr(X \leq m - t) \leq 2\exp\left(-\frac{t^2}{4m}\right),$$

$$\Pr(X \geq m + t) \leq 2\exp\left(-\frac{t^2}{4(m+t)}\right).$$

Proof. From the above theorem, renormalize t by a factor of $\sqrt{r}$: now we have

$$\Pr(f \le r - t) \Pr(f \ge r) \le e^{-t^2/(4r)}.$$

Setting $r = m$ gives the lower bound, and setting $R = m + t$ gets the lower bound. $\square$

So now this applies directly to X for our increasing subsequence, since we can "witness" our event by just showing the subsequence itself. Note also that the median of X is $O(\sqrt{n})$, so now we know that

$$\Pr(|X - \mathbb{E}[X]| < s) = 1 - o(1) \text{ if } s \gg n^{1/4},$$

meaning we've found $\sqrt[4]{n}$-concentration.

But this is not the best possible result! In 1985, Vershik-Kerov showed that X is concentrated around $2\sqrt{n}$, and in fact, the limiting distribution was found by Baik-Deift-Johansson in 1999 to be

$$\frac{X - 2\sqrt{n}}{n^{1/6}} \to \text{Tracy-Widon distribution.}$$

(As we may remember, this is also the fluctuation of the top eigenvalue of a random matrix.)

10 Entropy methods

10.1 Information entropy

We're going to shift away from concentration results now. This next concept was essentially invented by Shannon, and we'll focus on its combinatorial applications.

Definition 10.1

Let X be a discrete random variable taking values in some set S. Then the **entropy** of X is

$$H(X) = \sum_{s \in S} -p_s \log_2 p_s,$$

where $p_s = \Pr(X = s)$.

Intuitively, entropy is supposed to measure the amount of randomness or information in the random variable X.

Because we're doing combinatorics, we'll work with base-2 logarithms - this is really more of a convention than anything else, and all logs in this section mean base 2.

Example 10.2

The entropy of a Bernoulli variable $\mathrm{Ber}(p)$ is just $-p \log_2 p - (1-p)\log_2(1-p)$, which has a maximum of 1 at $p = \frac{1}{2}$.

Basically, this tracks how "surprised" we are when we hear a sample from the distribution. This idea essentially comes from trying to encode messages efficiently: for example, if a coin only comes up heads 1% of the time, encoding it as a binary string directly is not the most efficient way.

Lemma 10.3

$H(X) \le \log_2 |\mathrm{range}(X)|$.

Proof. This is convexity of the function $x \to x \log_2 x$. □

Equality holds when we have the uniform distribution: then $H(X)$ tells us the number of binary bits needed to specify which choice of X we pick out.

Denote by $H(X, Y)$ the entropy of the joint random variable $Z = (X, Y)$, where X and Y are not necessarily independent. This means we have

$$H(X, Y) = \sum_{(x,y)} -\Pr(X = x, Y = y) \log_2 \Pr(X = x, Y = y).$$

Lemma 10.4 (Subadditivity)

Given any two random variables X, Y, $H(X, Y) \le H(X) + H(Y)$.

Proof. Expanding $H(X) + H(Y) - H(X, Y)$ out, this gives

$$H(X) + H(Y) - H(X, Y) = \sum_{x,y} (-p(x, y) \log_2 p(x) - p(x, y) \log_2 p(y) + p(x, y) \log_2 p(x, y)) \quad = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}.$$

Let $f(t) = t \log t$, which is a convex function. Then by Jensen's, we can bound this as

$$= \sum_{x,y} p(x)p(y) f\left(\frac{p(x,y)}{p(x)p(y)}\right) \ge f(1) = 0.$$

□

Basically, there's at least as much information in X and Y individually as when we put them together.

$$H(X) + H(Y) - H(X,Y) = I(X,Y)$$

is called the **mutual information**, and it's always nonnegative.

In particular, if X and Y are independent, then $H(X,Y) = H(X) + H(Y)$. In this case, the amount of information in our variable X is just the sum of the individual parts.

Corollary 10.5

For any random variables $X_1, \cdots, X_n$,

$$H(X_1, \cdots, X_n) \le H(X_1) + \cdots + H(X_n).$$

There's also a notion of "conditional entropy:" let E be an event with positive probability, and then we have

$$H(X|E) = \sum_x -\Pr(X = x|E) \log_2 \Pr(X = x|E).$$

What's really important to us, though, is when we condition on a second random variable: if X and Y are jointly distributed, we define

$$H(X|Y) = \mathbb{E}_y\left[H(X|Y = y)\right].$$

Essentially, this is how much new information we get given a certain piece of information about Y.

Lemma 10.6 (Chain rule)

For any random variables X, Y, $H(X|Y) = H(X,Y) - H(Y)$.

Proof.

$$\begin{aligned}
H(X|Y) &= \mathbb{E}_y\left[H(X|Y = y)\right] \\
&= \sum_y \Pr(Y = y) H(X|Y = y) = \sum_y -p(y) \sum_x p(x|y) \log_2 p(x|y) \\
&= \sum_{x,y} -p(x,y) \log_2 p(x,y) + \sum_{x,y} p(x,y) \log_2 (y) \\
&= \sum_{x,y} -p(x,y) \log_2 p(x,y) + \sum_y p(y) \log_2 (y),
\end{aligned}$$

where the first equality follows from Bayes' rule and the last because $\sum_x p(x,y) = p(y)$.

□

In other words, the conditional entropy is just the total entropy minus what we "already knew about Y." In particular, if $X = Y$, or if $X = f(Y)$ (so we know X given Y), the conditional entropy is 0. On the other hand, if X and Y are independent, the conditional entropy is just $H(X)$.

Lemma 10.7 (Dropping conditioning)

For any random variables X, Y, Z, $H(X|Y) \le H(X)$ and $H(X|Y,Z) \le H(X|Z)$.

Proof. These follow from the chain rule (Lemma 10.6) and subadditivity (Lemma 10.4). For example,

$$H(X|Y) = H(X,Y) - H(Y) \le H(X).$$

□

10.2 Various direct applications

Let's start to see how this can be useful! Entropy's use primarily comes up in tail bounds. Here's a philosophy: we want to show an upper bound on some quantity, so we start by taking the log of both sides. The left side is the log of some quantity, so take a uniform probability distribution on the things we want to count: we now have an entropy.

Theorem 10.8

Let $\mathcal{F}$ be a collection of subsets of $[n]$, and let p_i be the fraction of subsets in $\mathcal{F}$ that contain the element i. Then

$$\log_2 |\mathcal{F}| \le \sum_{i=1}^{n} H(p_i),$$

where $H(p)$ is the binary entropy of the Bernoulli variable

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p).$$

Proof. Let $X = (X_1, \cdots, X_n)$ be the characteristic vector for a uniform random element $F \in \mathcal{F}$: this means that X_i is 1 if $i \in F$ and 0 otherwise. The entries aren't necessarily independent here, so we can play with this with entropy: $\log_2 |\mathcal{F}|$ is just $H(X)$, because we have a uniform distribution (this is the equality case of Lemma 10.3).

By subadditivity, this is at most $H(X_1) + \cdots + H(X_n)$. Each X_i is a Bernoulli random variable with probability p_i, which is what we want. □

Theorem 10.9

Let $k \le \frac{n}{2}$. Then

$$\sum_{0 \le i \le k} \binom{n}{i} \le 2^{H\left(\frac{k}{n}\right)n}.$$

Proof. Let $X = (X_1, \cdots, X_n) \in \{0,1\}^n$ be the uniform random vector conditioned on $X_1 + \cdots + X_n \le k$. By Lemma 10.3, the logarithm of the left hand side is $H(X)$, and by subadditivity, this is at most $H(X_1) + \cdots + H(X_n)$. Conditioning on the sum of the X_i being exactly m, each X_i is a Bernoulli variable with probability $\frac{m}{n}$. Since the sum is always at most k, we can say that X_i is Bernoulli with probability at most $\frac{k}{n}$. Since this is less than $\frac{1}{2}$ by assumption and the entropy of a Bernoulli increases until $p = 1/2$, we have that $H(X_i) \le H\left(\frac{k}{n}\right)$.

Now there are n copies of this term, and rearranging gives the result. □

We get a similar result if we don't pick everything with probability $\frac{1}{2}$ but instead with probability p: then we get a relative entropy called the Kullback-Leibler divergence.

Theorem 10.10 (This was problem 32 from our problem set)
Let $S_1, \cdots, S_k$ be subsets of $[n]$, and suppose that for every pair of distinct subsets $A, B \subseteq [n]$, there exists an i such that

$$|S_i \cap A| \neq |S_i \cap B|.$$

Then $k \geq (2 - o(1))\frac{n}{\log_2 n}$.

This is called a coin weighing problem, because we can imagine that we have two types of coins, where one is a little heavier than the other. We can then weigh k times, and we want to be able to tell how many counterfeit coins we have. Well, if there always exists an i that distinguishes them, then we know exactly which coins we want. It turns out we need at least $\approx \frac{2n}{\log_2 n}$ weighings to do the job.

The main idea here is that there's some information that we're gaining on each comparison S_i: can we get enough to deduce the set of coins?

Proof. Let X be a uniform random subset of $[n]$. Since there are 2^n different possibilities that are uniformly weighted, the entropy of X is just n. Observe that X contains the same information as the sizes of all $|X \cap S_i|$ for $1 \leq i \leq k$: in particular, this is an injective map, since no two subsets have the same set of intersections. By subadditivity,

$$H(X) = H(|X \cap S_1|, \cdots, |X \cap S_k|) \leq H(|X \cap S_1|) + \cdots + H(|X \cap S_k|).$$

Because X is a uniform subset of 1 through n, $|X \cap S_i|$ is binomial with distribution $\text{Bin}\left(|S_i|, \frac{1}{2}\right)$. The entropy of such a binomial distribution is bounded by $\log_2 |S_i|$, and $|S_i| \leq n$, so this gives

$$n = H(X) \leq k \log_2 n,$$

which is enough to give everything except for the factor of 2.

However, note that the binomial distribution is not uniform: it's highly concentrated, and thus we should have much less entropy than a uniform distribution! Heuristically, we know that the binomial distribution is concentrated in a $\sqrt{|S_i|}$-interval, so the entropy should be essentially related to $\log_2(|S_i|)$. This turns out to be true if we work out the calculations, and that gives us

$$H\left(\text{Bin}\left(|S_i|, \frac{1}{2}\right)\right) \leq \left(\frac{1}{2} + o(1)\right) \log_2 m$$

and now rearranging gives the result that we want. □

As a sidenote, the actual entropy of the Binomial distribution is $\frac{1}{2}\log_2 m + O(1)$.

10.3 Bregman's theorem

Definition 10.11

The **permanent** of an $n \times n$ matrix is

$$\text{per } A = \sum_{\sigma \in S_n} \prod a_{i,\sigma_i}.$$

In contrast, the **determinant** is similar but includes a sign:

$$\det A = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod a_{i,\sigma_i}.$$

These are very different quantities - in particular, the determinant is believed to be much easier to calculate.

Let's only consider matrices $A \subset \{0,1\}^{n \times n}$: any such matrix can be encoded by a bipartite graph with n row nodes and n column nodes, where row i and column j are connected if and only if there's a 1 in the corresponding entry.

Lemma 10.12

The permanent of a matrix $A \subset \{0,1\}^{n \times n}$ is equal to the number of perfect matchings in the corresponding bipartite graph.

Proof. The permanent expands over all permutations, and we count a permutation if and only if every edge we try to use exists (giving us a product of 1). □

So here's a natural question to ask: if we have some degree distribution (for example, d-regular), what is the maximum number of perfect matchings that are possible? One possible extremal graph is a union of complete bipartite graphs: in the d-regular case, the number of perfect matchings is just $d!$ to some power. Is this the best we can do in general?

Theorem 10.13 (Bregman)

Given a matrix $A \in \{0,1\}^{n \times n}$ whose i^{th} row sums to d_i for all i,

$$\text{per } A \le \prod_{i=1}^{n} (d_i!)^{1/d_i}.$$

Note that a disjoint union of complete bipartite graphs $K_{s,s}$ gives the equality case.

Proof by Radhakrishnan. Let σ be a uniform permutation of $[n]$, conditioned on all A_{i,σ_i} being 1 in our matrix. In other words, we are picking a uniform random perfect matching! By Lemma 10.12, $H(\sigma) = \log_2(\text{per} A)$.

Attempt 1 (Subadditivity): σ has n different coordinates, one for the entry σ_i picked in each row, so each coordinate is a random variable. As we've done in the previous examples, we can try to apply subadditivity here, bounding the entropy for the i^{th} coordinate by $H(\sigma_i)$. If there are d_i 1s in that row, we can say that $H(\sigma_i) \le \log_2 d_i$ (we may not have equality because we don't have a uniform distribution on the σ_is). Unfortunately, this is not enough! Because the σ_i are not chosen independently (e.g. picking something in the first row affects the others), applying subadditivity directly costs us a lot.

Attempt 2 (Randomization + chain rule): Instead, let's reveal the rows in a uniform random order and then apply the chain rule. If τ is a uniform permutation in S_n, we now have

$$H(\sigma) = H(\sigma_{\tau_1}) + H(\sigma_{\tau_2} | \sigma_{\tau_1}) + \cdots + H(\sigma_{\tau_n} | \sigma_{\tau_1}, \cdots, \sigma_{\tau_{n-1}}).$$

Take expectations on both sides. We know that the left hand side is independent of the ordering - at the end of the process, we still see all the rows, so the information we get is the same. Thus,

$$H(\sigma) = \mathbb{E}[H(\sigma_{\tau_1})] + \mathbb{E}[H(\sigma_{\tau_2}|\sigma_{\tau_1})] + \cdots + \mathbb{E}[H(\sigma_{\tau_n}|\sigma_{\tau_1}, \cdots, \sigma_{\tau_{n-1}})].$$

What's the contribution of the i^{th} row of our original matrix to this sum? If the row appears in the k^{th} term of the sum, then it contributes $\mathbb{E}[H(\sigma_i|\cdots)]$, where $\cdots$ represents a uniform subset of $k-1$ other rows. Then,

$$\mathbb{E}[H(\sigma_i|\cdots)] \leq \mathbb{E}[\log_2(\text{number of available entries in row } i|\cdots)].$$

Since we only care about the ordering of the d_i rows whose entries conflict with the d_i 1s in row i, and each ordering is equally likely,

$$= \frac{1}{d_i}(\log_2 1 + \log_2 2 + \cdots + \log_2 d_i) = \frac{1}{d_i}\log_2(d_i!).$$

Plugging this back into the sum,

$$\mathbb{E}[H(\sigma)] \leq \sum_{i=1}^{n} \frac{1}{d_i}\log_2(d_i!),$$

and exponentiating both sides yields the result. □

10.4 A useful entropy lemma

Lemma 10.14 (Shearer's lemma (special))

For any random variables X, Y, Z,

$$2H(X,Y,Z) \leq H(X,Y) + H(X,Z) + H(Y,Z).$$

Proof. By the chain rule,

$$H(X,Y,Z) = H(X) + H(Y|X) + H(Z|X,Y).$$

Now, add up the following:

$$H(X,Y) = H(X) + H(Y|X)$$
$$H(X,Z) = H(X) + H(Z|X)$$
$$H(Y,Z) = H(Y) + H(Z|Y).$$

Dropping conditioning on $H(X,Y,Z)$ yields the result. □

What are some applications of this?

Corollary 10.15

Given a finite set $S \subset \mathbb{R}^3$, consider the orthogonal projections $\pi_{xy}(S)$ onto the xy-plane (and similarly for the xz and yz-planes). We have

$$|S|^2 \leq \pi_{xy}(S)\pi_{xz}(S)\pi_{yz}(S).$$

Equality holds for a Cartesian box.

Proof. Let (X, Y, Z) be a uniform point in S. Then $\log_2 |S|$ is the entropy $H(X, Y, Z)$, and by Shearer, this entropy is at most

$$2\log_2 |S| \le H(X, Y) + H(X, Z) + H(Y, Z).$$

The shadow distribution doesn't need to be uniform, but we can upper bound its entropy with that of the uniform distribution:

$$\le \log_2 \pi_{xy}(S) + \log_2 \pi_{xz}(S) + \log_2 \pi_{yz}(S).$$

Taking 2 to the power of both sides gives the result. □

Remark. *We can actually get the same result for a volume $S \subset \mathbb{R}^3$: the volume of S squared is at most the areas of the projections onto the planes. This can be proved by approximating S as a union of grid boxes!*

Let's now look at Shearer's inequality in its general form.

Theorem 10.16 (Shearer's lemma, general)
Let $A_1, \cdots, A_s \subseteq [n]$, where each $i \in [n]$ appears in at least k different A_js. Let $X_1, \cdots, X_n$ be random variables and define the joint random variables

$$X_{A_j} = (X_i)_{i \in A_j}.$$

Then

$$kH(X_1, \cdots, X_n) \le H(X_{A_1}) + \cdots + H(X_{A_S}).$$

In the special case (Lemma 10.14), A_1, A_2, and A_3 are just the two-element subsets of $(1, 2, 3)$. The proof of the general case is the same! Let's establish a corollary analogous to Corollary 10.15:

Corollary 10.17
Let $A_1, \cdots, A_s \subseteq \Omega$, where each $i \in \Omega$ appears in at least k different A_j's. Then for every family $\mathcal{F}$ of subsets of Ω,

$$|\mathcal{F}|^k \le \prod_{j=1}^{S} |\mathcal{F}|_{A_j}|,$$

where the notation $\mathcal{F}|_{A_j}$ means $\mathcal{F}$ restricted to the elements of A_j: $\{S \cap A : S \in \mathcal{F}\}$.

Proof. Let $(X_1, \cdots, X_n) \in \{0, 1\}^n$ be the indicator vector of a uniform random $F \in \mathcal{F}$. Then

$$k \log_2 |\mathcal{F}| = kH(X_1, \cdots, X_n) \le \sum_{j=1}^{S} H(X_{A_j}).$$

Again, we can upper bound by the uniform entropy:

$$k \log_2 |\mathcal{F}| \le \sum_{j=1}^{s} \log_2 |\mathcal{F}|_{A_j}|,$$

and exponentiate both sides to get the desired result. □

Let's use this for a combinatorial application: in particular, what was the problem that inspired this inequality?

Problem 10.18 (Easy)

What is the largest intersecting family of subsets of m elements, where "intersecting family" means every pair has a nonempty intersection?

The answer is 2^{m-1}: we can just pick every subset that contains the element 1. This is maximal, because any set A and its complement $[m] \setminus A$ can't both appear. If we look back to the beginning of class, the original problem restricted us to only k-element subsets; without this restriction, the problem is easy.

Problem 10.19

What is the largest set of graphs on n labeled vertices so that every pair has a common triangle?

We can get $\frac{1}{8}$ of the total: fix a triangle, and pick all graphs containing that fixed triangle. We also know that it's less than $\frac{1}{2}$ of the total, because we can't pick both a graph and its complement.

Theorem 10.20 (Chung-Frankl-Graham-Shearer)

Every triangle-intersecting family of graphs on n labeled vertices has at most $2^{\binom{n}{2}-2}$ elements.

Proof. Let $\mathcal{G}$ be a triangle-intersecting family on n vertices. Notice that if we restrict ourselves to half our graph and look at the shadow on the two cliques, we must still have an edge-intersecting family, because what's left is a complete bipartite graph.

More concretely, let $m = \binom{n}{2}$. Pick a subset $S \subseteq [n]$ with $|S| = \left\lfloor \frac{n}{2} \right\rfloor$, and let A_S be the union of cliques on S and $[n] \setminus S$. $K_n \setminus A$ is triangle free, so $\mathcal{G}|_A$ must be intersecting. This means that $|\mathcal{G}|_{A_S}| \le 2^{|A_S|-1}$ by the logic above: now if we look at all possible Ss, each edge of K_n appears in k different A_Ss, where $k = \frac{r}{m}\binom{n}{\lfloor n/2 \rfloor}$, $r = |A_S|$.

Now by Shearer's lemma,

$$|\mathcal{G}|^k \le \prod |\mathcal{G}|_{A_S}| = \left(2^{r-1}\right)^{\binom{n}{\lfloor n/2 \rfloor}}.$$

This simplifies to $|\mathcal{G}| \le 2^{m-\frac{m}{r}}$, where $\frac{m}{r}$ is the inverse of the edge density, and since $\frac{m}{r} \ge 2$ for all n, this yields the desired result. □

What's the truth, though?

Theorem 10.21 (Ellis-Filmus-Friedgut, 2012)

Every triangle-intersecting family of graphs on n labeled vertices has at most $\frac{1}{8} \cdot 2^{\binom{n}{2}}$ elements.

The proof of this more refined result uses Fourier analysis!

10.5 Entropy in graph theory

Problem 10.22

Among all d-regular graphs G, how can we maximize the quantity

$$i(G)^{1/v(G)},$$

where $i(G)$ is the number of independent sets and $v(G)$ is the number of vertices of G?

It turns out that this quantity is maximized for a disjoint union of copies of $K_{d,d}$. Let's start by doing this in a special case:

Theorem 10.23 (Kahn)

For a bipartite n-vertex d-regular graph G,

$$i(G) \le [i(K_{d,d})]^{n/2d}.$$

Equality holds if and only if G is the disjoint union of copies of $K_{d,d}$.

Proof. Pick a bipartition of $V(G) = A \cup B$, and let $X = (X_v)_{v \in V(G)}$ be the indicator vector for an independent set of G chosen uniformly at random. (In other words, pick a random independent set, and put a 1 for each vertex in the set and 0 everywhere else.) Then the entropy of this variable is just $H(X) = \log_2(i(G))$.

How can we upper bound this? X is not necessarily uniform or independent on the vertices, but we can still write

$$\log_2(i(G)) = H(X) = H(X_A) + H(X_B|X_A).$$

Observe that because the graph is d-regular and bipartite, each vertex in A lies in the neighbor sets of d vertices in B. Therefore, we can simplify the first term using Theorem 10.16 and also bound the second term by subadditivity:

$$H(X) \le \frac{1}{d}\sum_{b \in B} H(X_{N(b)}) + \sum_{b \in B} H(X_b|X_A)$$

Dropping conditioning on the second term (forgetting about the non-neighbors),

$$H(X) \le \frac{1}{d}\sum_{b \in B} H(X_{N(b)}) + \sum_{b \in B} H(X_b|X_{N(b)}).$$

Fix a $b \in B$. We upper bound the expression

$$H(X_{N(b)}) + dH(X_b|X_{N(b)}).$$

We want to relate this to the entropy of $i(K_{d,d})$ somehow: we will do so by replacing X_b with d identical independent variables $X_b^{(1)}, \dots, X_b^{(d)}$ that have the same distribution given $X_{N(b)}$ as the original X_b. Then,

$$\begin{aligned} H(X_{N(b)}) + dH(X_b|X_{N(b)}) &= H(X_{N(b)}) + H\left(X_b^{(1)}|X_{N(b)}\right) + \cdots + H\left(X_b^{(d)}|X_{N(b)}\right) \\ &= H(X_{N(b)}) + H\left(X_b^{(1)}, \cdots, X_b^{(d)}|X_{N(b)}\right) \\ &= H(X_{N(b)}, X_b^{(1)}, \cdots, X_b^{(d)}), \end{aligned}$$

where the last equality follows from the chain rule. The key observation is that the joint random variable $Y = (X_{N(b)}, X_b^{(1)}, \cdots, X_b^{(d)})$ is the indicator variable of some random independent set of $K_{d,d}$: $X_{N(b)}$ corresponds to the d vertices on the left side and the d variables $X_b^{(i)}$ correspond to d different vertices on the right side! The values that Y takes correspond to independent sets, because the original X_b (and thus none of the copies) is never 1 if there's a 1 in any coordinate of $X_{N(b)}$.

This distribution of Y may not be uniform, but we can still upper bound its entropy by the entropy of the uniform distribution over independent sets of $K_{d,d}$, which is (by Lemma 10.3 as always) $\log_2(i(K_{d,d}))$.

Our graph G is d-regular, so the two pieces of the bipartition have size $\frac{n}{2}$. Because the above bound holds for every $b \in B$,

$$\log_2 i(G) \le \frac{n}{2d}\log_2(i(K_{d,d}))$$

as desired. □

In this proof, we used almost nothing about independent sets, and that motivates us to generalize this result.

Definition 10.24

A **graph homomorphism** $G \to H$ is a map of the vertex set $V(G) \to V(H)$ such that every edge $uv \in G$ is mapped to an edge $\phi(u)\phi(v)$ in H.

Example 10.25

Here are two examples of graph homomorphisms:

- **Independent sets:** Let H be the graph on two vertices $\{0, 1\}$ with an edge between 0 and 1 and a self-loop on 0. Then, a map $\phi\colon (V(G)) \to H$ induces a homomorphism if and only if $\phi^{-1}(1)$ forms an independent set.
- **q-colorings:** Let $H = K_q$. The proper q-colorings of a graph G correspond to homomorphisms from G to H: color each vertex in G mapping to i with the color i.

Theorem 10.26 (Galvin-Tetai)

Let G be an n-vertex, d-regular bipartite graph, and let H be any (possibly looped) graph. Let $\mathrm{Hom}(G, H)$ to be the set of homomorphisms from G to H: then

$$|\mathrm{Hom}(G, H)| \le |\mathrm{Hom}(K_{d,d}, H)|^{n/2d}.$$

The proof of this result is identical to the proof of Theorem 10.23.

Corollary 10.27

Let G be an n-vertex d-regular bipartite graph, and let $q \in \mathbb{N}$. Let $c_q(G)$ denote the number of proper q-colorings of G: then

$$c_q(G) \le c_q(K_{d,d})^{v(G)}.$$

Proof. Let X be the vector of colors of a uniformly random coloring of G, and the rest follows as above. □

Is it possible to prove an analog of Theorem 10.23 for general (not necessarily bipartite) graphs? The answer is yes!

Theorem 10.28

For a n-vertex d-regular graph G,

$$i(G) \le [i(K_{d,d})]^{n/2d}.$$

Equality holds if (and only if) G is the disjoint union of copies of $K_{d,d}$.

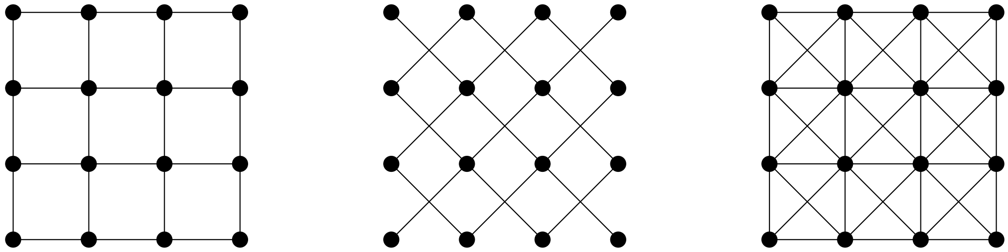
Proof. We will reduce to the bipartite case.

Lemma 10.29 (Zhao)

For all G,

$$i(G)^2 \le i(G \times K_2).$$

Remark. *There's lots of ways to denote a graph product. Given two paths G and H on 4 vertices, there's three main ways to construct a graph product of those paths:*



These, naturally, should be denoted $G \square H$, $G \times H$, and $G \boxtimes H$, respectively.

Proof of lemma. We will construct an injection from $\mathcal{I}(G \sqcup G)$, the collection of independent sets in two disjoint copies of G, to $\mathcal{I}(G \times K_2)$. Think of $G \sqcup G$ as two copies of G, one above the other, and $G \times K_2$ as the same thing but with parallel edges replaced with crosses.

Let's say we have some independent set $S \in \mathcal{I}(G \sqcup G)$: if we take those same vertices in $G \times K_2$, we might not have an independent set, because there are some bad edges: treating $G \sqcup G$ as two layers 0 and 1,

$$E_{\text{bad}} = \{uv \in E(G) : (u, 0), (v, 1) \in S\}.$$

All edges in E_{bad} correspond to an edge in $G \times K_2$ with one endpoint in $\{u \in V(G) : (u, 0) \in S\}$, which is the set of vertices of S (our not-quite independent set) in the top layer. Fix some ordering of the subsets of $V(G)$ (for example, lexicographical, and take Q to be the first subset (in our ordering) of $V(G)$ such that each bad edge in E_{bad} has exactly one endpoint in Q. In other words, we're finding some canonical subset that "shows" our bipartition.

Now swap each pair of $V(G \times K_2)$ in Q (in other words, replace $(v, 0)$ with $(v, 1)$ and vice versa): we can check that this gives us an independent set in $G \times K_2$. In addition, this mapping is injective: find the edges of E that correspond to E_{bad}, and then we can find Q and reverse all of the swaps that we did. □

The graph $G \times K_2$ is d-regular and bipartite with $2n$ vertices, so we can apply Theorem 10.23. This gives an inequality

$$i(G) \le i(G \times K_2)^{\frac{1}{2}} \le i(K_{d,d})^{n/2d},$$

and we're done. □

This means that for independent sets, we can drop the bipartite hypothesis: can we do the same in general for graph homomorphisms? The answer is no!

Example 10.30

Take H to be two disjoint loops. Any graph homomorphism into H sends each connected component to one of the two vertices of H, so the graph with the most graph homomorphisms into H is not a union of copies of $K_{d,d}$ but rather a union of cliques K_{d+1}, since we're just trying to maximize the number of connected components.

The above bipartite swapping trick does not work for some variants of the problem, such as the number of q-colorings instead of the number of independent sets. Recently, the problem for the number of proper colorings was settled using a different method by Sah, Sawhney, Stoner, and Zhao.

Also, we can reduce "bipartite" to "triangle-free" in the graph homomorphism theorem. On the flip side, for any G with triangles, there exists a graph H for which the theorem is not true! However, we don't have good conjectures on classifications of the graph H.

10.6 More on graph homomorphisms: Sidorenko's conjecture

Definition 10.31

Let $t(H, G)$ denote the number of homomorphisms from G to H, divided by the total number of vertex maps $|V(G)|^{|V(H)|}$.

In other words, this is the probability that a uniform random vertex map induces a graph homomorphism.

Conjecture 10.32 (Sidorenko)

If H is a bipartite graph, then for all G, the homomorphism density

$$t(H, G) \geq t(K_2, G)^{e(H)},$$

where $t(K_2, G)$ is the edge density.

Rephrased, this can be phrased another way: among all graphs G with a fixed edge density, which G has the minimum number of copies of H? Sidorenko's conjecture says (informally) that this is a "random" G. This is still an open problem, but let's look at a specific case.

Theorem 10.33

Let G be a graph with n vertices and m edges, and let P_4 be a three-edge path. Then

$$\hom(P_4, G) \geq n^3 \left(\frac{2m}{n^2}\right)^3 = \frac{8m^3}{n^2}.$$

Proof. We'll use the entropy method, but the proof will look slightly different from the techniques that have been used so far. We're trying to lower-bound our quantity this time, so we don't necessarily want to start with a uniform distribution.

Basically, our goal is to construct a probability distribution on the set of homomorphisms $\mathrm{Hom}(P_4, G)$ with entropy at least $\log_2\left(\frac{(2m)^3}{n^2}\right)$. Then by the uniform inequality, we can find that the entropy of the uniform distribution, which is $\log_2$ of the number of homomorphisms, is at least that quantity. Note that a homomorphism is just a 4-vertex path.

Construct X, Y, Z, W to be a 4-vertex walk on G in the following way: let XY be a unfirom edge of the graph, Z be a uniform neighbor of Y (allowing X), and W be a uniform neighbor of Z. The entropy of this distribution is, by the chain rule,

$$H(X, Y, Z, W) = H(X) + H(Y|X) + H(Z|X, Y) + H(W|X, Y, Z).$$

Note that if XY is a uniform edge, YZ and ZW are also uniformly distributed. This is because the vertex probability distribution of X is proportional to $d(v)$: specifically,

$$\Pr(X = v) = \frac{d(v)}{2m}.$$

This is true for Y as well, and now the distribution of Z as a uniform neighbor of Y is the same as the distribution of X as a uniform neighbor of Y: $Z|Y \sim X|Y$. So YZ is uniform, and so is ZW by the same argument. That means

$$H(X, Y, Z, W) = H(X) + H(Y|X) + H(Z|Y) + H(W|Z) = H(X) + 3H(Y|X),$$

and this is (by definition)

$$\sum_v \frac{-d(v)}{2m} \log_2 \frac{d(v)}{2m} + 3\sum_v \frac{d(v)}{2m} \log_2 d(v),$$

where $\log_2 d(v)$ is $H(Y|X = v)$. Expanding and applying convexity, this is

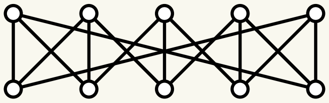
$$= \log_2(2m) + 2\sum_v \frac{d(v)}{2m} \log_2 d(v) \geq \log_2(2m) + \frac{2n}{2m} \cdot \frac{2m}{n} \log_2 \frac{2m}{n},$$

and rearranging gives an entropy of $\log_2 \frac{(2m)^3}{2n}$. □

It turns out that this proof works for every tree. For what kinds of graphs is Sidorenko's conjecture harder to resolve?

Fact 10.34

The smallest open case of Sidorenko's conjecture is the following Mobius graph: it's $K_{5,5}$ minus a Hamiltonian cycle.



It turns out this is the incidence graph of the smallest simplicial complex of the Mobius strip. One side is the set of vertices, and the other is the set of faces.

Notably, the Mobius graph doesn't fit the conditions of the following theorem, which resolves Sidorenko's conjecture for certain graphs:

Theorem 10.35 (Conlon-Fox-Sudakov)

Sidorenko's conjecture holds for a graph H if there exists a bipartition $H = A \sqcup B$ such that there exists a vertex $a \in A$ with $N(a) = B$.

There are also ways to interpret Sidorenko's conjecture beyond graph theory! It turns out Sidorenko's conjecture where H is a three-edge path (Theorem 10.33) is equivalent to the following inequality:

Proposition 10.36

Given a function $f : [0, 1]^2 \to [0, \infty]$,

$$\int_{[0,1]^4} f(x,y)f(y,z)f(z,w)dxdydzdw \geq \left(\int_{[0,1]^2} f(x,y)dxdy\right)^3.$$

As a grad student, Professor Zhao posted on Math Overflow a few years ago asking for a Cauchy-Schwarz proof of this. A week ago, Sidorenko actually answered it!

Sidorenko. Think of $g(x) = \int f(x,y)dy$ as representing the "degree of x." Then the left hand side becomes

$$\int f(x,y)f(z,y)g(z)dxdydz$$

but we can also rewrite the graph as a path $u \to x \to y \to z$, so the left hand side is also

$$\int g(x)f(x,y)f(z,y)dxdydz.$$

Applying Cauchy-Schwarz,

$$\text{LHS} \geq \int g(x)^{1/2} f(x,y) f(z,y) g(z)^{1/2}$$

and since this integral is symmetric with respect to x and z, we can write this as

$$= \int \left(\int g(x)^{1/2} f(x,y) dx \right)^2 dy \geq \left(\int g(x)^{1/2} f(x,y) dx dy \right)^2$$

by Cauchy-Schwarz, and now we can integrate out

$$\left(\int g(x)^{3/2} dx \right)^2 \geq \left(\int g(x) dx \right)^3 = \left(\int f(x,y) dx dy \right)^3,$$

and we're done. $\square$

11 The occupancy method

11.1 Introducing the technique

Let's think more about how to approach independent sets in graphs – in physics, this is related to the hard-core model. Imagine a physical system with solid balls that cannot overlap (for example, in a system of atomic particles). This corresponds to having "independent sets," since we don't want adjacent vertices to be occupied.

Consider picking a random independent set $I \subset V(G)$, so that every set I is chosen with probability proportional to $\lambda^{|I|}$. Here, λ is known as the **fugacity**. Normalizing this probability, the denominator is something that will come up frequently:

Definition 11.1

Define the **independence polynomial** or **partition function**

$$P_G(\lambda) = \sum_{I \text{ independent set}} \lambda^{|I|}.$$

Note that the probability of a set I being picked is

$$P(I) = \frac{\lambda^{|I|}}{P_G(\lambda)}.$$

In particular, if $\lambda = 1$, then $P_G(1) = i(G)$ (the number of independent sets) and we're choosing our independent set uniformly at random. We actually have a result bounding the value of our partition function across all graphs G:

Theorem 11.2 (Kahn, Zhao, Galvin-Tetali)

If G is an n-vertex d-regular graph, and $\lambda \geq 0$ is our parameter,

$$P_G(\lambda)^{1/n} \leq P_{K_{d,d}}(\lambda)^{1/2d}.$$

Previously, we proved this theorem for $\lambda = 1$. A proof for general λ was recently found (credited to Davies-Jenssen-Perkins-Roberts) that is very probabilistic! Here, we have a change of perspective: instead of trying to count everything directly, we find an "observable" by sampling at random.

Definition 11.3

Let the **occupancy fraction** be the expected fraction of $V(G)$ contained in the random independent set I that we've chosen:

$$\overline{\alpha_G}(\lambda) = \frac{1}{|V(G)|}\mathbb{E}[|I|].$$

We can explicitly write this out and then rewrite it as a derivative:

$$= \frac{1}{nP_G(\lambda)} \sum_I |I|\lambda^{|I|} = \frac{\lambda}{n} \frac{d}{d\lambda} \log P_G(\lambda).$$

This is related to the cumulant, which is the logarithm of the moment generating function.

Theorem 11.4

Let G be a d-regular graph, and let $\lambda \geq 0$ be our fugacity parameter. Then

$$\overline{\alpha_G}(\lambda) \leq \overline{\alpha_{K_{d,d}}}(\lambda).$$

So sampling an independent set, the expected number of vertices is maximized at $K_{d,d}$. In particular, note that

$$\frac{\log P_\lambda(G)}{|V(G)|} = \int_0^\lambda \frac{\overline{\alpha_G}(\tau)}{\tau} d\tau, \quad \frac{\log P_\lambda(K_{d,d})}{2d} = \int_0^\lambda \frac{\overline{\alpha_{K_{d,d}}}(\tau)}{\tau} d\tau,$$

and now we can bound over the integral to get the following, which is what we're trying to show in Theorem 11.2 above about the partition function:

Corollary 11.5 (implies Theorem 11.2)

$$\frac{\log P_G(\lambda)}{|V(G)|} \leq \frac{\log P_{K_{d,d}(\lambda)}}{2d}.$$

Let's first prove a special case of our theorem:

Proof of Theorem 11.4 for G triangle-free. Pick a random independent set I (under our hard-core model) on G with fugacity λ. Call a vertex v **occupied** if $v \in I$ and **uncovered** if $N(v) \cap I = \varnothing$; that is, v has no occupied neighbors (though v can still be occupied).

If v is occupied, it must be uncovered (because a vertex and its neighbor can't both be in an independent set), and we also have the following facts:

- If v is uncovered, v can be occupied or not: the probability

  $$\Pr[v \text{ occupied} | v \text{ uncovered}] = \frac{\lambda}{1+\lambda},$$

  because v acts in isolation.
- Meanwhile, what about

  $$\Pr[v \text{ uncovered} | v \text{ has exactly } j \text{ uncovered neighbors}]?$$

  The covered neighbors can't be occupied, so we don't need to worry about them. We know that G is triangle-free, so $N(v)$ is an independent set, and if we condition on everything farther away from v than $N(v)$, the j uncovered neighbors are independent of each other. Thus, this is just

  $$= \frac{1}{(1+\lambda)^j}.$$

So now let's find $\overline{\alpha_G}(\lambda)$ in two different ways. For one, this is just

$$= \frac{1}{n} \sum_{v \in V} \Pr(v \text{ occupied}) = \frac{1}{n} \cdot \frac{\lambda}{1+\lambda} \sum_{v \in V} \Pr(v \text{ uncovered}),$$

since all occupied vertices are uncovered. But by the second fact above, we also know that this is equal to

$$= \frac{1}{n} \cdot \frac{\lambda}{1+\lambda} \sum_{v \in V} \sum_{j=0}^{d} \Pr(v \text{ has } j \text{ uncovered neighbors})(1+\lambda)^{-j}$$

by Bayes' rule.

Now here's another way to compute this quantity:

$$\overline{\alpha_G}(\lambda) = \frac{1}{nd}\sum_v \sum_{u \in N(v)} \Pr(u \text{ occupied}),$$

using the fact that G is d-regular, so each vertex is equally likely to come up. This can then be written as

$$= \frac{1}{nd}\frac{\lambda}{1+\lambda}\sum_v \sum_{u \in N(v)} \Pr(u \text{ uncovered}).$$

This can be thought of as running a two-part experiment: pick I from the hard-core model, and then pick v to be a uniform random vertex of G. Let Y be the random variable equal to the number of uncovered neighbors of v with respect to I.

On one hand, the occupancy fraction $\overline{\alpha_G}(\lambda)$ is

$$\overline{\alpha}_G(\lambda) = \frac{\lambda}{1+\lambda}\mathbb{E}[(1+\lambda)^{-Y}]$$

from the first calculation. But by the second calculation, we also have

$$\overline{\alpha_G}(\lambda) = \frac{1}{d}\frac{\lambda}{1+\lambda}\mathbb{E}[Y].$$

Setting these equal,

$$\mathbb{E}[(1+\lambda)^{-Y}] = \frac{1}{d}\mathbb{E}[Y].$$

where we should remember that Y is a random variable supported on $\{0, 1, \cdots, d\}$, since our graph is d-regular.

But now instead of considering Y coming from this specific process, let's imagine Y is any probability distribution supported on $\{0, 1, \cdots, d\}$ that satisfies the condition we've just derived. Our goal is to show that

$$\frac{1}{d}\frac{\lambda}{1+\lambda}\mathbb{E}[Y] \le \overline{\alpha_{K_{d,d}}}(\lambda).$$

Note that this can be simplified as a linear program: let $x_k = \Pr(Y = k)$, and then we want to maximize

$$\mathbb{E}[Y] = \sum_{k=0}^{d} k x_k$$

under the linear constraints

$$x_k > 0, \sum_{k=0}^{k} x_k = 1, \sum_{k=0}^{d} x_k \left((1+\lambda)^{-k} - \frac{k}{d}\right) = 0.$$

It turns out that the maximium for this linear program occurs for the value that arises from $G = K_{d,d}$. This is because we have convexity of $Y \to (1+\lambda)^{-Y}$: maximizing $(1+\lambda)^{-Y}$ when conditioned on the total expectation happens when we concentrate everything at $\{0, d\}$. In addition, for every λ, there is a unique random variable Y supported on $\{0, d\}$ satisfying the constraints we want. The one that results from $K_{d,d}$ satisfies all such constraints: thus, it must be the maximizer. □

But remember that we've been doing this for G triangle-free. How are things different when we look at G in general?

Proof in general. Again, we do a two-part experiment: pick I to be a random independent set from the hard-core model, and choose v to be a uniform random vertex of G. Let H be the graph induced by uncovered neighbors of v (beforehand, we only cared about the number of such neighbors for calculations).

So now we repeat the calculation: the first calculation becomes

$$\overline{\alpha_G}(\lambda) = \frac{\lambda}{1+\lambda}\mathbb{E}[P_H(\lambda)^{-1}]$$

and the second becomes

$$\frac{1}{d}\mathbb{E}[\text{number of occupied neighbors of } v] = \frac{\lambda}{d}\mathbb{E}\left[\frac{P_H'(\lambda)}{P_H(\lambda)}\right].$$

So now instead of enumerating over all possible distributions on Y, we can set up a similar linear program: our main constraint is that the two ways of finding $\overline{\alpha_G}(\lambda)$ are equal. Then we just need to show that our guess for the optimal solution is correct, and this can be done by showing that all of the dual constraints are satisfied. □

11.2 An alternative approach to the above problem

Here's another way to set up a linear program! Again, let's consider $\overline{\alpha_G}(\lambda)$ in two different ways (we'll drop G and λ throughout for sake of notation). Then

$$\overline{\alpha} = \Pr(v \text{ occupied}) = \frac{\lambda}{1+\lambda}\Pr(v \text{ uncovered}).$$

Letting X be the number of occupied neighbors of v, we have a random variable that depends both on our independent set I and our vertex v. Note that v is uncovered is the same as saying that $X = 0$.

On the other hand, if we pick v and then pick a uniform neighbor u, then u is also a uniform random vertex (because G is d-regular). Now

$$\overline{\alpha} = \frac{1}{d}\sum_{u \in N(v)} \Pr(u \text{ occupied}) = \frac{1}{d}\cdot\mathbb{E}[X]$$

by linearity of expectation. We know that X takes on one of the values $\{0, 1, \cdots, d\}$: denote p_k to be the probability $\Pr(X = k)$. Since we counted $\overline{\alpha}$ in two ways, we can set the values equal:

$$\frac{\lambda}{1+\lambda}\Pr(X = 0) = \frac{1}{d}\mathbb{E}[X],$$

and plugging in values yields

$$\frac{\lambda}{1+\lambda}p_0 = \frac{1}{d}(p_1 + 2p_2 + \cdots + dp_d).$$

X is some variable with various constraints, such as this one, and we can relax the problem by throwing in more constraints to form a linear program. But somehow this isn't capturing the whole system, since this doesn't give a strong enough answer.

So what else can we do? We can consider the probabilities p_k and p_{k-1} and try to come up with relations between them. If exactly k neighbors of v are occupied, we can produce another independent set I' with $k-1$ neighbors by removing one of them from I. There are k ways to remove a neighbor, and there are at most $d-k+1$ ways to go back to I: putting this together with the λ factor from larger independent sets, we have

$$\frac{\lambda p_{k-1}}{k} \geq \frac{p_k}{d-k+1}$$

for all $2 \leq k \leq d$. Throw these into our linear program as well: we now want to maximize $\frac{\lambda}{1+\lambda}p_0$ given $p_0, \cdots, p_k \geq 0, p_0 + \cdots + p_k = 1,$, plus the constraints we've found above. This will give us some upper bound to $\overline{\alpha_G}(\lambda)$ for a d-regular graph G – it may not be optimal, since we've only considered some of the constraints.

But it turns out this is indeed enough:

Lemma 11.6

If $(p_0, \cdots, p_d)$ is a maximizer of our linear program, then every inequality of the form $\frac{\lambda p_{k-1}}{k} \geq \frac{p_k}{d-k+1}$ is an equality.

Proof. Otherwise, if there is some k with a strict inequality

$$\frac{\lambda p_{k-1}}{k} > \frac{p_k}{d-k+1},$$

and we can perturb our ps a bit: increase p_0 by ε, decrease p_{k-1} by $\left(\frac{d\lambda}{1+\lambda} + k\right)\varepsilon$, and increase p_k by $\left(\frac{d\lambda}{1+\lambda} + (k-1)\right)\varepsilon$ and we have a new maximizer (contradiction). □

So now we have a full-rank system of linear equalities, so there exists a unique solution. Indeed $G = K_{d,d}$ satisfies all of the equalities, and we're done.

The lesson here is that picking a uniform random v and considering it locally gives linear constraints. By only looking at those constraints, we can get some bound on the occupancy fraction, and this is usually enough to solve the problem.

11.3 Further bounds with the occupancy method

Let's try to lower bound the occupancy fraction instead of upper bounding it. This was initially developed by Shearer: "triangle-free graphs have large independent sets."

Theorem 11.7

Fix a parameter $\lambda \geq 0$, and let G be a triangle-free graph with maximum degree d. Then

$$\overline{\alpha_G}(\lambda) \geq (1 + o_{d\to\infty}(1))\frac{\log d}{d}.$$

In comparison, remember that every max-degree-d graph has an independent set of size at least $\frac{n}{d+1}$ (take a vertex, remove it and its neighbors). That means that we're gaining a factor of $\log d$ here by having G be triangle-free.

Proof. Pick I according to the hard-core model again, and let v be a uniform vertex of $V(G)$. Let Y be the number of uncovered neighbors of v, so we now also care about the neighbors of neighbors of v.

Just like last time, there's two different ways to write the occupancy fraction. Because we have a triangle-free graph, the neighbors of G behave independently of each other, so the number of uncovered neighbors satisfies

$$\overline{\alpha_G}(\lambda) = \frac{\lambda}{1+\lambda}\mathbb{E}[(1+\lambda)^{-Y}]$$

where the expected value term is the probability that none of the neighbors of v are in I. By convexity, we can bound this:

$$\geq \boxed{\frac{\lambda}{1+\lambda}(1+\lambda)^{-\mathbb{E}[Y]}}.$$

Now since G has maximum degree d, which is similar to being d-regular, we also know that

$$\overline{\alpha_G}(\lambda) = \mathbb{E}[\Pr(v \text{ occupied})] \geq \frac{1}{d}\mathbb{E}\left[\sum_{u\in N(v)} \Pr(u \text{ occupied})\right] = \boxed{\frac{1}{d}\frac{\lambda}{1+\lambda}\mathbb{E}[Y]}$$

by linearity of expectation. The occupancy ratio is then at least the maximum of the two estimates:

$$\overline{\alpha} \geq \frac{\lambda}{1+\lambda} \max\left\{(1+\lambda)^{-\mathbb{E}[Y]}, \frac{1}{d}\mathbb{E}[Y]\right\}.$$

Note that one of these expressions decreases with $\mathbb{E}[Y]$ and the other increases. That means that there is some absolute constant we can find here:

$$\geq \frac{\lambda}{1+\lambda} \min_{y>0} \max\left\{(1+\lambda)^{-y}, \frac{1}{d}y\right\},$$

and after some optimization, this indeed yields the result we want. □

What consequences does this have? First of all, there are other situations where similar techniques and theorems apply: in fact, we saw one earlier when we were talking about dense sphere-packings in high dimensions. The **hard-core model** models non-overlapping spheres, and we can set up problems in similar ways where we draw a sphere packing according to some distribution to find the expected fraction of space taken up. Doing the calculations, we found a sphere-packing density of $n2^{-n}$ in $\mathbb{R}^n$. This is close to the best we know for almost all n, and it's definitely better than the 2^{-n} that we got with a greedy packing. (Notice that we have the same characteristic log term as in our graph theory problem.)

Similarly, we can pack spherical caps on a sphere, which is like saying that we want points on a sphere that are pairwise separated by some angle. The most prominent case of this is called the **kissing problem**, which asks about the maximum number of unit balls that are nonoverlapping but all touch a central unit ball. This problem is interesting even in 3 dimensions!

11.4 A useful corollary: Ramsey numbers

We can translate the occupancy number statement directly into a graph theoretic statement:

Corollary 11.8

Every triangle-free graph on n vertices with max degree at most d contains an independent set of size at least $(1 + o_{d\to\infty}(1))\frac{\log d}{d}n$.

This actually gives us a bound for the Ramsey numbers:

Corollary 11.9

We have

$$R(3, k) \leq (1 + o(1))\frac{k^2}{\log k}.$$

In other words, there exists an $n \sim \frac{k^2}{\log k}$ such that every graph on n vertices has either a triangle or a large independent set.

Proof. If our graph is triangle-free, every neighborhood is an independent set. So if any vertex has degree k, we automatically have an independent set of the desired size. Otherwise, by the corollary, there exists an independent set of size at least $(1 + o(1))\frac{\log k}{k} \cdot n$, and choosing $n = (1 + o(1))\frac{k^2}{\log k}$ provides us with the independent set of desired size. □

This is essentially the best upper bound we have, and we also know a pretty close lower bound:

$$R(3,k) \geq \left(\frac{1}{4} + o(1)\right) \frac{k^2}{\log k}.$$

To construct a graph with this many vertices, remember that we found lower bounds to $R(k,k)$ by taking a random graph. Use a similar philosophy here: construct our graphs randomly, but use a **triangle-free process**. Start with an empty graph on n vertices, and keep adding uniform random edges subject to the constraint "don't make triangles."

Remark. *In contrast, we don't even know the order of magnitude for $R(4,k)$ yet.*

11.5 Back to independent sets

We found an upper bound on $i(G)^{1/V(G)}$ earlier on: can we find a way to minimize this quantity? Which graphs have the minimum number of independent sets?

Remark. *Two different students guessed "K_{d+1}" and "random."*

It turns out that as stated, among all d-regular graphs G, the minimizer is K_{d+1}. However, if we restrict ourselves to bipartite d-regular graphs G, the answer becomes "random" or "a d-regular infinite tree."

One way to think of this is to consider a "2-lift" G': take two copies of G, and replace edges with their "crossed" versions. Then we have $i(G') \leq i(G)^2$, and now by repeatedly lifting to destroy small cycles in the graph, we find that the number of independent sets, normalized, approaches some constant:

$$i(G_n)^{1/v(G_n)} \to c_d$$

if the girth of G_n goes to infinity. In other words, we have a "tree-like" graph!

Let's modify the question so that neither of these answers is allowed:

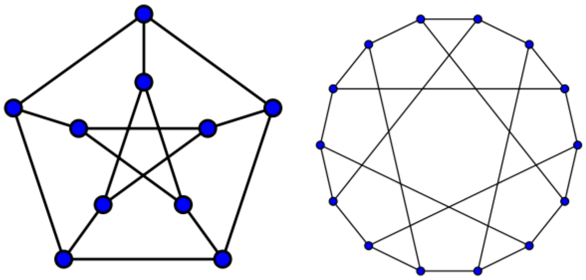
Problem 11.10

Let G be a 3-regular graph. How do we maximize $i(G)^{1/v(G)}$ if we're not allowed to have 4-cycles? Similarly, how do we minimize this quantity if our graph is triangle-free?

Theorem 11.11 (Perernau-Perkins, 2018)

Among 3-regular graphs G without cycles of length 4, $i(G)^{1/v(G)}$ is minimized by the Peterson graph and maximized by the Heawood graph.

Here are the Peterson and Heawood graphs, respectively:



These are **Moore graphs** - they are essentially the smallest graphs that are d-regular with a specific girth condition. The idea here is that we can throw girth conditions into our linear program, because the additional constraints are local.

11.6 Proper colorings in graphs

Let's look at one more example: let $C_q(G)$ be the number of proper q-colorings of our graph G. Recall that we want to maximize $(C_q(G))^{1/V(G)}$ across d-regular graphs G: this can be done using the **Potts model** in statistical physics. Basically, sample a coloring $c : V \to [q]<$ not necessarily proper, where a coloring occurs with probability proportional to $\beta^{m(c)}$, where $m(c)$ is the number of monochromatic edges. (Note that the parameter β acts sort of like the parameter λ, and in the end, we can set $\beta = 0$. Notably, $\beta = 1$ gives a uniform coloring.)

So now we have a partition function for the Potts model

$$Z_{G,q}(\beta) = \sum_{c:V\to[q]} \beta^{m(c)};$$

note now that the log derivative

$$U_{G,q}(\beta) = \frac{\beta}{e(G)} \frac{d}{d\beta} \log Z = \frac{\mathbb{E}[m(c)]}{e(G)}$$

is the expected fraction of monochromatic edges. In physics, this is known as the "internal energy."

Theorem 11.12

For all 3-regular graphs G, $q \geq 2$, and $0 \leq \beta \leq 1$,

$$U_{g,q}(\beta) \geq U_{K_{3,3},q}(\beta).$$

Integrating this (the inequality is flipped because we integrate from 1 to $\beta \leq 1$), we get

$$C_q(G)^{1/V(G)} \leq c_q(K_{3,3})^{1/6}.$$

Proving this requires a similar kind of trick of constructing the two-part experiment and finding linear constraints. However, this has lots of variables - one for each possible configuration. We don't actually know how to do this by hand, but we plug this into a computer, and indeed $K_{3,3}$ is the maximizer! Unfortunately, we don't know how to get this to work for d-regular in general: the computation time is far too large.

12 A teaser for "Graph Theory and Additive Combinatorics"

For this last lecture, titled "triangles and equations," we're going to be previewing 18.217, "Graph Theory and Additive Combinatorics," being taught next fall. (This was a class Professor Zhao taught in Fall 2017 as well!)

12.1 A glance at Fermat's last theorem

Like many other mathematicians of the time, Schur thought about Fermat's Last Theorem, which looks for solutions

$$X^n + Y^n + Z^n, n \geq 3.$$

He considered the following idea: why not reduce this mod p? If we can show that for infinitely many different primes, there are no solutions mod p, then it must not have any solutions in the integers. Unfortunately, this doesn't work, and he proved it doesn't work. Instead, we got the following result:

Theorem 12.1 (FLT mod p)
For all n, the equation

$$X^n + Y^n = Z^n \bmod p$$

has a nontrivial solution (where p does not divide XYZ) for all sufficiently large p.

In fact, Schur proved a more combinatorial Ramsey-type result:

Theorem 12.2 (Schur)
For all r, there exists an integer N such that if we color $\{1, \cdots, N\}$ with r colors, then there exists a monochromatic solution to $X + Y = Z$.

The modern way to view this is that we can reduce to Ramsey's theorem.

Proof. Given a coloring of $\phi : [n] \to [r]$, we color a complete graph K_{n+1} with vertices $[N+1]$ by coloring an edge (i, j) with colors $\phi(|i-j|)$. By Ramsey's theorem, if N is sufficiently large, there exists a monochromatic triangle (think of this as using Pigeonhole principle). Then if those vertices are $i \leq j \leq k$, then

$$\phi(k-i), \phi(k-j), \phi(j-i)$$

are all the same color, and now we've found a monochromatic $X + Y = Z$: let $x = j - i, y = k - j, z = k - i$. □

So this is a connection between a number-theoretic problem and the corresponding graph-theory problem.

Fact 12.3
This implies FLT mod p: let H be the subgroup of $(\mathbb{Z}/p\mathbb{Z})^*$ consisting of nth powers $\{a^n : a \in (\mathbb{Z}/p\mathbb{Z})*\}$. Then partition the numbers $\{1, 2, \cdots, (p-1)\}$ into H-cosets: this uses at most n colors. If p is sufficiently large in terms of n, Schur's theorem tells us that there exists some coset with a monochromatic solution: if the coset is aH, then

$$aX^n + aY^n = aZ^n \bmod p;$$

multiply by the multiplicative inverse of a to get the result.

This is a "baby example" of what is called additive combinatorics, which also goes under the name of "combinatorial number theory." Usually, number theory is about multiplying primes together: here, we care more about combinatorial properties of the numbers. Let's do a few more examples.

12.2 Turán's theorem and more

The question here is to find the **maximum number of edges in an n-vertex triangle-free graph.** The answer is that we want a completely bipartite graph $K_{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$.

Well, the analogous problem in graph theory is to find the maximum subset of $[N]$ without a solution $x + y = z$. Picking odd numbers, or picking the numbers greater than $\frac{N}{2}$, gives half (rounded up) of the total subset. We can't do better, because we can let M be the maximum element in our subset: then both x and $M - x$ can't appear in our set, so we can have at most half density.

Here's a way to make that question a bit harder:

Problem 12.4

What's the maximum size subset of $[N]$ without a solution $x + y = 2z$ (where x, y, z aren't all equal)? In other words, we want a three-term arithmetic progression?

This problem is hard to answer, and it's even hard to find good guesses. One thing we can do is try this greedily: add 1 and 2, and then successively add numbers if they don't create 3-term arithmetic progressions. This gives only the numbers with digits 1 and 2 in base-3 representation, and it has density $N^{\log_3 2}$, since we pick 2^k of the first 3^k positive integers.

This is nowhere near the best, though: there exists an $N^{1-o(1)}$ construction, which we won't discuss here. On the other hand, due to Roth, we also know that the size is sublinear: we cannot get a positive proportion of $[N]$.

$$N^{1-o(1)} \le \text{size of subset} \le o(N).$$

This is probably Roth's second most important result, and it's driven a lot of research in additive combinatorics.

But now, what's the analogous question for the 3-term AP problem in graph theory?

Problem 12.5

What's the maximum number of edges in an n-vertex graph, where every edge is contained in a unique triangle?

Analogously, the answer here is

$$n^{2-o(1)} \le \text{number of edges} \le o(n^2).$$

Let's show the lower bound:

Proposition 12.6

We can find a graph with $n^{2-o(1)}$ edges where every edge is contained in a unique triangle.

Solution. Let's say $A \subset [\mathbb{Z}/N\mathbb{Z}]$ is a subset without three-term arithmetic progressions, where N is odd (just to avoid technicalities). Then we can actually construct a graph where every edge is contained in a unique triangle: we have three vertex sets, $X = Y = Z = \mathbb{Z}/n\mathbb{Z}$, and we put an edge between $x \in X$ and $y \in Y$ if $y - x \in A$, an edge between $y \in Y$ and $z \in Z$ if $z - y \in A$, and an edge between $x \in X$ and $z \in Z$ if $\frac{z-x}{2} \in A$.

What are the triangles in this graph? Note that $y-x, z-y, \frac{z-x}{2}$ form an arithmetic progression, but A doesn't have any 3-APs except for the trivial ones: x, x, x. So every edge here lies in exactly one triangle! This same construction also proves that the upper bound of the graph theory version implies the upper bound in the AP-subset problem. □

We haven't discussed how to prove any of the bounds, but we'll do that in the course next semester. A lot of interesting tools are used to achieve this, and generalizations and extensions have blossomed into a new field.

12.3 A generalization: more modern approaches

Theorem 12.7 (van der Waerden)
For all r and k, there exists N such that if $[N]$ is colored with r colors, then there is a monochromatic k-term arithmetic progression.

Erdös and Turán believed that the real reason for van der Waerden's theorem is not because we use k colors, but because one of our color classes has positive density. This led them to a conjecture in 1936 that was only resolved by Szemerédi in 1975, resulting in the following landmark theorem:

Theorem 12.8
For every δ, there exists N such that every subset $A \subset [N]$ with $|A| \geq \delta N$ contains a k-term arithmetic progression.

The proof is difficult and involved enough that we won't even prove it next semester. But this theorem has been looked at from other directions, and this has led to some success: the results can also be shown with ergodic theory, and this turns out to be more general in some sense. In addition, a Fourier analytic approach (by Roth) also works, but it doesn't work for 4-term arithmetic progressions. (We may have also heard of the "Hardy-Littlewood circle method.") Recently, a newer approach was found that generalizes Roth's proof to "higher-order Fourier analysis."

Remark. *Normal Fourier analysis considers correlations of a function with an exponential phase*

$$\mathbb{E}[f(x)e^{i\alpha x}].$$

In contrast, quadratic Fourier analysis looks at correlations with quadratic exponential phases:

$$\mathbb{E}[f(x)e^{i\alpha x^2}],$$

and these turn out to be essential when studying four-term arithmetic progressions.

12.4 A principle about approaching complicated problems

One last idea that developed out of Szemerédi's theorem is the "regularity lemma." Each of these approaches has its own tools, but overall, there are some connections. The idea here is **structure versus randomness**, or **signal versus noise**: the idea is that a system should be able to be written down as a piece that is structured, plus a "pseudo-random piece."

Example 12.9

If we want to understand 3-term arithmetic progressions in $[N]$, we may want to instead consider functions $f : \mathbb{Z}/n \to \mathbb{R}$. These can be written via the Fourier inversion formula

$$f(x) = \sum_r \hat{f}(r)\omega^{rx},$$

where ω is an Nth root of unity. The coefficients $\hat{f}(r) = \mathbb{E}[f(x)\omega^{-rx}]$ are generally not large (in some sense), so we can write out our sum as a sum of parts where $|\hat{f}(r)|$ is large (structured, few of them) and where $|\hat{f}(r)|$ is small (looks pseudorandom).

Example 12.10

If we start with a symmetric matrix $A \in \mathbb{R}^{n\times n}$, we can decompose it in terms of its spectrum:

$$A = \sum_i \lambda_i v_i v_i^T.$$

We can again separate this into terms where the eigenvalues are large versus small.

For graph theory, there's a similar notion: we can always start by representing a graph by its adjacency matrix, but there's also more combinatorial ways to do this. This leads us to a powerful tool in graph theory:

Theorem 12.11 (Szemerédi's regularity lemma)

Informally, every graph can be decomposed into a bounded number of vertex parts (in terms of some error), so that almost all pairs of parts look "pseudorandom."

In other words, we can think of two vertex parts as being essentially defined by edge densities. Then the "structure" part of this is the density we assign, and the "randomness" is the rest of the graph.

This is a powerful tool: it actually allows us to prove the $o(n^2)$ result for Problem 12.5.

12.5 Graph limits

Let's say we have a sequence of graphs $G_1, G_2, \cdots$ and ask the question of "do these graphs converge?" If we say that $G_n = G\left(n, \frac{1}{2}\right)$, we can say that the sequence converges to a limit, which is some constant function $\frac{1}{2}$. (In other words, we don't care about the specific edges, but only global macroscopic pictures.)

Definition 12.12

A sequence of graphs $(G_n)_n$ **converges** if for all graphs F, the density $t(F, G_n)$ converges to some constant c_F as $n \to \infty$.

This is a very local property, but how exactly do we represent this convergence?

Definition 12.13

A **graphon** is a symmetric, measurable function $W : [0,1]^2 \to [0,1]$.

We can think of our graphs as adjacency matrices: they'll have a bunch of 0s and 1s. Think of the 1s as black squares and the 0s as white squares: as $n \to \infty$, and our eyesight becomes poor, we see a grayscale image.

This isn't **quite** correct yet: for example, what's the limit of $K_{n/2,n/2}$? This can either look like a blown up version of $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, or it can look like a checkerboard! The latter begins to look a lot like $\frac{1}{2}$, but the former is the actually correct answer. So there's some subtleties that we're skipping over.

Here's some more motivation: let's say I want to maximize $x - x^3$ for $x \in [0, 1]$, but I only know about rational numbers. We can still say the answer, but it's a lot more contrived: we have to take some sequence of rationals to get to the answer.

Well, instead, let's say we want to minimize the 4-cycle density in a graph with edge density $\frac{1}{2}$. This is similar to the path of length 3 problem: the answer is to take a sequence of pseudorandom graphs with edge density $\frac{1}{2}$ and number of vertices going to ∞. That's kind of annoying to say, though: is there a nicer way to state the limit? The beauty of using graph limits is that we can just say our answer as $W = \frac{1}{2}$.

Proving these things exist requires Szemerédi's regularity lemma to represent graphs: this allows us to view graphs with this structure versus randomness decomposition. It's a nice fact, by the way, that every sequence of graphs contains at least one graph limit.

12.6 A few open problems

In 18.217, we'll discuss the structure of set addition: let $A + A$ be the set of all numbers $\{a + b : a, b \in A\}$. We can ask questions like "what is the size of $A + A$ if $|A| = n$?" In the integers, the minimum is attained for $A = [n]$, and the maximium is attained with random large numbers: this gives

$$2n - 1 \le |A + A| \le \binom{n+1}{2}.$$

But now, what can we say about A if $A + A$ is small? For example, are there any properties that we know if $|A + A| \le 100|A|$?

This means our set is not too random, or else we'd have quadratic pairwise sums. So there's some arithmetic structure in A:

Theorem 12.14 (Freiman)

A must be contained in a "small" **generalized arithmetic progression**; that is, numbers of the form $a + c_1 d_1 + c_2 d_2 + \cdots + c_k d_k$.

But there's still open problems around this theorem. In particular, the following is considered one of the most important conjectures in the field:

Conjecture 12.15 (Polynomial Freiman-Ruzsa conjecture)

There are two equivalent forms of this: we'll only state this over $\mathbb{F}_2^n$.

- Let $A \subset \mathbb{F}_2^n$ be a subset with small doubling: $|A + A| \le K|A|$. Then there exists a subspace V where $|V| \le |A|$, such that $|V \cap A| \ge K^{-O(1)}|A|$.
- Given a function $\mathbb{F}_2^n \to \mathbb{F}_2^n$ which is "almost linear" – $(f(x + y) - f(x) - f(y))$ takes on at most K values – then there exists a **linear** function $g : \mathbb{F}_2^n \to \mathbb{F}_2^n$ such that $f(x) - g(x)$ takes on at most $K^{O(1)}$ different values.

In the second statement, it's easy to show that we can get 2^K: we just have g agree with f on a basis! Then all errors need to lie in the subspace spanned by $f(x+y)-f(x)-f(y)$. The best result that is known is a quasi-polynomial bound: we have $e^{(\log k)^{O(1)}}$.

Finally, let's consider both addition and multiplication together: much like we define $A+A=\{a+b : a, b \in A\}$, define $AA = \{ab : a, b \in A\}$. Then $|A+A|$ and $|AA|$ can separately be made linear, but there's a conjecture that this can't happen simultaneously:

Conjecture 12.16

Suppose $|A| = n$. Then

$$\max\{|A|+|A|, |AA|\} \geq n^{2-o(1)}.$$

Recent improvements have gotten us from $n^{4/3-o(1)}$ to $n^{4/3+c}$ for a small constant $c > 0$, which is still very far from what we think is the truth. This is another example of the connections between graph theory and additive combinatorics: earlier on, we saw the Szemerédi-Trotter theorem about incidences between points and lines, and it turns out we can connect the earlier material here as well. The idea is that slopes of lines involve both addition and multiplication, so encoding that information into this problem here allows us to use point-line incidences to deduce results about sums and products.