18.226 — Probabilistic Methods in Combinatorics

Class by Yufei Zhao

Notes by Sanjana Das

Fall 2022

Notes for the MIT class **18.226** (Probabilistic Methods in Combinatorics), taught by Yufei Zhao. All errors are my responsibility.

Contents

§1 Introduction

This course is taught by Prof. Yufei Zhao, and the course homepage is math.mit.edu/18.226.

§2 The Probabilistic Method

The Probablistic Method is a powerful technique in combinatorics where to show something exists, we try to construct it randomly, and show that we succeed with positive probability. This is an amazing idea. We'll see lots of examples and applications, beginning with the following illustrative example.

Theorem 2.1

For any graph $G = (V, E)$, there is a subgraph $G' = (V, E')$ (using the same vertex set, and a subset of the edges) such that G' is bipartite, and $|E'| \geq |E|/2$.

So if we start with any graph, we can always find a large bipartite subgraph.

It turns out that we can construct this bipartite graph at *random*.

Proof. Assign every vertex a color black or white at random (uniformly and independently). Take E' to be the set of edges with one black endpoint and one white endpoint. Then E' always creates a bipartite graph — we can put all the black vertices on one side, and the white vertices on the other, and all edges in E' go from one side to the other.

Each edge lies in E' with probability $\frac{1}{2}$. So $\mathbb{E}(|E'|) = \frac{1}{2}|E|$. In particular, there exists some choice of vertex coloring so that $|E'| \geq \frac{1}{2}|E|$. □

This is a very simple but elegant application of the probabilistic method. We didn't really even need to think much about the structure of the graph — this works no matter what the graph is. The goal of this course is to look at this method and see how far we can go. In this first lecture, we'll look at a specific problem — finding lower bounds on Ramsey numbers — and see several more advanced applications of the probabilistic method, as a starting point which will guide the rest of the class.

§3 Ramsey Numbers

Definition 3.1. The **Ramsey number** $R(k, \ell)$ is the smallest positive integer n such that every red-blue edge-coloring of K_n has either a red K_k or a blue K_ℓ.

Example 3.2

$R(3, 3) = 6$: if we take any complete graph on 6 vertices, no matter how we color the edges red and blue, there always exists either a red triangle or a blue triangle. Meanwhile, there is a way to color K_5 without any red or blue triangle — color the outer pentagon red, and the inside star blue.

SUPpose instead of 6 vertices, we had a *lot* of vertices, and we want to color the edges of our large complete graph so that there are no large monochromatic cliques. How can we do this? One attempt is to generalize the answer for $(3, 3)$ and 5, but it turns out this is really hard — finding a structured way to color large graphs is an open problem. Instead, the best method we know today is the probabilistic method. This started with an important paepr of Erdös in 1947, modestly called "Some Remarks on the Theory of Graphs" where he presented an argument that gives a good lower bound on these Ramsey numbers. To some extent, this is still more or less the best bound we know.

Remark 3.3. Frank Ramsey and Paul Erdös are both very interesting figures. Ramsey died young, at the age of 26, but even in his short lifespan he contributed seminal ideas to mathematics, philosophy, and economics. Erdös was one of the most prolific mathematicians — he wrote about 1600 papers in his lifetime (many coauthored — you may have heard of Erdös numbers (Prof. Zhao's is two)). There was a nice New Yorker article about Ramsey ("The Man Who Thought Too Fast") and a lot written about Erdös (n is a number, The Man Who Loved Only Numbers). Erdös is the father of the subject of the probabilistic method (in favt, the textbook has a picture of him on the cover).

Remark 3.4. There are many Hungarian mathematicians, so you should know how to pronounce their names. S is pronounced sh (for example Erdös) and sz is pronounced like an s (for example Lovasz). Also, you typeset `Erd\H{o}s` and not `Erd\"os`.

Theorem 3.5

If n and k are positive integers such that

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

then $R(k,k) > n$.

What this means is that there exists a red-blue edge-coloring of K_n without a monochromatic K_k.

Proof. Color the edges of K_n red and blue randomly (uniformly and independently). For each S which is a k-vertex subset of the vertices, let A_S denote the event that S induces a monochromatic clique. Then

$$\mathbb{P}(A_s) = 2^{1-\binom{k}{2}},$$

since there are $\binom{k}{2}$ edges and there's a $2^{-\binom{k}{2}}$ probability they're all red, and the same probability they're all blue.

We want *none* of these events to occur. One simple way we can bound the probability is the union bound — we have

$$P(\text{Exists a monochromatic } K_k) = P\left(\bigcup_S A_S\right) \le \sum_S \mathbb{P}(A_S).$$

There are exactly $\binom{n}{k}$ k-element subsets, and each probability is $2^{1-\binom{k}{2}}$. The hypothesis states that this sum is less than 1.

So if we do everything randomly, the bad event occurs with probability less than one; thus we have a positive probability of choosing a coloring with no monochromatic K_k. So some such coloring must exist. □

The theorem is stated rather implicitly, with some relation on k and n. We can do some analysis (using Stirling's formula) to get the quantitative bound:

Corollary 3.6

We have

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

We usually care about the exponential growth, so Erdős proved a lower bound of $R(k,k) > 2^{k/2}$ (this is true without any asymptotics). It's still a major open problem whether we can improve this exponent — for example, is $R(k,k) > 2^{0.500001k}$? In the rest of today's lecture, we'll see a few ways to improve the constant.

Remark 3.7. We phrased Erdős's proof in terms of probabilistic arguments. But in the original proof, he doesn't use probability — he phrases everything in terms of counting. This can be done, since we're working in a finite probability space. In that time, the idea of the probabilistic method was so novel that it may not have been accepted as a probabilistic proof then, but times are different now.

Question 3.8. Is there an algorithmically efficient method of finding Ramsey colorings?

This could mean writing down a mathematical description (the Cayley graph of quadratic residues), or a description in the language of computer science. This has surprising connections to the world of *randomness extractors* — suppose you have a world with some weak sources of randomness (the intensity of sunlight), and you want to use that to produce something that's uniformly random (since generating true random numbers is an important and difficult task). It turns out this is related to finding Ramsey colorings, and some recent algorithmic advances in finding Ramsey colorings come from here.

We don't actually have a good way of doing this — checking a k-clique takes a long time. If n is large, the probability of a bad event is very small; so if you do a random coloring, you succeed with very high probability. But it's like trying to find hay in a haystack — anywhere you look there will be hay, but you don't know if you've found it.

Remark 3.9. Ramsey proved that the Ramsey numbers are finite — there exists some large enough n such that any coloring of K_n has a monochromatic large clique. An important topic in Ramsey theory is understanding these bounds.

A result by Erdős–Szekeres (proved using induction) is the bound

$$R(k+1, \ell+1) \le \binom{k+\ell}{k}.$$

The case $k = \ell$ is of particular interest; then this is roughly 4^k. There have been some improvements; the most recent, due to Ashwin Sah, is the bound

$$e^{-c(\log k)^2}\binom{2k}{k}.$$

(We will not discuss upper bounds in this class.)

We will now see a couple of modifications of this proof, as a preview of different variants of the probabilistic method.

§3.1 The Alteration Method

Instead of just taking something at random in one go, suppose we do it in a couple of steps:

1. Construct randomly.
2. Try to fix blemishes.

Theorem 3.10

For all k and n,

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

Corollary 3.11

We have

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

Proof. First we randomly color edges red and blue, as before.

Now delete a vertex from every monochromatic K_k (in an arbitrary way). The end result always has no monochromatic clique K_k (since the second step destroys all of them).

Of course, the danger with this approach is that this may destroy a large part of the graph, so then we may not end up with a large n. So we want to see how many vertices are left.

Let X be a random variable encoding the number of monochromatic K_k after the random coloring. Then each K_k is monochromatic with a probability we've computed before, so $\mathbb{E}X = \binom{n}{k}2^{1-\binom{k}{2}}$.

In the second step, we delete *at most* $|X|$ vertices (it's possible that some cliques overlap, but that's okay). So the total number of vertices remaining is at least $n - |X|$. And

$$\mathbb{E}(n - |X|) = n - \binom{n}{k}2^{1-\binom{k}{2}}.$$

So in expectation we have at least this many vertices left, which means there exists some instance with this many vertices remaining, which is what we wanted. □

Student Question. Is there a conjectured value for $R(k,k)$?

No — we know the base of the exponent is between $\sqrt{2}$ and 4, but we don't have good evidence for any specific one.

§3.2 Lovasz Local Lemma

Now we'll look at another tool — the Lovasz local lemma. In many situations, we want to avoid some collection of bad events $E_1, \ldots, E_n$.

There are some easy cases — if all the bad events have small probability, then we can apply the union bound — if $\sum \mathbb{P}(E_i) < 1$ then by the union bound we can avoid all the bad events ismultaneously.

ANother easy case is if they're all independent. Then if all have probability strictly less than 1, the probabiltiy of avoiding all of them is positive.

The hard and most interesting cases are what happens in between — when the union bound doesn't apply and the events are not independent, but often there's some structure in the problem that means there's a "small amount of independence" — everything is independent from everything else except for a small number.

We will state a version of the lemma; and we will prove it later and discuss it in more detail. (There are more general versions we'll see later.)

Theorem 3.12 (Lovasz Local Lemma)

Let $X_1, \dots, X_N$ be independent random variables, and let $B_1, \dots, B_m$ be subsets of $[N] = \{1, 2, \dots, N\}$. Let E_i be an event that depends only on variables indexed by B_i (in other words, we only have to look at X_j for $j \in B_i$ to see whether E_i holds).

Suppose for every $i \in [m]$, B_i has nonempty intersection with at most d other sets B_j (not including itself), and $\mathbb{P}(E_i) \le \frac{1}{(d+1)e}$ (where e is the constant). Then with positive probability, none of the E_i occur.

The setup is somehwere in between the two extreme cases — each bad event is not too likely, but the total number of events could be significantly larger than d, so the union bound won't give anything useful. But there's only a small amount of dependence — each event overlaps only with d other events. Given these bounds together, LLL tells us you can avoid all these bad events.

Student Question. Does d have to be the same for all the events?

Yes, for now; but we'll see the proof later and more general versions. In practice we usually think of d as a constant. Curiously e is optimal — the theorem is false if you replace e by a smaller number.

Theorem 3.13 (Spencer 1977)

If

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

then $R(k, k) > n$.

Corollary 3.14

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

This is the best bound we know today.

Proof. Color the edges randomly. For each k-element subset of the vertices S, let E_S be the event that S induces a monochromatic clique.

In the first proof we took a union bound. Now we have more powerful tools, so let's apply them.

In the random variable model, each variable is the color of an edge, and $N = \binom{n}{2}$. Then the variables for S and S' overlap if and only if they share two vertices — $|S \cap S'| \ge 2$. For each S, the number of such S' is at most $\binom{k}{2}\binom{n}{k-2}$ (you can count exactly but that's tedious) — we choose the two vertices to have the overlap, and if we're careless about the rest there are $\binom{n}{k-2}$ ways to choose the remaining $k - 2$. So that's our value of d.

Now if we plug in the LLL hypothesis, the probability of each individual bad event is $2^{1-\binom{k}{2}}$. SO if

$$2^{1-\binom{k}{2}} \le \frac{1}{e(d+1)}$$

then we can apply LLL. This inequality is precisely the one given in the theorem statement. □

Today we started with a simple example, and then we saw progressively more advanced techniques that gave better adn better lower bounds. In the next block we'll see what's coming in this course:

Next time we'll see a bunch of cute and neat applications of the probabilistic method. Then we'll move on to exploring Linearity of Expectations in more depth — we already saw this today every time we evaluated $\mathbb{E}$ of a random variable, but we'll explore it in more depth later. We'll then see the Alteration method (we saw an instance of that today — do something naively at random and then try to fix it). We'll then see the *Second Moment Method* — we wish to understand not just that $\mathbb{E}$ is small or large, but also that this variable is tightly controlled around its mean. We usually do this by looking at its *variance*. (In many appplications it's not just enough to understnad the expectation, we also need to control the concentrartion — that will be a major theme in the rest of hte class.) The second moment is often enough, but if you have complete independence you can do more — we'll explore the *Chernoff bound*, which says that if you have the sum of independent rnadom variables, you have extremely good control of its tails. We'll then return to LLL, and its variants and applications. We'll then see correlation inequalities — independnet random variables are often easy to analyze, but you may not get independence; instead your variables may all be positively corrleated with each other. But that's also not that bad — we'll see results you can get in this case. We'll then see the Janson inequalities, which allows you to understand for example the following:

> **Question 3.15.** Given a random garph, where every edge appears with probability p, what is the probability there are no triangles?

Triangles aren't independent — they coudl overlap — so we need additional tools, and Janson's inequalities are a tool for that.

We'll then see concentration of measure — in some way the Chernoff bound and second moment method are examples. There are some counterintuitive statements in high dimensions:

> **Fact 3.16** — If we take a very high-dimensional sphere and think about a Lipschitz function on the sphere (it has small derivatives), this function is almost constant.

In 3D we can think of all sorts of crazy stuff but that doesn't hapepn in high dimensions.

This part of the course will be less about trying to inject rnadomenss into combinatorial problems, and more about how to analyze that randomenss, which is important if you want to apply things.

We'll then come to the ENtropy method and use that to analyze combinatorial problems — there are neat applications. Finally if there's time we'll look at a recent method, the Container method — when you try to apply union bounds, if there's too many events it won't work. But sometimes you don't have to apply the union bound naively — you can first bucket the events into containser, so there's a small number of containers, and *then* you can apply teh union bound. This is recent (in the past decade) and quite influential.

> **Remark 3.17.** Administrative and logistics things:
>
> The course homepage is listed above. There is a textbook which you should take a look at. Prof. Zhao's lecture notes are publicly available. The problem sets are an important part of this class (it's a very problem-solving type of subject — you learn what's going on by practicing). There will be 6 problem sets. Each consists of some unstarred an starred problems. The unstarred problems are meant to practice the techniques demonstrated in class. Some are tricky but most should be doable once you understand what's going on in class. The starred problems are more challenging. You should try them especially if you want a deeper understanding or enjoy solving such problems.
>
> There are no exams, grades are entirely based on problem sets. For grades up to A-, only the unstarred problems count. If you do all those problems, you have a good undrestanding of course materia. To get A or A+ you have to solve a significant number of starred problems.
>
> You are allowed and encouraged to collaborate but this hsoud be done in a genuine way (you cannot split up the problem set, for example).

Remark 3.18. Students always wish they had started on the problem sets earlier.

Remark 3.19. This semester Joel Spencer will give a special lecture, on October 19th. (This lecture will be open to a broader audience.)

Remark 3.20. If you want to go beyond the topics of this class, you may be interested in the combinatorics seminar — research seminar primarily for grad students, there are regular resesarch presentations in the seminar, starting today at 4:15. This is typically WF but you should google it; there's a mailing list to get announcements.

Remark 3.21. If you have questions, don't email; ask after class or in breaks or in office hours, which will be announced.

§4 September 12, 2022

Last time we introduced the probabilistic method. It can be used to show some objects exist via a random construction, and showing that such a random constructino succeeds with positive probability. THat'll be a common theme, but we'll also see other variants.

Today we'll begin with a few gems in the probabilistic method, which have to do with set systems.

There's an area of combinatorics called *extremal set theory*, concerning how big a set system can get satisfying certain properties.

Definition 4.1. A **set system** $\mathcal{F}$ is a collection fo subsets of a ground set, usually $[n]$.

We wish to understand how big such a system can get under some constraints.

§4.1 Sperner's Theorem

Question 4.2. What is the maximum number of sets in an *antichain* of subsets of $[n]$?

An *antichain* is a collection where there are no two sets, one containing the other.

Example 4.3

$\{\{1,2\},\{1,3\},\{2,3,4\}\}$ is an antichain. $\{\{1,2\},\{1,3\},\{3\}\}$ is *not* an antichain.

One way to get lots of sets is to take all the subsets of a given size — take all k-element subsets for some fixed k. This has size $\binom{n}{k}$. We get to choose k, so picking one of the middle k — $k = \lfloor n/2 \rfloor$ or $k = \lceil n/2 \rceil$ — maximizes $\binom{n}{k}$.

This gives a construction for a large antichain. Sperner's Theorem tells us that we cannot do better. It's a very important result in combinatorics.

Theorem 4.4 (Sperner's Theorem (1928))

Every antichain of subsets of $[n]$ has size at most $\binom{n}{\lfloor n/2 \rfloor}$.

We will in fact prove a slightly stronger result:

Theorem 4.5 (LYM Inequality)
If $\mathcal{F}$ is an antichain of subsets of $[n]$, then

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

LYM stands for three people's names who discovered it independently.

To see why the LYM inequality implies Sperner's Theorem, note that $\binom{n}{|A|} \leq \binom{n}{\lfloor n/2 \rfloor}$. So each individual summand contributes at least $\binom{n}{\lfloor n/2 \rfloor}$, which means

$$\frac{|F|}{\binom{n}{\lfloor n/2 \rfloor}} \leq \sum_{A \in \mathcal{F}} \frac{1}{\binom{n}{|A|}} \leq 1.$$

This has a super clever proof using the probabilistic method.

Proof. Consider a random permutation $\sigma(1)$, . . . , $\sigma(n)$, and consider the chain

$$\varnothing \subset \{\sigma(1)\} \subset \{\sigma(1), \sigma(2)\} \subset \cdots \subset \{\sigma(1), \ldots, \sigma(n)\}.$$

For each subset $A \subseteq [n]$, let E_A be the event that A is an element of this chain. We can evaluate $\mathbb{P}(E_A)$ — E_A is the event that all elements of A appear first in σ, followed by non-elements of A. There are $|A|! \cdot (n - A)!$ ways to have such a permutation, so

$$\mathbb{P}(E_A) = \frac{|A|!\,(n - |A|)!}{n!} = \binom{n}{|A|}^{-1}.$$

Now the key observation is that if A and B are distinct elements of our antichain, then $A \not\subset B$ and $B \not\subset A$. So the two events E_A and E_B are *disjoint* — we cannot have both events happen at the same time. THis is because if A and B can both be found in the chain, then one necessarily contains the other.

SO as a result, we have a bunch of disjoint events. This means their probabilities can add to at most 1. This means

$$\sum \mathbb{P}(E_A) = \sum \binom{n}{|A|}^{-1} \leq 1,$$

which is exactly the theorem. □

To recap, we wanted to prove that in an antichain, the inequality holds. The clever part of this proof was to start with a random permutation of the ground set, and note that because we have an antichain, the event that A appears in our random chain is disjoint from other such events (we cannot have two elements of the antichain both appear in our random chain). That lets us conclude that the sum of their probabilities is at most 1, and therefore deduce the desired claim.

How does one come up with such an idea? It is very clever, and we should appreciate the beauty of this proof. BUt in this case, the result is *tight* — the bound we get is exactly the maximum. In general, with the probabilistic method, most of our results will be asymptotic in nature. So this is somewhat unusual.

Let A_1, . . . , A_m be r-element sets and B_1, . . . , B_m be s-element sets, such that $A_i \cap B_j$ is $\varnothing$ if $i = j$, and $A_i \cap B_j$ is nonempty if $i \neq j$.

Question 4.6. How big can m get?

Example 4.7

Suppose A_i ranges over all r-element subsets of $[r+s]$, and $B_i = [r+s] \setminus A_i$. Then the properties are satisfied — each A_i is disjoint from each B_i, and each A_i intersects all the other B_j. (There's only enough room to fit its own complement.) So i goes from 1 to $\binom{r+s}{r}$.

Theorem 4.8 (Bollabás Families Theorem (1965))

$m \le \binom{r+s}{r}$.

The original paper is written because of applications ot udnerstanding transversals — suppose you have a bunch of sets (such as boxes in the plane) and you would like to hit all of them using as few points as you can. This theorem says something about that.

We will prove a slightly stronger version:

Theorem 4.9

Suppose $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$ are finite sets such that $A_i \cap B_i = \varnothing$ and $A_i \cap B_j \ne \varnothing$ when $i \ne j$. Then

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

This implies the theorem, but it *also* implies LYM and Sperner — if $A_1, \ldots, A_m$ form an antichain of subsets of $[n]$, then taking $B_i = [n] \setminus A_i$ satisfies the hypothesis. So maybe it is unsurprising that the proof is similar.

Proof. Consider a random ordering of all the elements, and let E_i be the event that all elements of A_i appear before all elements of B_i in the ordering. We have

$$\mathbb{P}(E_i) = \binom{|A_i| + |B_i|}{|A_i|}^{-1}.$$

Meanwhile, these events E_i are all disjoint — suppose we have an event where all the A_1's appear before all the B_1's. Then we cannot have all the A_2's before the B_2's — this would break one of the intersection conditions. At this point we are basically done, since then

$$\sum \mathbb{P}(E_i) = \sum \binom{|A_i| + |B_i|}{|A_i|}^{-1} \le 1.$$

□

This has many extensions. For example, one extension is that you can replace this by the weaker condition $i < j$. That does *not* follow from this proof. However, there are other important proofs using linear algebraic methods, that allow even more extensions (such as various vector space extensions).

Student Question. Is it clear a priori that m is bounded?

Good question, left to us to figure out.

We see a theme that we start with a problem and inject some randomness. It may be quite surprising what kind of randomness we inject. On the homework we will have to come up with our own sources of randomness to inject.

§4.2 Intersecting Families

Definition 4.10. A family $\mathcal{F}$ is **intersecting** if $A \cap B \neq \varnothing$ for all $A, B \in \mathcal{F}$.

Question 4.11. How large can an intersecting family be?

Example 4.12

What is the size of a largest intersecting family of subsets of $[n]$?

You can take all sets with size greater than $\frac{n}{2}$. You can also take all sets that contain 1. If you take a subset of cardinality 3, you can also take all subsets containing at least two of these. So there are lots of constructions.

The second has size 2^{n-1}. The first sometimes gets 2^{n-1} (when n is odd), and does a bit worse when n is even.

Theorem 4.13

Every intersecting family of subsets of $[n]$ has size at most 2^{n-1}.

This is not hard — pair up A with its complement. We can have at most one in our intersecting family.

A trickier question is the followign:

Question 4.14. What is the largest intersecting family of k-element subsets of $[n]$?

So now we are restricting to only considering k-element sets.

We can still use the second construction — all sets containing 1. The cardinality is $\binom{n-1}{k-1}$.

Question 4.15. Is this the best we can do?

The answer is no. If $k > n/2$, then you can just take all $\binom{n}{k}$ sets. So we don't actually need to do anything.

Theorem 4.16 (Erdős–Ko–Rado Theorem (1961, 1938))

If $n \geq 2k$, then every intersecting family of k-element subsets of $[n]$ has size at most $\binom{n-1}{k-1}$.

So we cannot do better than picking a single element and considering all subsets containing it.

There are many proofs known but the prettiest is the probabilistic one.

Proof. Consider a random *cyclic* ordering of the elements — we order $[n]$ on a circle uniformly at random. For each k-element subset $A \subseteq [n]$, we say that A is *contiguous* if all elements of A form a contiguous block on the circular ordering. We consider the events that each individual set from this collection is contiguous.

Let E_A be the event that A is contiguous. Then

$$\mathbb{P}(E_A) = \frac{n}{\binom{n}{k}}.$$

So the expected number of contiguous sets is

$$\frac{n\,|\mathcal{F}|}{\binom{n}{k}}.$$

We can get k by shifting our window one step a bunch of times. We could also shift it to the right. But if you include all the sets drawn, those aren't intersecting — the blue ones pair up with the yellow ones. (If we start with 7 to 10, 4 to 7 pairs with 8 to 11.) So you cannot do better than simply taking k.

This means the expected number of contiguous sets is also at most k. So

$$\frac{n\,|\mathcal{F}|}{\binom{n}{k}} \le k,$$

which means

$$|\mathcal{F}| \le \frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1}.$$

□

These are very pretty and clever proofs. They were historically important, but they are also beautiful to look at.

We will now see a few more proofs which are more down-to-earth (in that you should be able to come up with them).

§4.3 k-Uniform Hypergraphs

Definition 4.17. A k-uniform hypergraph $H = (V, E)$, abbreviated k-graph, is a finite set of vertices V and a set E of edges, where each edge is a k-element subset of V.

This is synonymous with set systems, but from a different angle.

We will consider coloring the vertices of this hypergraph.

Definition 4.18. We say H is r-colorable if V can be colored using r colors, Such that no edge is monochromatic.

When $r = 2$ this corresponds to the usual notion of colorings.

Definition 4.19. $m(k)$ is the minimal number of edges in a k-graph that is not 2-colorable.

Example 4.20

$m(2) = 2$ — if we have two edges we can always color, but if we have three edges we could have a triangle.

Example 4.21

$m(3) = 7$, with the construction the *fano plane* — take the seven vertices to be a triangle, midpoints, and center, and draw the edges, medians, and incircle. Every 6-edge 3-uniform hypergraph is 2-colorable by exhaustive search.

Example 4.22

$m(4) = 23$, by exhaustive computer search.

We will now see some asymptotic bounds on this quantity.

One way to interpret this is that we're aksing for the minimal amount of structure so that we can force a conflict.

Theorem 4.23 (Erdős 1964)
$m(k) \geq 2^{k-1}$ for all $k \geq 2$.

In other words, every k-uniform hypergraph with fewer than 2^{k-1} edges is 2-colorable.

The lesson from last lecture is that in the absence of other information, we should do things uniformly at random. So let's 2-color at random — color the vertices uniformly and independently at random (black and white).

The probability that a specific edge ends up monochromatic is exactly 2^{-k+1}, since there are k elements in the edge which need to end up the same oclor. Since there are less than 2^{k-1} edges, the probabilitiy that *some* edge is monochromatic (which we don't want), by the union bound, is at most

$$\#\text{edges} \cdot 2^{-k+1} < 1.$$

So there exists a good coloring.

Later we will show a somewhat better bound using a more clever method of coloring than uniformly at random.

Now we will show an upper obund. Surprisingly, ti will also come from the probabilistic method.

Theorem 4.24 (Erdős 1964)
$m(k) = O(k^2 2^k)$.

In other words, there exists a k-graph with $O(k^2 2^k)$ edges that is not 2-colorable.

Phrased in this way, we're still trying to prove existence; so let's construct the k-graph randomly.

Proof. Let the vertex set have size $n = k^2$ (this choice will come out of the proof), and let H be the k-graph obtained by choosing m edges $S_1, \ldots, S_m$ independently and uniformly at random. (We are choosing with replacement, so we allow them to not all be distinct; but that only helps us, since we are trying to construct a graph with not too many edges.)

Given a 2-coloring $\chi \colon V \to \{0, 1\}$, let's consider the probability that χ is a valid 2-coloring of H. First, let's find

$$\mathbb{P}(S_1 \text{ is monochromatic under } \chi).$$

Note that S_1 is random; χ is not random. This depends on what χ is — if χ colors everything the same, then S_1 is definitely monochromatic. Suppose χ has a black and b white vertices. Then

$$\mathbb{P}(S_1 \text{ monochromatic}) = \frac{\binom{a}{k} + \binom{b}{k}}{\binom{n}{k}}.$$

But we know $a + b = n$, so using convexity, this is at least

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

The second factor is roughly constant (it converges to a limit) for our choice of $n = k^2$. SO this is around $c \cdot 2^{-k}$.

Now the probability that χ is a proper coloring, meaning that there's no monochromatic edges, si the probability that none of the edges are monochromatic. But tehse edges are independent, so this is

$$(1 - \mathbb{P})^m \leq (1 - c \cdot 2^{-k})^m,$$

But $1-x \le e^{-x}$, so for a single coloring we get a bound $e^{-c2^{-k}m}$.

Now we get a union bound. The probability that the random hypergraph we constructed has a proper coloring is at most

$$\sum_{\chi} \mathbb{P}(\chi \text{ is a proper coloring}) \le 2^n \cdot e^{-c \cdot 2^k \cdot m}.$$

By choosing $m = Ck^2 2^k$ for a large constant C, recall taht $n = k^2$. So then we find that this quantity is strictly less than 1 — in other words, the random hypergarph with positive probabiltiy does not have a proper coloring, so there exists a random hypergraph with teh prescribed parameters that is nto 2-colorable. This is exactly what we are looking for — we get a hypergraph with at most m edges (which only helps us, since we want a small number of edges). □

The idea is that the individual events are all not hard to analyze, but we want to do bookkeeping to understand what is and isn't random. We're not doing a random coloring — we want to check every coloring.

To recap, we looked at how many edges we had to put into a k-graph to make it not 2-colorable. The first part shows that if you have few — less than 2^{k-1} edges — then you can find a 2-coloring, by randomly coloring. In the second, we showed that you can find one which isn't 2-colorable with not too many edges, and we als oconsidered this at random by throwing down edges at random. But this proof is slightly more involved or tricky to think about, partly because we need to union bound over all colorings.

Combining these, we see that $m(k) \approx 2^k$, but there is a leading factor. We don't actually know precisely what it is. Later on we will see a proof that $m(k)$ is bounded below by around $\sqrt{k/\log k}2^k$, so the leading factor is roughly between $\sqrt{k}$ and k^2. But it's an open problem to determine the right order of growth.

§5 Linearity of Expectations

Theorem 5.1

For random variables $X_1, \dots, X_n$ and constants $c_1, \dots, c_n$,

$$\mathbb{E}[c_1X_1 + \cdots + c_nX_n] = c_1\mathbb{E}[X_1] + \cdots + c_n\mathbb{E}[X_n].$$

Note that this is true even if the X_i are not independent! This is very powerful, and we will spend the next few classes exploring how to use it.

Example 5.2

What's the expected number of fixed points in a random permutation of $[n]$?

This is a basic problem in enumeration — in fact, there's a whole subject about such problems, with powerful and fancy methods. But this problem has a one-line solution:

Solution. For each i, the probability that i is a fixed point is exactly $\frac{1}{n}$. Let the event that i is fixed be E_i. Then the number of fixed points is $1_{E_1} + \cdots + 1_{E_n}$, which means

$$\mathbb{E}[\# \text{ fixed points}] = \mathbb{E}[1_{E_n}] + \cdots + \mathbb{E}[1_{E_n}] = n \cdot \frac{1}{n} = 1.$$ □

§5.1 Hamilton Paths in Tournaments

Definition 5.3. A **tournament** is a directed complete graph.

We can imagine people playing in a tournament, with edges from the winners to losers.

Definition 5.4. A **Hamilton path** is a directed path through all vertices.

Question 5.5. What is the maximum and minimum number of Hamilton paths in a n-vertex graph?

The minimum is easy — take a *transitive* or *linear* tournament, where there's only one tournament. Meanwhile, every tournament has at least one Hamilton path. (This is a fun exercise; there is a hint in the lecture notes.)

The maximum problem is more difficult, and more interesting.

Theorem 5.6 (Szelle 1943)
For every n, there exists a n-vertex tournament with at least $\frac{n!}{2^{n-1}}$ Hamilton paths.

Trying to design by hand a tournament with many Hamilton paths is quite difficult. Rather, the way to do this is via the probabilistic method.

Proof. Consider a random tournament — we start with K_n and orient each edge uniformly at random. Then

$$\mathbb{E}\#\text{Hamilton paths} = \frac{n!}{2^{n-1}},$$

since there are $n!$ possible Hamilton paths and each one is directed correctly with probability 2^{n-1}. So there exists a tournament with at least this many Hamilton paths. □

This is actually historically the first example of an application of the probabilistic method (even predating Erdős on lower bounds for the Ramsey numbers).

Question 5.7. Is this bound tight?

This question is more difficult, but we know an answer!

Theorem 5.8 (Alon 1990)
Every tournament has at most $\frac{n!}{(2-o(1))^n}$ Hamiltonian paths.

So our random tournament is basically the best possible, which is somewhat surprising. This proof uses an important inequality which we will see later on in the course, when discussing the entropy method.

In this example, the source of randomness is pretty standard. Many more clever applications have more clever sources of randomness.

§5.2 Sum-Free Subsets

Definition 5.9. A **sum-free set** is a set A such that there do not exist $a, b, c \in A$ where $a + b = c$.

Question 5.10. Does every set of integers contain a large sum-free subset?

Theorem 5.11 (Erdős 1965)

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

This is pretty cool, because the set of integers you start with can be arbitrary — it can have very nasty structure. But nevertheless we can find a third of it which is sum-free.

The proof of this theorem is relaly a gem.

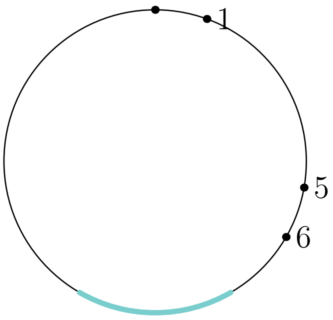
Proof. We use the probabilistic method, but the source of randomness is not at all obvious. (If you keep every integer at random, that's not a good idea.)

Let $A \subset \mathbb{Z} \setminus \{0\}$ with $|A| = n$. Now for a real number $\theta \in [0, 1]$, consider the following procedure: we try to embed A on a circle by going cyclically. So we define

$$A_\theta := \{m \in A \mid \{m\theta\} \in [1/3, 2/3].\}$$

Here $\{m\theta\}$ denotes the fractional part of $m\theta$.

Suppose we start with $A = \{1, 5, 6, \ldots\}$, and we start with some θ. Then we think about embedding A onto the circle:



We can wrap around when we write numbers on thsi circle. THen we want to keep the elements which lie in the middle third.

Note that A_θ is sum-free — if we had $a + b = c$ in A, then $a_\theta + b_\theta = c_\theta \mod 1$. So we'd have three elements in the middle third where two add up to the third, and that's impossible — essentially, our interval is itself sum-free.

So we've found a sum-free subset, but we want to make sure we get a *large* sum-free subset. If we choose θ to be really small, then *nothing* is in A_θ, so this is not guaranteed.

But now we choose θ uniformly at random in $[0, 1]$. Then we can see that for a fixed integer m, $\{m\theta\}$ is also uniformly in $[0, 1]$. So this event occurs with probability $\frac{1}{3}$ — each element of A lies in A_θ with probability exactly $\frac{1}{3}$. By linearity of expectation, $\mathbb{E}[|A_\theta|] = \frac{n}{3}$, so then there exists θ with $|A_\theta| \geq \frac{n}{3}$. □

Question 5.12. How good is this bound?

You can already improve this argument slightly — $\mathbb{E}\,|A_\theta| = \frac{n}{3}$, so unless it's constant at $\frac{n}{3}$, you should be able to get something *strictly* greater. But if θ is very small, then A_θ is empty (all the points are embedded very close to 0), and hence there exists θ such that $|A_\theta| > n/3$, and therefore $|A_\theta| \geq \frac{n+1}{3}$. This is only an improvement if $3 \mid n$.

By a very sophisticated argument using Fourier analysis, Bourgain 1997 improved the bound by an extra third — to $\frac{n+2}{3}$. And that's the best bound we know.

There was also a breakthrough Eberhard–Green–Manners 2011 that shows the 3 cannot be improved — there exist n-element sets of integers whose largest su-mfree subset has size at most $\left(\frac{1}{3} + o(1)\right) n$. So if we replace $\frac{1}{3}$ by 0.334, the statement is false.

But there's still a big gap in the second-order term, and it's an open problem to try to get some bound $\frac{n+g(n)}{3}$, where $g(n) \to \infty$.

The proof of the upper bound uses pretty sophisticated technology from Fourier analysis.

§5.3 Turan's Theorem and Independent Sets

Question 5.13. What is the maximum number of edges in a triangle-free graph, or more generally in a K_{r+1}-free graph?

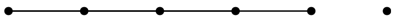
Since taking the complement of a graph changes cliques to independent sets, and vice versa, we can formulate an equivalent problem about independent sets in class.

The result we'll see is the following theorem:

Theorem 5.14 (Caro–Wei)
Every graph G contains an independent set of size at least $\sum_v \frac{1}{d_v+1}$, where $d_v = \deg v$.

In particular, if we mostly have small degrees, we are guaranteed to have some large independent sets.

Proof. Consider a random ordering of vertices.



Now we can read frmo left to right, and try to put vertices into our independent set using a greedy algorithm.

We can put the first vertex into the set. We can put the second in, since it's not adjacent to the first. THen we can't put in the third or fourth because it's adjacent to the first or second, and we can put in the fifth.

This gets us a random independent set I. Now we want to find how large I is.

For every vertex v, the probability that v is in I is kind of tricky — that depends on not just the neighbors, but whether those neighbors got put in. So there's propagations, which takes some work.

So we will modify our step here, and do something even simpler, which doesn't involve a greedy process and just looks locally.

We construct

$$I = \{v \mid v \text{ appears before all its neighbors}\}.$$

Then I must be an independent set — if u and v were adjacent, one would have to appear before the other.

Now the probability that v appears in I is the probabiliyt it appears before all its neighbors, which is $\frac{1}{d_v+1}$. By linearity of expectations,

$$\mathbb{E}\,|I| = \sum \mathbb{P}(v \in I) = \sum \frac{1}{d_v + 1}.$$

□

This proves the Caro–Wei theorem.

Question 5.15. Is this theorem tight?

Yes — it's tight for disjoint unions of cliques. If you think about he proof, you can see that this is the entire equality case.

Question 5.16. Can we do this algorithmically?

This proof tells us to take a random permutation.

Claim — This isn't a good algorithm to guarantee a set of at least this size.

There can be a very heavy tail — a very small probability of getting a huge independent set.

But we can actually get a deterministic algorithm. This is called *derandomization*. Some proofs — particularly those involving linearity of expectation (although not always) can be derandomized.

The general method to do so is to try to do some things greedily, rather than probabilistically.

Start with the vertex with the *least* number of neighbors. Then we choose that vertex, and throw out all its neighbors. But because we chose the vertex with teh *least* number of neighbors, all the summands we threw out add to at most 1. Then we can repeat.

This is a greedy algorithm — include the minimum degree vertex, take it out and take out its neighbors, and repeat. There's no randomness at all. Even though we used a clever setup, our proof is sort of actually doing just that.

Even the very first proof we saw in the class — how to find a very large bipartite subgraph (with size at least $|E|/2$) — can be derandomized. But some examples we don't know how to derandomize at all, such as lower bounds for the Ramsey numbers.

The Caro–Wei inequality allows us to deduce some corollaries about cliques, by taking the complement.

Corollary 5.17

Every n-vertex graph contains a clique of size at least

$$\sum_v \frac{1}{n-d_v}.$$

The equality case is the complement of a disjoint union of cliques — a *complete multipartite graph.*

Now we can answer our earlier question forbidding a clique of a given size.

Theorem 5.18 (Turan's Theorem 1941)

A n-vertex graph that is K_{r+1}-free has a tmost $\left(1-\frac{1}{r}\right)\frac{n^2}{2}$ edges.

The proof essentially follows directly from the corollary, and by doing a convexity argument on the denominator. This bound is essentially tight — partition the vertex set into r parts of equal size (when $r \mid n$ — when $r \nmid n$ they may differ by 1), and plug in a complete multipartite graph (edges between all vertices in different parts). This is a K_{r+1}-free graph with $(1-\frac{1}{r})\frac{n^2}{2}$ edges.

Remark 5.19. There's lots of open problems that are extensions of this; Prof. Zhao wrote a book on such things.

Now we'll move on to a deceivingly similar-looking problem that's notoriously difficult.

Question 5.20. What about hypergraphs?

Question 5.21 (Hypergraph Turán Problem)**.** What is the maximum number of edges in a tetrahedron-free 3-graph?

Recall that edges are triples of vertices. So a tetrahedron is the complete graph on 4 vertices, where all four sets of three vertices are present — the faces of a tetrahedron.

For triangles, we saw that the maximum number of edges is precisely given by the complete bipartite graph with equal parts. Meanwhile here we don't know the answer.

Conjecture 5.22 (Turan) — The answer is $(\frac{5}{9}+o(1))\binom{n}{3}$.

Take three parts, and draw edges where one vertex is in each part, or when two are in the first and one in the second, and cyclically. Then there's no tetrahedron, and you can calculate that the edge density is $\frac{5}{9}$. This is the best construction we know, but we don't know any bound that's close to this.

Part of what makes this problem difficult is that unlike forbidding a triangle, where the extremal graph is unique, here there's a whole *family* of constructions achieving the same bound. This usually makes the problem more difficult, since any argument has to respect the equality cases.

We won't see anything close to the state of the art, but we'll see *some* nontrivial upper bounds — we're seeing this to illustrate a basic technique, called *sampling*.

Consider S to be a 4-element set chosen uniformly at random. If the original 3-graph H has $p\binom{n}{3}$ edges, then the expected number of edges induced by S — there are 4 possible edges, each has a probability p of actually being an edge, so $\mathbb{E}[\#\text{edges induced by } S] = 4p$.

But since H is tetrahedron-free, we also know that #edges induced by$S$ cannot be 4 — if it were 4, we'd see a tetrahedron. So it's at most 3, and putting htis together, we get $p \le \frac{3}{4}$ — i.e. the number of edges is at most $\frac{3}{4}\binom{n}{3}$.

This is the idea of sampling — to get a global upper bound, sometimes it's enough to look locally. Here analyzing a constant sample was enough.

Of course, we're not restricted to taking a 4-vertex subset — what happens if we sample a 5-vertex subset? Then we have to understand the local question — how many edges can be induced by a 5-element subset? This is a bit less obvious. But one trick is to take complements of edges — instead of viewing things as a hypergraph, now we have a graph, and a tetrahedron becomes a vertex of degree 4.

In 5 vertices, we can have at most $\lfloor 3\cdot 5/2\rfloor = 7$ edges. Plugging this in gives an upper bound of $\frac{7}{10}\binom{n}{3}$.

We can keep on playing this — sampling larger and larger subsets, doing analysis on the finite sample, and getting a bound. This bound always gets better and better (or at least, doesn't get worse). In principle you could get arbitrarily close to the truth by analyzing larger and larger samples, but in reality this is too difficult computationally.

Using something more sophisticated but still with the help of a computer — a technique known as *flag algebras* — the best bound currently known is around $(0.561\ldots)\binom{n}{3}$.

§5.4 Unbalancing Lights

Question 5.23. We have a large $n \times n$ square grid. On each entry, there's a lightbulb. Some of the lightbulbs are on, and some are off.

We are allowed to flip a switch along each row or along each column, that changes all the lights i that row or int hat column.

Someone gives you a configuration of lights, and you'd like to do this to get as many lights on as you can. How many lights can you guarantee to turn on?

A naive thing to do might be to flip randomly. This gives $\frac{n^2}{2}$, since every light ends up on or off with probability $\frac{1}{2}$. But we can actually do much better.

Theorem 5.24

Suppose we have a $n \times n$ matrix (a_{ij}) of ± 1. Then there exist $x_i, y_j \in \{-1, +1\}$ such that

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

This is the number of on-lights minus the number of off-lights. If you do randomly, you get expectation 0. But we can actually get positive expectation, on the order of $n^{3/2}$.

If we chose x_i and y_j uniformly at random, the LHS would have expectation 0. So let's not do that.

Let's choose the y_j uniformly at random — so we flip the columns uniformly at random. Now if the first row has 6 ons and 4 offs, we should keep it. If the next row has 3 ons and 7 offs, we should flip it. So each row you decide what to do greedily — we choose x_i to make the ith row sum nonnegative.

Let's set $R_i = \sum_{j=1}^{n} a_{ij} y_j$. Then the resulting sum is $\sum |R_i|$. So we would like to understand $\mathbb{E}[\sum |R_i|]$, which means by linearity we would like to understand $\mathbb{E}\,|R_i|$.

But restricted to a single row, we now have a sum of IID ± 1's — since $a_{ij} y_j$ are all independent for fixed i and varying j. (Between different rows there is correlation, but in rows they're independent).

Then this is easy to analyze — it has a binomial distribution. If S_n is the sum of n iid ± 1's, then $\mathbb{E} \frac{|S_n|}{\sqrt{n}} \to \mathbb{E}\,|Z|$ where Z is a random normal variable in $(0, 1)$, which is $\sqrt{2/\pi}$. So we get $(\sqrt{2/\pi} + o(1))\sqrt{n}$. (There is an explicit binomial coefficient expression that is unnecessarily complicated.)

Now by linearity of expectatino, $\mathbb{E}\,|R| = \sum \mathbb{E}\,|R_i| = n \cdot (\sqrt{2/pi} + o(1))\sqrt{n}$, which concludes the proof.

The moral of this example is that you shouldn't choose everything naively at random — you could introduce some randomness, but leave other things deterministic. Later on, we'll see examples of highly nontrivial uses where we have an intricate combination of randomness and non-randomness.

Next time we'll see a couple of geometric applications of linearity of expectation.

§6 September 19, 2022

Theorem 6.1

Let $k \geq 2$, and $V = V_1 \sqcup V_2 \sqcup \cdots \sqcup V_k$, where the V_i are disjoint and $|V_i| = n$. Consider a complete k-uniform hypergraph on V, and color its edges red or blue. Suppose that every edge that contains one vertex from each V_i is colored blue. Then there exists a subset $S \subseteq V$ such that $|\#\text{red edges in } S - \#\text{blue edges in } S| \geq C_k n^k$, where C_k is a positive constant that only depends on k.

Consider $k = 3$, so we have 3 blobs. Then if we have an edge which is transversal to all three vertex, sets (a triangle witih one vertex i neach), then it's automatically blue.

Example 6.2

When $k = 2$, we have a complete graph, and the edges between the two parts are all colored blue; the edges within a part could be anything.

If all the edges there are colored red, the goal is to find a subset S of vertices where there's a significant difference between the number of red and blue edges. If we took S to be the whole set, that would not be a good choice — the number of blue and red edges differ linearly. We can take one part; that's all red.

Now suppose again all edges between the two parts are blue, but inside each set, it's red or blue randomly. Now the previous choice would not be a good choice. So now we should choose the whole thing.

So the choice of S depends on the graph, and we want to show there exists S with a large discrepancy.

Proof. For notational simplicity we'll illustrate this for $k = 3$, but the proof works in general.

We choose S somewhat randomly — each vertex is chosen to be included in S — but we see that how we choose S may have to depend on the graph itself. So we need to be somewhat flexible.

Let $p_1, p_2, p_3 \in [0, 1]$ to be decided. Then we pick S by including each vertex of V_i with probability p_i.

Now the number of blue edges in S minus the number of red edges in S is a random variable, because S was chosen randomly. So let's evaluate $\mathbb{E}[X]$. We will use linearity of expectations.

For each possible edge, we need to think about what its contribution is to the quantity. By linearity of expectations, we see that this quantity is

$$\sum_{i \le k \le k} (\#\text{blue edges } i\text{-}j\text{-}k - \#\text{red edges } i\text{-}j\text{-}k) p_i p_j p_k.$$

Here if $i, j, k = 1, 2, 3$ then $a_{123} = n^3$ since there are n^3 triples between 1, 2, 3 and all of them are blue. (Here i, j, k are all 1, 2, 3.) Similarly a_{112} is the number of blue edges with two vertices in V_1 and 1 in V_2, minus the number of such red edges.

Let's call this quantity a_{ijk}. Then we have a polynomial f in three variables p_1, p_2, p_3, with some coefficients that come from teh coloring we're initially given.

We know that $a_{123} = n^3$, and all the others are at most n^3 in absolute value.

We shouldn't have chosen p_1, p_2, p_3 in the beginning because then an adversary could come up with a graph where this isn't a good choice, as we've seen. So we're going to choose p_1, p_2, p_3 *after* seeing the graph. We want this polynomial to be large in absolute value — we will be done if the following is true.

Lemma 6.3

Let P_3 be the set of polynomials $g(p_1, p_2, p_3)$ of degree 3 whose coefficients all have absolute value at most 1, and the coefficient of the cross-term $p_1p_2p_3$ is exactly 1. Then there exists p_1, p_2, p_3 in $[0, 1]$ so that $|g(p_1, p_2, p_3)|$ has absolute value at least c_k, where c_k is a constant bigger than 0.

Here 1 is because we divided out by n^3. So given a polynomial of degree 3 with this property, we can find ways to plug in the inputs to this polynomial so that the polynomial evaluates to something bounded away from 0.

If this exists, we can plug it in to get the value of S we want.

How are we going to prove the existence of p_1, p_2, p_3? We'll use a compactness argument.

Proof. Let $M(g) = \sup_{p_1,p_2,p_3 \in [0,1]} |g(p_1, p_2, p_3)|$. Note that sup is actually a maximum — it is achieved due to compactness (g is polynomial and thus continuous, and a continuous function over a compact domain achieves its supremum).

For a given polynomial, since g is nonzero (its coefficient of $p_1p_2p_3$ is nonzero, so it's certainly a nonzero polynomial), we have $M(g) > 0$.

Since P_3 is compact, and M is continuous on the space of polynomials, M attains its minimum. Let the minimum value be C_3, so $C_3 = M(g)$ for some g, and therefore $C_3 > 0$. And we're done. □

So that is a compactness argument showing there exists a positive constant C, and that finishes the proof. □

Student Question. Why is P_3 compact?

View it as a subspace of $\mathbb{R}$ to the number of coefficients.

Student Question. Are there quantitative bounds on the argument?

This proof has the interesting feature that it doesn't give any quantitative bounds — you know there exists a constant but the proof can't give you one. Sometimes it's really hard to convert that into an actual constant. In this case, Prof. Zhao hasn't tried, but he thinks it shouldn't be that hard, but will involve quite a bit of analysis.

Student Question. Why is the minimum value what we're looking for here?

That implies $M(g) \geq C_3$ for all g.

This is kind of sneaky — it's using the probabilistic method but with an unknown probability distribution.

Student Question. Why do we need that the top sup thing is actually a max?

You might be able to get away without it. But you do need to be able to justify that M is continuous.

§6.1 Crossing Number Inequality

In graph theory, we usually consider graphs as combinatorial objects. But usually we like to draw graphs. ANd some drawings may be better than others. FOr example, if you draw K_4 as a quadrilateral two edges might cross, but you might not like that; so draw it as a triangle and center.

Today we would like to understand the **crossing number** $\text{cr}(G)$ — the minimum number of crossings needed to draw the graph in $\mathbb{R}^2$. (The edges have to be continuous curves.)

Definition 6.4. The graph is planar if $\text{cr}(G) = 0$.

The study of planar graphs is a classical topic in graph theory. Here the situation is well-understood:

Fact 6.5 — $K_{3,3}$ and K_5 are each not planar (can be proved by Euler's theorem).

We actually have a complete characterization — knowing $K_{3,3}$ is not planar tells you that some additional stuff are non-planar. Any graph topologically homeomorphic to $K_{3,3}$ is also non-planar — if we divide an edge into a path (adding extra vertices) that's still nonplanar.

Theorem 6.6 (Kuratowski's Theorem 1930)

That is the only obstruction — every non-planar graph contains a subgraph that is topologically homeomorphic to $K_{3,3}$ or K_5.

So there is a subgraph which can be produced by taking $K_{3,3}$ or K_5 and dividing up the edges.

There is a related theorem:

Theorem 6.7 (Wagner's Theorem 1937)

Every non-planar graph contains a $K_{3,3}$ or K_5 minor.

We will not explain what a minor is but it's an important concept in graph theory. (A subgraph is obtained by deleting vertices and deleting edges. A minor allows you to do these, and also to do contractions — we can take an edge and squeeze it together.) In fact, these two theorems are equivalent.

Now we will explore what happens beyond planarity. If we have lots of edges, is it true we have lots of crossings?

Question 6.8. Is it true that lots of edges automatically imply lots of crossings?

Example 6.9

The complete graph has a lot of edges; can you draw it in a way that doesn't have a lot of crossings? What's $\mathrm{cr}(K_n)$ asymptotically?

That is actualy a famous open problem; same with $K_{n,n}$. We will not get exact answers but we will get some easy asymptotics.

If we draw carelessly, we would expect n^4 possible pairs of edges. So we would like a statement like $e(G) \gtrsim n^2$ implying $\mathrm{cr}(G) \gtrsim n^4$. It turns out this is true and will follow from the main theorem we will state.

Theorem 6.10 (Crossing Number Inequality)

In a graph $G = (V, E)$, if $|E| \geq 4|V|$, then

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

A few remarks: why do we need this hypothesis? 4 is really anything bigger than 3. If you have a planar graph you can have average degree around 3 (a hexagonal grid). You can then throw vertices inside the hexagon to make a triangular grid — that has average degree close to 6, so $|E| \approx 3|V|$. So you need some hypothesis like this.

§7 September 21, 2022 — Alterations

We saw the alteration method in the first lecture, when we gave a lower bound for Ramsey numbers.

Alterations involve two steps:

- Take a random construction.
- Fix the blemishes.

Today we will see several more examples of how to do alterations.

§7.1 Dominating Sets

Theorem 7.1

Every n-vertex graph with minimum degree $\delta > 1$ has a dominating set of size at most $()n$.

Definition 7.2. A **dominating set** is a subset $W \subseteq V$ of vertices such that every vertex outside W is adjacent to a vertex of W — W and its neighbors cover the entire graph.

We would like to show that if you have a graph with not-too-small minimal degree, we should be able to dominate the graph by choosing a fairly small subset of vertices.

The first idea we can try is to greedily pick out vertices — let's greedily pick large degree vertices.

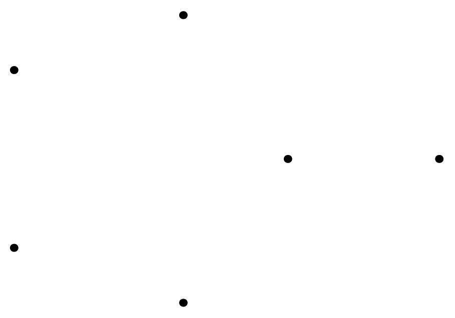
The first vertex and its neighbors will eliminate at least $\delta+1$ vertices, which is pretty good progress. Perhaps we can try to keep going. If every step can eliminate some large number of vertices, then we get a dominating set of size at most $n/(1+\delta)$. But the issue with this approach is when we delete a vertex and its neighbors, the minimal degree could go down; so subsequent steps may not make as much progress.

But this gives us a target in mind — around n/δ would be a good thing to aim for.

Proof. First, let's choose vertices at random, each with probability p (independently), for some p to be decided.

Now we've thrown in some vertices. This may not be dominating, but we can fix that by adding in additional vertices to make the selection dominating.

Now let's try to analyze what happens. Let X be the random subset from the first step. Then we have a graph, and we chose some subset of vertices.



Now let Y be the set $V \setminus (X \cup N(X))$ — the vertices not in X and its neighbors — and note that $X \cup Y$ is a dominating set.

Now let's analyze $\mathbb{E}[X \cup Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. We have $\mathbb{E}[X] = np$. Meanwhile for Y, a vertex is in Y if it's not selected, and neither is any of its neighbors. This set has size at least $1+\delta$. So each vertex has probability *at most* $(1-p)^{1+\delta}$ of being included in Y. This gives

$$\mathbb{E}[X \cup Y] \leq \left(p + (1-p)^{1+\delta}\right) n.$$

Now we want to pick p that makes this as small as possible. We can try to minimize this expression in p by taking the derivative and solving, but usually we don't need such great precision; it's often better to make some estimations that make your life easier without sacrificing the result. One common estimate you should remember is $1 + x \leq e^x$ — you can convert easily between additive errors and exponential errors. We'll use that to bound this quantity —

$$p + (1-p)^{1+\delta} \leq (p + e^{-p(1+\delta)})n.$$

Then you will find by taking the derivative that you want $p = \log(\delta+1)/\delta+1$. Plugging that in, we get

$$\mathbb{E}[X+Y] \le \left(\frac{1+\log(\delta+1)}{\delta+1}\right)n.$$

This is actually pretty close to our incorrect approach — except for this extra log factor. So ew have done pretty well. □

§7.2 Heilbronn Triangle Problem

Problem 7.3. How can one place n points in the unit square $[0,1]^2$ without three points forming a triangle of small area?

This belongs to a general class of problems called *discrepancy* or *hyperuniformity*. When we think of something uniform, we might want nothing clustered together in no particular sense; then it becomes unclear how you should do it.

One naive guess is to place a grid — this would be our guess for placing uniformly. BUt this is bad for our purpose, because we have lots of zero-area triangles. So in some sense we're trying to be more uniform than we can do naively.

Theorem 7.4

For every n, there exists n points in $[0,1]^2$ such that all triangles have area $\gtrsim 1/n^2$.

The $\gtrsim$ means we are omitting an absolute constant factor.

Proof. First we put the points at random, and then we try to fix it. One way to fix is to destroy a vertex of a small triangle.

So we choose $2n$ points uniformly and independently at random, and now we want to understand the probability that some point is involved in a triangle of small area. There are some simple geometric arguments you can do to show that if we have three points p, q, r, then

$$\mathbb{P}(\text{area}(pqr) \le \varepsilon) = O(\varepsilon).$$

Now let X denote the number of triangles with area at most ε, where ε is to be chosen. Then

$$\mathbb{E}[X] = O(\varepsilon n^3).$$

So if we choose ε to be c/n^2 for $c > 0$ small enough, then we can make sure that $\mathbb{E}[X] \le n$. Now we delete a point from every triangle with area at most ε. The resulting set will be free of triangles with small area.

But we want to know that we still have at least n points left. The number of remaining points is at least $2n - X$ — we delete at most one point for each small triangle — which means

$$\mathbb{E}[\text{remaining points}] \ge 2n - \mathbb{E}[X] \ge n.$$

So there exists a configuration with at least n points and no triangle of small area (where "small" means less than $\varepsilon = c/n^2$). □

What happens here is a little bit different from over there. It's a good trick to try to choose a few extra things and delete — and theh point is to check that we haven't removed too many things.

This is still an open problem — what's the biggest number we can replace $1/n^2$ with? The current best bound is $\gtrsim n^{-2}\log n$, by a very careful analysis of this strategy. The current best *upper* bound is $\le n^{-8/7+o(1)}$. Both results are form teh 1980s, and that's the state of the art. There is also an algebraic proof giving an explicit construction with the same n^{-2} bound, but it doesn't leave room for improvement.

§7.3 High Girth and High Chromatic Number

Suppose someone gives you a graph, and you have local access — you can look at a vertex, and its k-radius neighborhood. Can you decide, just from this local information, whether the graph has small chromatic number?

There's a few simple observations:

- If a graph has a k-clique, then its chromatic number is at least k.
- The converse is not true — if you don't have any triangles, you might still not be able to 2-color.
- More generally, if a graph has high chromatic number $\chi(G)$, is it possible to certify this from some local data (looking at some small subgraph)?

This is a pretty interesting problem. It's natural in the study of local graph algorithms.

Surprisingly, the answer is no. The proof we'll give will give a graph with high girth and high chromatic number.

> **Definition 7.5.** The **girth** of a graph is the length of the shortest cycle.

Theorem 7.6 (Erdős, 1959)
For all k and ℓ, there exists a graph with girth greater than ℓ and chromatic number greater than k.

Having high girth is another way of saying that the graph is locally tree-like. If you're only allowed to look at neighborhoods of radius at most $\ell/2$, you will never see anything other than a tree. But despite this graph being locally tree-like, it has high chromatic number. THis is a very important construction, showing that chromatic number is a very intricate statistic that doesn't just depend on local information.

Before Erdős's work, you might have seen some specific constructions — which let you do this for some very specific paramters. But Erdős was the first to realize it could be done for *any* k and ℓ.

Proof. Start with the random graph $G \sim G(n,p)$, the n-vertex graph where every edge is ichosen with probability p. We will choose

$$p = \frac{(\log n)^2}{n}.$$

We will see that as long as

$$\frac{\log n}{n} \ll p \ll n^{-1+1/\ell}$$

we are good. (Note that here $\ll$ means o, not O. In number theory it is used as O.)

A random graph will probably not have high girth. So this is the primary issue we need to fix.

This graph *will* have lots of triangles, and 4-cycles, and lots of other small cycles. Those exist. But we want to get rid of them. So let's analyze how many there are — let X be the number of cycles of length at most ℓ in G. Then to compute $\mathbb{E}X$, we can consider each possible length from 3 to ℓ. For each length i, the number of cycles is exactly $\binom{n}{i}(i-1)!\,p^i$ — we choose our vertices, arrange them in a cycle in $(i-1)!$, and want all i edges to appear. So

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

(The $\frac{1}{2}$ is for cycles going backwards.) Since $p \ll n^{1-1/\ell}$, this is $\ell \cdot o(n) = o(n)$, since ℓ is a constant. So there are very *few* cycles of length at most ℓ — small compared to the totla number of vertices.

We will need the following important tool.

Theorem 7.7 (Markov's Inequality)

If $X \geq 0$ is a nonnegative random variable, then for every positive real a,

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

Proof. $\mathbb{E}[X] \geq \mathbb{E}[X \cdot 1_{X \geq a}]$, since this is true pointwise. But when $X \geq a$, we can replace X by a to get a lower bound — so $\mathbb{E}[X] \geq \mathbb{E}[a 1_{X \geq a}] = a\mathbb{P}(X \geq a)$. □

The message from Markov's inequality is that given a random variable X that is nonnegative, if $\mathbb{E}X$ is very small, then typically X is small. And that is what happens here — X is small in expectation, so the probability X is large is quite small. In particular,

$$\mathbb{P}(X \geq n/2) \leq \frac{\mathbb{E}[X]}{n/2} = o(1).$$

This morally allows us to focus on the case where $X < n/2$, and we can remove all of them without sacrificing too much of the graph.

That's how one can get rid of short cycles — by choosing p small enough, we can get rid of short cycles but still have many remaining vertices.

Now we want to understand the chromatic number. Later we will see much more precise analysis of the chromatic number of a random graph.

The chromatic number is an important statistic, but it's also a *difficult* statistic — having a lower bound on $\chi(G)$ is a statement about what you cannot do, which is difficult to certify. So we need some cheaper ways to lower bound $\chi(G)$.

One of the most useful ways to do this is via its *independence number.*

Lemma 7.8

$\chi(G) \geq \frac{|V|}{\alpha(G)}$, where $\alpha(G)$ is the size of the largest independent set.

This is because in every proper coloring, every color class is an independent set.

The independence number is usually easy to certify.

Choose $x = \frac{3}{p}\log n$ (in hindsight); we will now bound $\mathbb{P}(\alpha(G) \geq x)$. We do this using the union bound: we union over all possible x-element subsets, and for each set we look at the probability there are no edges among the x vertices. So

$$\mathbb{P}(\alpha(G) \geq x) \leq \binom{n}{x}\mathbb{P}(\text{no edges among these } x \text{ vertices}) = \binom{n}{x}(1-p)^{\binom{x}{2}}.$$

Now we use some cheap inequalities to bound these by

$$n^x e^{-px(x-1)/2} = (ne^{-p(x-1)/2})^x = o(1).$$

These are not bad bounds because x is quite small, so $\binom{n}{x}$ is really like n^x.

So with high probability, the independence number of G is small, and therefore the chromatic number must be large.

Now let's put everything together. We have two events which are high probability. So by choosing n to be large enough, we can make sure that

$$\mathbb{P}(X \geq n/2) < 1/2 \text{ and } \mathbb{P}(\alpha(G) \geq x) < 1/2.$$

In the complement, there exists some graph G with fewer than $n/2$ cycles of length at most ℓ, and whose $\alpha(G) \leq \frac{3}{p} \log n$.

To get high girth, we want to destroy our cycles. So we let G' be G with a vertex removed from each short cycle. (Here short means length at most ℓ.) Then we see $|V(G)| \geq n/2$.

Also, $\alpha(G') \leq \alpha(G)$, since G' is an induced subgraph. And we know $\alpha(G) \leq \frac{3}{p} \log n$, so

$$\chi(G') \geq \frac{|V(G')|}{\alpha(G')} \geq \frac{n/2}{\frac{3}{p}\log n} = \frac{\log n}{6} > k$$

if n is large enough. And that finishes the proof — we have constructed a graph G' with girth larger than ℓ and chromatic number larger than k, by starting with a random graph and removing one vertex from short cycles. (If you remove *edges* from short cycles instead, then you run into difficulties.) That takes care of girth, but we need to deal with χ, and the way to do this is to check that with high probability G does not have large independent sets. □

Remark 7.9. This certifies that you cannot certify high $\chi(G)$ from looking at local data. But in fact, it's even worse — another result due to Erdős is tha tfor all k there exists $\epsilon = \epsilon_k > 0$ such that for all large n, there exists a n-vertex graph G with $\chi(G) > k$ but every subgraph on ϵn of the vertices is 3-colorable. SO we cannot distingiush 3-colorabioity vs k-colorability by looking at a sublinear part of the graph — we need to see a constant fraction of the graph! So colorabiltiy is in many ways a difficult quantity to study on graphs in general.

§7.4 2-Coloring a Hypergraph

Previously, we saw the following:

Theorem 7.10

Every k-uniform hypergraph with $< 2^{k-1}$ edges is 2-colorable.

This means we can assign a red-blue coloring to vertices so that no edge is monochromatic.

Proof. Take a uniform random coloring. The expected number of monochromatic edges is less than 1, so there exists some coloring without a monochromatic edge. □

Today we will get a slightly better bound:

Theorem 7.11

Every k-graph with $\leq c\sqrt{\frac{k}{\log k}} 2^k$ edges is 2-colorable.

So we have gained a factor of $\sqrt{k/\log k}$.

Now we cannot use linearity of expectations any more. We also saw that if we replace this with a big constant and k^2, then this is false — there exists a hypergarph with $k^2 2^k$ edges that is not 2-colorable.

The idea is *random greedy coloring*. This bound is quite recent — from 2015. It combines several ingredients we've seen so far.

Proof. Let H be the k-graph with m edges, that we wish to color. Then we map $V(H) \to [0,1]$ independently and uniformly at random.

Now let's color the vertices greedily, from left to right. (So far we are only using the order.) We color blue *unless* this would create a monochromatic (blue) edge; and if so, we color red.

This might not always work — the resulting color has no blue edges by design, but it might have red edges. Then each of our vertices which are red come from a configuration (BB(R)RR) — in particular, there must be two edges such that the last vertex of e equals the first vertex of f. Call such (e,f) *conflicting* — given an ordering of vertices.

If an ordering results in no conflicting edges, then our greedy algorithm always succeeds. So we would like to show that with positive probability, we have no conflicting edges.

Given two edges with exactly one vertex in common,

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

So if we union bound, we only need to union bound over edges that have one vertex in common. Let's be generous, and union bound over all $\le n^2$. Then the probability that there exists a conflicting pair si

$$\mathbb{P}(\text{exists a conflicting pair}) \lesssim m^2k^{-1/2}2^{-2k}.$$

This gives us $m \le k^{1/4}2^k$.

So this is nto quite as good as what we claimed, but it is something — it is better than what we got using a much more naive argument (and this was a previous result before the proof we are about to see). So far, we have not accessed the interval; now we are going to access the interval.

Split the interval into three parts L, M, and R, where we chop off at $\frac{1-p}{2}$, and $\frac{1+p}{2}$. Then

$$\mathbb{P}(\text{a given edge is contained entirely on L}) = \left(\frac{1-p}{2}\right)^k.$$

This is the same as hte probability an edge si contained in R. (These are both supposed to be small quantities.) So

$$\mathbb{P}(\text{some edge lies entirely in } LorR) \le 2m\left(\frac{1-p}{2}\right)^k.$$

Now suppose that there are no edges entirely contained in L or R. Then two edges which conflict must have their common vertex in the middle.

For a given (e,f) with one vertex in common,

$$\mathbb{P}(e, f\text{conflict}) = \int_{(1-p)/2}^{(1+p)/2} x^{k-1}(1-x)^{k-1}\,dx$$

by thinking about where the common vertex ends up — if it ends up in x then all the remaining vertices have to be left of x, and in f right of x. Using that $x(1-x) \le \frac{1}{4}$, we get a bound

$$p \cdot 4^{-k+1}.$$

Therefore putting these together,

$$\mathbb{P}(\exists\text{conflicting pair}) \le 2m\left(\frac{1-p}{2}\right)^k + m^2p4^{-k+1}.$$

Once again, we convert $1-p$ to e^{-p} and get the bound

$$2^{-k+1}e^{-pk} + (2^{-k+1}m)^2p.$$

Taking the derivative, we can now set $p = \frac{\log(2^{-k+2}k/m)}{k}$, and this is less than 1 for $m = c2^k\sqrt{k/\log k}$. □

To recap, we color blue as long as it doesn't create a blue edge, and color red otherwise. If this fails, then we have two edges intersecting in exactly one vertex, and that happens to be the last vertex of the left edge and the first vertex of the right edge. So to bound the probability that the coloring fails, we should bound the probability of seeing a conflicting pair.

To bound that, we divide our interval up, and first think about the event where some edge is entirely on the left segment, or entirely on the right segment. In the complement, the shared vertex must lie in the middle, and according to where it lies, we can get a bound by integrating over all the possible places for x. Putting these together, we can choose the parameters so that each occurs with small probability, and there is less than 1 probability of having a conflicting pair. That finishes the proof.

§8 September 26, 2022 — The Second Moment Method

The second moment method is the first time we'll see the idea of *concentration* — we'll use it to show that a random variable is concentrated around its mean. Previously, we've only looked at the *first* moment, the *expectation* of the variable. Now we're going to use the second moment to show concentration.

Question 8.1 (Motivating Question)**.** We have a random graph $G(n, p)$ (the Erdős–Rényi random graph, with n vertices and each edge drawn with probability p), where p may depend on n. We'd like to know, does $G(n, p)$ contain a triangle with high probability?

If you draw $G(n, p)$, will we typically find a triangle in the graph?

We will commonly see "with high probability" (sometimes abbreviated whp) — this means with probability $1 - o(1)$, or probability approaching 1 as $n \to \infty$. So we're talking about a *sequence*, not just one graph (this is also called "asymptotically almost surely").

Using the tools we already have, we can compute $\mathbb{E}[\#\text{triangles}]$. Let $X = \#\text{triangles}$. We can compute its first moment using linearity — there are $\binom{n}{3}$ triples and each appears with probability p^3, so

$$\mathbb{E}[X] = \binom{n}{3} p^3 \asymp n^3 p^3$$

(the notation $\asymp$ means Θ — in other words, bounded below and above by a constant).

So if $np \to 0$, then the expected number of triangles goes to 0, and hence by Markov's inequality (which we saw last time), the probability that there is at least one triangle is at most

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

So the probability that there is any triangle becomes very small — in other words, if $np \to 0$, then $G_{n,p}$ is triangle-free with high probability.

Then a natural question is what happens when $np \to \infty$ — when $p \gg 1/n$? Looking at this calculation, we see that if $np \to \infty$, then $\mathbb{E}[X] \to \infty$.

Question 8.2. Can we conclude that typically there is at least one triangle?

The answer is no — we cannot yet conclude $X > 0$ with high probability, since we have not ruled out the possiblity that this occurs very rarely, but when it occurs, X is extremely large — that would be consistent with having high expectation. So we don't have enough information to conclude that X is positive with high probability.

The tool we'll see today *will* allow us to conclude this. We'll see that not only does X have a large mean, but X is typically close to $\mathbb{E}[X]$ in some precise, quantitative sense; and that *will* imply X is positive with high probability.

This is the idea of *concentration inequalities*, of which the second moment method is one of the most basic and useful methods.

It's useful for many purposes, but it also gives fairly modest bounds. Later we will see other concentration inequalities that are more powerful but require more assumptions.

The idea is to show $\mathbb{E}[X]$ is typically near the mean by bounding its *variance* (which is related to the second moment).

Definition 8.3. Given a random variable X, we define its **variance** to be

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

Expanding the square and applying linearity of expectations, we get the equivalent formulation

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

Another important statistic we'll see is the *covariance*:

Definition 8.4. The **covariance** between two random variables X and Y is

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

We can see that if X and Y are independent then $\mathrm{Cov}(X, Y) = 0$.

First let's see why it helps to bound the variance of a random variable.

SOme conventions: we will use μ for the mean and σ^2 for the variance; σ is sometimes called the **standard deviation**.

The tool for bounding tail probabilities we'll use is *Chebyshev's Inequality*:

Theorem 8.5 (Chebyshev's Inequality)

If X is a random variable with mean μ and variance σ^2, then for any $\lambda > 0$,

$$\mathbb{P}(|X - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}.$$

Proof. Use Markov's inequality — squaring both sides and using Markov's gives

$$\mathbb{P}((X - \mu)^2 \geq \lambda^2) \geq \frac{\mathbb{E}[(X - \mu)^2]}{\lambda^2\sigma^2} = \frac{1}{\lambda^2}. \qquad \square$$

Here we *only* know the variance, adn no other information. There are some random variables for which this is basically teh best you can deduce. If you have additional information — such as it being the sum of independent random variables — then you can deduce much better bounds. Bounds of this form are called **tail bounds**. If you have higher moments, you can get better bounds, using the same proof; later we will see the Chernoff bound, which is exponential decay (intuitively by taking arbitrary moments).

Corollary 8.6

For any X,

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

Proof. First, if $X = 0$ then X is certainly quite far away from its mean —

$$\mathbb{P}(X = 0) \le \mathbb{P}(|X - \mu| \ge \mu) \le \frac{\sigma^2}{\mu^2}$$

by Chebyshev's inequality. □

So if our variance is significantly smaller than our mean squared, we can derive that the probability of non-existence is small. And that is what we will use for the rest of today's lecture — we will check via calculation that this quantity is indeed small when X is the number of triangles in a random graph.

Often this quantity will not just be *somewhat* small, but *very* small; you can use the same inequality to deduce that X is *concentrated* around its mean, but we won't focus on that so much. More explicitly:

Corollary 8.7

If $\mathbb{E}X > 0$ and $\operatorname{Var} X = o(\mathbb{E}X)^2$, then $X > 0$ with high probability.

Unlike the statements we've written before, this is an asymptotic statement; it's not really about a single random variable, but a sequence of random variables X_n. If all have positive mean and this asymptotic holds for the sequence, then $X > 0$ with probbaility approaching 1.

All that is great, but now we need to know how to compute $\operatorname{Var}(X)$. The nice thing about the second moment method is that the variance is often quite straightforward to compute.

The reason why variance is quite easy to compute is that *covariance* is bilinear:

Theorem 8.8

$\operatorname{Cov}[a_1X_1 + \cdots, b_nY_1 + \cdots] = \sum_{i,j} a_ib_j \operatorname{Cov}[X_i, Y_j]$.

Often, we'll be dealing with functions where X is counting something; then we can write X as a sum of *indicator* random variables, and split using bilinearity. Note that there are *no assumptions* on the random variables here — just like linearity of expectations, this is true always. (You can see that $\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$ is bilinear in X and Y — if you fix Y it's linear in X, and vice versa.)

So if X counts something, we can write $X = X_1 + \cdots + X_m$, and compute

$$\operatorname{Var} X = \operatorname{Cov}[X, X] = \sum_{i,j} \operatorname{Cov}[X_i, X_j].$$

It will often be quite easy to compute this object.

Where you gain is that this thing here is often qiute small. It can be small for different reasons — in the case of triangles, two edge-disjoint triangles are independent and contribute 0, so most terms are 0, and there will be very few terms which are nonzero. That's why the variance is small in this case.

We will also see different aplications where the covariance individually are never 0, but are always small — maybe you only have a weak amount of interactions.

First let's do a simple computation to convince ourselves why this is a good idea.

Example 8.9

Take $X = X_1 + X_2 + \cdots + X_n$ to be the sum of n independent Bernoulli variables $X_i \sim \text{Ber}(p)$ (so X_i is 0 or 1, and 1 with probability p). Then $\mathbb{E}X = np$, and

$$\text{Var}(X) = \sum \text{Cov}[X_i, X_j].$$

But all the cross-terms are 0 because the X_i and X_j are independent, so we're left with

$$\text{Var}(X) = \sum_i \text{Var}(X_i) = n(p - p^2).$$

So if $np \to \infty$, we see that $\sigma = o(\mu)$, and by applying Chebyshev we see that X is concentrated around its mean with high probability.

We actually know much better concentration for that example (via the Chernoff bound, which we'll see next chapter).

Now let's return to triangles. Here we do not have a sum of independent random variables, but we can still use this method.

Theorem 8.10

If $np \to \infty$, then $G(n, p)$ contains a triangle with high probability.

Proof. Label the vertices by 1 through n, and let X_{ij} be the indicator random variable for the edge ij — X_{ij} is 1 if ij appears as an edge in the random graph and 0 otherwise.

It'll also be helpful to write indicators X_{ijk} for triples, so X_{ijk} is the indicator random variable that's 1 if ijk is a triangle and 0 otherwise.

We would like to compute Var X, where X is the total number of triangles. So

$$X = \sum_{i<j<k} X_{ijk} = \#\text{triangles},$$

and we can compute this by expanding the covariance with itself by bilinearity. To compute each term, let's first analyze what happens term by term: suppose T_1 and T_2 are 3-vertex sets, and we want to find $\text{Cov}[X_{T_1}, X_{T_2}]$ corresponding to two triangles that can come up. Then there are a few possibilities: recall that

$$\text{Cov}[X_{T_1}, X_{T_2}] = \mathbb{E}[X_{T_1}X_{T_2}] - (\mathbb{E}X_{T_1})(\mathbb{E}X_{T_2}).$$

The second term is always $p^3 \cdot p^3$. But the first term depends on how the two triples are situated. Then $\mathbb{E}[X_{T_1}X_{T_2}]$ is p raised to the number of total edges. There are a few cases:

- If T_1 and T_2 share at most one vertex — $|T_1 \cap T_2| \le 1$, so that they're edge-disjoint, then we have 6 edges and p^6.
- If they share two vertices, then we have a diamond configuration with 5 edges, so the answer is p^5.
- If the triangles coincide exactly, then we have 3 edges and we get p^3.

So then $\text{Cov}[X_{T_1}, X_{T_2}]$ is either 0, $p^5 - p^6$, or $p^3 - p^6$ depending on the number of vertices where T_1 and T_2 intersect.

Now we can plug the individual term calculation itno the sum, to complete our claculation fo the variance — thus the variance is

$$\text{Var}\, X = \sum_{T_1, T_2} \text{Cov}[X_{T_1}, X_{T_2}].$$

We only need to worry about the second and third possibilities, since the first contributes 0. We have

$$\operatorname{Var} X \lesssim n^3 \cdot (p^3 - p^6) + n^4 \cdot (p^5 - p^6)$$

(omitting constant factors, this is the number of ways to choose the vertices). We may as well drop the subtractions as well, so

$$\operatorname{Var} X \lesssim n^3p^3 + n^4p^5 = o(n^6p^6)$$

since $np \to \infty$ (this is true for both terms).

So then $\operatorname{Var} X = o(\mathbb{E}X)^2$, and by what we've just showed, we get that $X > 0$ with high probability. □

The conclusion we see here is that

$$\mathbb{P}(G(n,p) \supseteq K_3) \to \begin{cases} 0 & np \to 0 \\ 1 & np \to \infty \end{cases}.$$

In other words, it approaches 0 if $p \ll 1/n$ and 1 if $p \gg 1/n$. Here we say that $1/n$ is a **threshold** for containing a triangle, in the sense that if $p \ll 1/n$ then with high probability we don't contain a triangle, and if $p \gg 1/n$ then with high probability we do contain a triangle.

This is one of the central types of questions in probabilitsic combinatorics — we have some property we care about (e.g. containing a triangle, connectivity, containing a matching) and we'd like to understand what is the threshold for that property — what is the typical edge density you need so that you expect that property to occur? Next lecture we will have a more detailed discussion of thresholds, but as an example this shows that this is an interesting question, and there's in fact a lot of serious research and open problems on teh topic of thresholds. (For example, the threshold for a graph to be 3-colorable is open.)

We haven't yet defined this precisely, but in this formulation $2/n$ is also a threshold. Next class we will care about constant factors, but right now we will not.

Question 8.11. What happens when $np \to c$?

In other words, $p = c/n$? Here we do know the answer, and it will be on the homework. The method here is that the number of triangles converges to the Poisson distribution. For basic probems in probabilistic combinatorics, there's only two types of limiting distributiosn — Gaussian (normal) or Poisson, and these occur 99% of the time (for integer distributions with bounded mean). This will be proved using the method of moments.

In this case, when $np \to \infty$, the limiting distribution will be Gaussian (asymptotically normal); that will be another homework problem. Later on in the course, we will go beyond Gaussianity and derive ways to prove really good tail boudns for the number of triangles in a random graph, that iwll go much beyond the second moment method.

Question 8.12. What is the threshold for $G(n,p)$ to contain a fixed graph H as a subgraph?

Previously we took H to be a triangle; now it can be an arbitrary graph.

In a way, what we're about to do is straightforward, in the sense that if you understood what happens in triangles, you can do the same calculation. But for an arbitrary H, the different possibilities you have to analyze and keep track of may be difficult.

But there will be a bit of a surprise in what the answer is.

We want to find a function q_n such that if $p_n/q_n \to 0$ then $G(n,p_n)$ does not contain H with high probability, and if $p_n/q_n \to \infty$ then $G(n,p)$ does contain H with high probability. These are sometimes called the 0-statement and the 1-statement.

It should not yet be clear why such a q_n exists. Perhaps you need a gap of $\log n$ before you transition from one to the other — so far we have not excluded that property. But next class we will see that every monotone property has a threhsold — the existence of a threshold is a general abstract phenomenon, but in this case it'll come out of our concrete calculation.

Let's first do some setup. Suppose $X = X_1 + \cdots + X_n$, where X_i is the indicator random variable for the event A_i (so it's 0 or 1, depending on whether A_i occurs); these may not be independent. Then write $i \sim j$ if $i \neq j$ and the events (A_i, A_j) are not independent (A_i is related to A_j); these generally come in the form of overlapping edges, in our context.

If $i \neq j$ and $i \not\sim j$, then A_i and A_j are independent, so $\operatorname{Cov}[X_i, X_j] = 0$. Otherwise,

$$\operatorname{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - (\mathbb{E}X_i)(\mathbb{E}X_j).$$

Now we'll do some simplifications because p is usually quite small — when the thing is nonzero, the second term usually doesn't contribute once, so $\operatorname{Cov}[X_i, X_j] \leq \mathbb{E}[X_i X_j] = \mathbb{P}(A_i A_j)$ (this notation denotes $A_i \cap A_j$).

With that in mind, we see that

$$\operatorname{Var} X = \operatorname{Cov}[X, X] = \sum_{i,j} \operatorname{Cov}[X_i, X_j] \leq \sum \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}[X_i, X_j].$$

The first term is bounded above by $\mathbb{E}X$, and the second term is bounded by

$$\Delta = \sum_{i \sim j} \mathbb{P}(A_i A_j).$$

Then

$$\operatorname{Var} X \leq \mathbb{E}X + \Delta.$$

(To see why $\sum \operatorname{Var} \leq \mathbb{E}$, note that taking $i = j$ gives that $\operatorname{Cov}[X_i, X_j] \leq \mathbb{P}(A_i)$ by the above thing.) We're losing a lot of terms here, but it turns out none of those terms are important. So we can rewrite this as a conditional probability

$$\Delta \leq \sum_i \mathbb{P}(A_i) \sum_{j \sim i} \mathbb{P}(A_j \mid A_i).$$

Define

$$\Delta^* = \max_i \sum_{j \sim i} \mathbb{P}(A_j \mid A_i).$$

In practice, when we have symmetry in the problem (which will be the case for us), this quantity does not actually depend on i, and is usually pretty clean to compute.

We will see an example with the calculation soon. First putting everything here together, we see that we can replace the second sum by Δ^*, and the first sum is just $\mathbb{E}X$ — so

$$\Delta \leq (\mathbb{E}X)\Delta^*.$$

Now recall that having $\mathbb{E}X > 0$ and $\operatorname{Var} X = o((\mathbb{E}X)^2)$ implies that $X > 0$ with high probability. In the above setup, we see that if $\mathbb{E}X \to \infty$ and $\Delta^* = o(\mathbb{E}X)$, then $X > 0$ with high probability. And that's what we will verify in our problem of determining the threshold for containing a fixed subgraph — we will check that $\mathbb{E}X \to \infty$ and $\Delta^* = o(\mathbb{E}X)$. This is kind of a proxy for $\operatorname{Var}/\mathbb{E}$.

This is all just for setup; we will see an application soon.

Theorem 8.13

The threshold for containing K_4 is $n^{-2/3}$.

This means when $p \ll n^{-2/3}$ then $G(n,p)$ doesn't contain K_4 with high probability, and when $p \gg n^{-2/3}$ then $G(n,p)$ does contain K_4 with high probability.

Proof. The 0 statement is usually easier. Let X be the number of copies of K_4 in $G_{n,p}$. Then

$$\mathbb{E}X = \binom{n}{4} p^6 \asymp n^4 p^6.$$

So if $p \ll n^{-2/3}$, then $\mathbb{E}X = o(1)$, and therefore $\mathbb{P}(X \geq 1) \leq \mathbb{E}X = o(1)$ by Markov; so we have proved the 0-statement, that the grpah is K_4-free with high probability.

The other direction is where we need the second moment calculation — now suppose $p \gg n^{-2/3}$. We see already that $\mathbb{E}X \to \infty$. And for each 4-vertex subset of the vertices, let A_S be the event that S is a clique (meaning K_4) in $G(n,p)$.

We have $\binom{n}{4}$ such events, and we want to understand their interactions.

For each fixed S, one has $S' \sim S$ if and only if they share edges. Now we need to analyze in what ways they can share edges — $S' \sim S$ iff $|S \cap S'| \geq 2$, meaning they share edges.

There are two possibilities — they could share exactly 2 vertices, or they could share 3 (we are assuming $S \neq S'$). The number of S''s that share exactly two vertices with S is $O(n^2)$ (we will ignore constant factors — we need to select each of the two remaining vertices, which can be done in around n ways). For each such S', we need to evaluate $\mathbb{P}(A_{S'} \mid A_s)$. This is saying that if the blue S already appeared in the random graph, what's the probability the red 4-clique appears? We need to add 5 extra edges, so this is p^5. Similarly, in the second case, the number of S' is $O(n)$, and for each such S' the conditional probability $\mathbb{P}(A_{S'} \mid A_S) = p^3$ since that's the number of additional red edges that we need to add.

And now to compute Δ^*, we just have to sum over everything — so we get

$$\Delta^* = \sum_{|S' \cap S| = 2 \text{ or } 3} \mathbb{P}(A_{S'} \mid A_s) = O(n^2 p^5 + np^3).$$

We can check (we won't do this here) that this quantity is $\ll n^4 p^6 \asymp \mathbb{E}[X]$, provided that $np \to \infty$. So we have checked what we needed to check — that $\Delta^* \ll \mathbb{E}[X]$ and therefore the number of K_4's is positive with high probability. □

Note that $2/3$ comes from the expectation calculation — in both K_3 and K_4, the threshold came from the computation for expectation — whether $\mathbb{E}X \to 0$ or ∞. When it goes to 0, you really don't expect this event to occur. But when it goes to ∞ it's not a priori clear, but we did the second moment method to show that it is.

Question 8.14. Does this always happen?

Example 8.15

Let H be K_4 with an extra edge hanging off it.

Let X_H be the number of copies of H. Then $\mathbb{E}X_H \sim n^5 p^7$. So if $\mathbb{E}X = o(1)$ then we know that $X = 0$ with high probability (by Markov). But what about in the other direction — what if $\mathbb{E}X \to \infty$? Can we conclude that we typically have at least one copy of H?

You can check that $\mathbb{E}X \to \infty$ gives the "prediction" $p \asymp n^{-5/7}$. But now consider when $n^{-5/7} \ll p \ll n^{-2/3}$. Since $n^{-2/3}$ is what we just found to be the threshold for containing a K_4, we know that below this threshold $G(n,p)$ is K_4-free with high probability. And if it's K_4-free, it can certainly not contain our new object H. So it's also H-free.

That's the surprise mentioned in the beginning — you can't just compute the expectation, because it might not be the dominating control.

In this case, what we really need to look at is the *densest part* of the graph.

It turns out that the threshold for K_4 is *also* the threshold for K_4 plus an edge. This should not be so surprising — once you contain a K_4, it's "free" to add an extra edge.

We have a complete answer. First let's define something to capture the notion fo the "densest subgraph."

Definition 8.16. For a graph, define the **edge-vertex ratio** as $\rho(H) = e(H)/v(H)$, and define $m(H)$ as the maximum possible edge-vertex ratio among all subgraphs —

$$m(H) = \max_{H' \subseteq H} \rho(H').$$

You should only delete vertices when taking the subgraph — you shouldn't delete edges because that would be strictly worse. So if we have a clique we should keep it, but if we have a sparse graph we should look at the dense parts.

Theorem 8.17 (Bollobás 1981)

For a fixed graph H, the threshold for $G(n, p)$ containing H as a subgraph is $p^{-1/m(H)}$.

This is the same story as we saw for our K_4 with an edge.

The proof will be left to next time. But the moral of the story here is that if you do the second moment calculation, when you do this calculation you find that the dominant term is what comes out of this definition. Just as an example, in the case of H being this thing, the densest subgraph is K_4, which has edge-vertex ratio $6/4 = 3/2$.

Student Question. Does this mean if there's 1 triangle, there's 1000 triangles?

Previously we've shown that $np \to \infty$ means $X_{K_3} > 0$ whp. But this same proof also shows you that $X \sim EX$ with high probabiltiy (it's concentrated around its mean — i.e. for any ϵ, $(1-\varepsilon)\mathbb{E}X \le X \le (1+\varepsilon)\mathbb{E}X$ with high probability. So the number of triangles must actually be quite close to teh mean. In fact, there's better results by using better concentration bounds. In the homework we will demonstrate the asymptotic normality — we will show $(X - \mathbb{E}X)/\sqrt{\operatorname{Var} X} \to N(0,1)$, although this doesn't follow from the standard central limit theorem (since that's about sums of independent random variables). For this problem, the historical method of how this was shown was the *method of moments* — you check for every k that the kth moment of this quantity converges to the kth moemnt of this standard Gaussian. We'll see another method in the homework, the method of projectiosn — we will show X is close to Y which is the sum of independent random variables, and since Y satisfies the CLT so does X. And we will prove this by checking $X - Y$ has small variance. (This will be somewhat calculation-heavy, but this is very important and a powerful tool.)

§9 September 28, 2022 — Thresholds

Last time, we started talking about the second moment method, adn in particular, we showed the following:

$$\mathbb{P}(G(n,p) \text{ contains a triangle}) \to \begin{cases} 0 & np \to 0 \\ 1 & np \to \infty \end{cases}.$$

The 0-statement comes from looking at the *expected* number of trianlges, and the 1-statement comes from the second moment method.

This is an example of a threshold. We haven't yet formally defined thresholds, but we can see that there is a threshold at $p = 1/n$ — we have one phenomenon when $p \ll 1/n$ and a completely different one when $p \gg 1/n$, so there's a phase transition at $1/n$.

Then we looked at what happens when we replace a traingel wtih some other graph. Basically the same thing happens, but with a twist — we have to look at the *densest subgraph*. SO the threshold for $G(n, p)$ to contain some fixed graph H is at $n^{-1/m(H)}$, where $m(H)$ is the edge-vertex ratio of the densest subgraph of H.

We will not prove this in class, since there aren't really any new ideas. Iinstead, we will move on to discussing thresholds more generally.

Thresholds are not just about random graphs — we can define this notion more abstractly, adn consider some questions.

0. Is there always a threshold? — The answer is yes — there's a general result that says there's always a threshold for monotone properties.
1. Where is the threshold?
2. What is the nature of the phase transition? — How quickly do we transition from one behavior to the other (do we transition very quickly, or slowly and gradually?)

We will initially study this quite abstractly, but these are usually studied for very concrete properties — given a random graph, does it contain a Hamilton cycle? A perfect matching? Some of these questions have been solved and some are still open research problems; and understanding thresholds is at the center of probabilistic combinatorics.

§9.1 The General Setup

Suppose we have a ground set Ω, which is some finite set. We'll write Ω_p to be a subset of Ω, where every element is included with probability p independently.

Example 9.1

The random graph model $G(n, p)$ corresponds to taking Ω to be the set of all 2-element subsets — $\Omega = \binom{[n]}{2}$.

We would like to know if some subset of Ω satisfies a certain property.

Definition 9.2. An **increasing property** (also known as a *monotone property*) on subsets of Ω is some binary property such that if $A \subseteq \Omega$ satisfies the property, then all supersets of A also satisfy the property.

So adding elements cannot destroy having this property.

Definition 9.3. A property is **trivial** if either all sets satisfy it, or all sets do *not* satisfy it.

A property is trivial if it is always true or always false. These are uninteresting; we will always consider nontrivial properties.

We will be particularly interested in *graph* properties — innitially properties of vertex labelled graphs, but these properties should only depend on the isomorphism class (it should not odepend on what labels you assign to the vertices). For example:

- Containing a triangle

- Being connected
- Having a perfect matching
- Having a Hamilton cycle
- Not being 3-colorable

In each of these properties, adding more edges can never destroy the property, so all of these are increasing properties.

There is an equivalent way to view the concept of increasing properties: consider a family $\mathcal{F}$ of subsets of Ω (so $\mathcal{F} \subseteq \mathcal{P}(\omega)$). Then we say $\mathcal{F}$ is an *up-set* if whenever $A \in \mathcal{F}$, all supersets of A are also in $\mathcal{F}$. So being an up-set is basically teh same thing as being an increasing property; we'll use these terms rather interchangeably.

Student Question. Why do we use this model of ER random graphs in general?

It's nice that the edges are idnependent. The original paper used n vertices and m edges, which turns out to be basically the same — not exactly the same, but for all practical purposes they're similar.

Is this model good at modelling real world situatiosn? Not really. But we view it as a way to study probabilistic methods. There is a hwole world of studies called network science about ways to model actual networks, that we will not get into.

Definition 9.4 (Thresholds)**.** Let $\Omega = \Omega^{(n)}$ be a sequence of finite sets, and $\mathcal{F} = \mathcal{F}^{(n)}$ a sequence of properties (where each is a monotone property on subsets of Ω). We say that q_n is a **threshold** for $\mathcal{F}$ if the following is true:

$$\mathbb{P}(\Omega_{p_n} \in \mathcal{F}) \to \begin{cases} 0 & \text{if } p_n/q_n \to 0 \\ 1 & \text{if } p_n/q_n \to \infty \end{cases}.$$

This generalizes the thing we saw for triangles.

Student Question. Do they have to be realted?

In teh abstract notion, not necessarily — we could have containing a triangle for n even and containing K_4 for n odd.

This is a direct generalization of the triangle statement.

A few remarks: we have only defined this for increasing properties, you can similarly define it for decreasing properties by switching 0 and 1. If the property is not monotone, you may or may not have a threshold. For example, containing some given graph as an *induced* graph. If you have no edges or all edges that's not good; somewhere in the middle is good. So this is nto monotone.

The other thing is that here we say $p = 1/n$ is a threshold, but we could equally say $p = 2/n$ is a threshold. But if q_n and q_n' are both thresholds for the same property, then we must hav $ecq_n' \le q_n \le Cq_n'$. So when we say *a* threshold, we can say *the* threhsold and ignore constant factors.

Question 9.5. Does a threshold always exist?

First, why is this a nontrivial question? We could imagine that the probability of success, as p goes from 0 to 1, may hover for a very long time at some intermediate region. For example, maybe between ϵ and $1-\epsilon$, there's a very big gap. It turns out that never happens.

But before even addressing that, there's an even more basic question:

Question 9.6. Is this probability even an increasing function of p?

This seems obvious, but the proof is not trivial.

Theorem 9.7 (Monotonicity of Satisfying Probability)

For Ω finite and $\mathcal{F}$ a nontrivial monotone increasing property, then $p \mapsto \mathbb{P}(\Omega_p \in \mathcal{F})$ is strictly increasing on $p \in [0, 1]$.

This should be intuitively obvious — your property only gets better as you add more edges. But if you have two different probabilities p and q, and you draw one from p and one from q, it's not a priori clear why one should be greater than the other. So intuitively this is clear but it requires a proof.

We will see a couple of related proofs. They are both neat and show important lessons. Both proofs use the idea of *coupling*, which is an important technique in probability that lalows you to analyze two random processes by couplling them together into a joint distribution.

Proof 1. Let $0 \le p < q \le 1$, and for each $x \in \Omega$, first generate a uniform $t_x \in [0, 1]$ indepdnently for all x. Then consider the set $A = \{x \in \Omega \mid t_x \le p\}$, and $B = \{x \in \Omega \mid t_x \le q\}$. So $A, B \subseteq \Omega$.

Now A has the same distribution as Ω_p — when we generate it, that's the same as picking each element with probability p. Similarly, B has the same distribution as Ω_p. But now these processes are coupled — we're not picking them independently. ANd from the definition, when an element is in A, it's automatically in B.

SO we know $\mathbb{P}(A \in \mathcal{F}) \le \mathbb{P}(B \in \mathcal{F})$, because $A \in \mathcal{F}$ implies $B \in \mathcal{F}$ by monotonicity. We also know that A has the same distribution as Ω_p, so this proves

$$\mathbb{P}(\Omega_p \in \mathcal{F}) \le \mathbb{P}(\Omega_q \in \mathcal{F}).$$

To check this is strict, we will show that there's a nonzero probability that A is $\varnothing$ and B is Ω. This happens precisely when all the labels are $t_x \in (p, q]$, and this occurs with positive probability. □

This is a different way to think about probabilities — if you sampled the two things independently, it woudl be hard to compare them.

There's another proof, which looks different but has the same idea.

Proof (2-round exposure). As earlier, set $0 \le p < q \le 1$. Note that $B = \Omega_q$ has the same distribution as the union of independent $A = \Omega_p$ and $A' = \Omega_{p'}$, where p' is chosen carefully to satisfy the equation

$$1 - q = (1 - p)(1 - p').$$

In a concrete example, flipping a coin two times and taking the OR is the same as flipping a single 75% coin. So to generate a set with element probability q, this is the same as the union of two independent coin flips with some approopriately chosen probability.

In other words, we're using two rounds of exposure to generate our thing.

As a result, $\mathbb{P}(A \in \mathcal{F}) \le \mathbb{P}(A \cup A' \in \mathcal{F}) = \mathbb{P}(B \in \mathcal{F})$, since $A \cup A'$ has the same distribution as $\mathcal{F}$. That gives us weak monotonicity. To get the strict inequality, we observe that with positive probability we have $A \notin \mathcal{F}$ and $A \cup A' \in \mathcal{F}$ (similarly to before). □

So this was again through a coupling argument, where we coupled generating a q-random subset by showing it's the asme as the union fo two independent random subsets with appropriately chosen probabilitites.

SO far, we know that the satisfying probability is an increasing function of p, which is intuitively obvious but has a neat proof.

Theorem 9.8 (Existence of Threshold, Bollobás–Thomason 1987)

Every nontrivial monotone property has a threshold.

This is really about the sequence, rather than specific set — we're looking at a sequence of set and sequence of properties.

The key claim is the following non-asymptotic claim:

Lemma 9.9

IF Ω is a finite set and $\mathcal{F}$ a nontrivial monotone property, and if $p \in [0,1]$ and m a nonnegative integer, then

$$\mathbb{P}(\Omega_p \notin \mathcal{F}) \le \mathbb{P}(\Omega_{p/m} \notin \mathcal{F})^m.$$

First let's prove this lemma. The proof is basically a modification of the coupling proof, although it's a bit confusing because of the negations.

Proof. Consider m independent copies of $\Omega_{p/m}$, and let Y be their union. Since $\mathcal{F}$ is increasing, if $Y \notin \mathcal{F}$, then none of these m copies of $\Omega_{p/m}$ are in $\mathcal{F}$ — if any copy forming Y satisfied the property, then Y woudl automatically satisfy the property. This gives

$$\mathbb{P}(Y \notin \mathcal{F}) \le \mathbb{P}(\Omega_{p/m} \notin \mathcal{F})^m$$

(because that's the probability that *none* of the constituents satisfy the property, by independence).

But as before, Y is the union fo m independent copies of things chosen with probability p/m. So then Y has the same distribution as Ω_q where $q \le p$ (you can calculate exactly what it is, but it's at most p), by finding the probability each elemnent appears. So by monotonicity, we have

$$\mathbb{P}(\Omega_p \notin \mathcal{F}) \le \mathbb{P}(\Omega_q \notin \mathcal{F}) = \mathbb{P}(Y \notin \mathcal{F}) \le \mathbb{P}(\Omega_{p/m} \notin \mathcal{F})^m.$$

□

Now we are ready to prove the existence of thresholds.

Proof of THeorem. We saw earlier that the function $p \mapsto \mathbb{P}(\Omega_p \in \mathcal{F})$ is increasing. We can also see that it starts from 0 and ends at 1, since the property is nontrivial — the empty set doesn't satisfy the property, and the full set does. Furthermore, it is continuous. (One way to see this is that if you're told what Ω and $\mathcal{F}$ are, you can write down an exact formula — it's actually polynomial in p.) Therefore, by the IVT, there exists a **critical probability** such that $\mathbb{P}(\Omega_{p_c} \in \mathcal{F}) = 1/2$.

It remains to check for every ε, there exists sufficiently large $m = m(\varepsilon)$ (which does not depend on the set or property) such that

$$\mathbb{P}(\Omega_{p_c/m} \notin \mathcal{F}) \ge 1 - \varepsilon,$$

and

$$\mathbb{P}(\Omega_{p_c m} \notin \mathcal{F}) \le \varepsilon.$$

This is a nonasymptotic statement, but n ow if $p/q \to 0$ that corresponds to m getting larger and large,r which allows you to take ε to be smaller and smaller. So having this statement here would imply the existence of thresholds — in particular p_c is a threshold.

But our lemma directly implies that (you can move symbols around a bit — this allows you to scale the probability by dividing and multiplyign by m). □

So for monotone properties, thresholds always exist.

§9.2 History of Thresholds

Now we will go on a tour through some of the things we know about thresholds of graph properties. There's a lot of interesting and deep mysteries behind thresholds; we'll see that in the second half of this lecture.

Example 9.10

The property of containing a triangle.

We saw that the threshold is $1/n$. Specifically,

$$\mathbb{P}(G(n,p)\text{contains a triangle}) \to \begin{cases} 0 & \text{if } np \to 0 \\ 1 & \text{if } np \to \infty \end{cases}.$$

It's legitimate to ask what happens if $np \to c$ for some $c > 0$? We analyze this in teh homework; the limiting distribuiton is Poisson, which means you can deduce the probability of having a triangle. The answer is $1 - e^{\text{something}}$ — tis' some number that goes from 0 to 1 as c goes from 0 to ∞.

What does the transition look like? As p increases on the scale of $1/n$, does the probability increase slowly or quickly from 0 to 1? This seems to be a quite gradual icnrase.

This is an example of a **coarse** threshold — a slow transition from 0 to 1.

A similar thing happens for containing H (for some other graph H) — there's a coarse transition on the order of the threshold, where the threshold is $p \asymp n^{-1/m(H)}$.

Example 9.11

The property of having no isolated vertices.

Here this is not hard to find using the second moment method: suppose $p = (\log n + c)/n$. Then

$$\mathbb{P}(G(n,p) \text{ has no isolated vertices}) \to \begin{cases} 0 & c \to \infty \\ 1 & c \to 0 \\ 1 - e^{-e^{-c}} & c \text{constant} \end{cases}.$$

Now first, the threshold is at $\log n/n$. But the nature of the threshold is quite different. If we plot p at around $\log n/n$, we see that rather than a slow transition from 0 to 1, there is a very rapid transition. In fact, the window of this transition is $\Theta(1/\sqrt{n})$.

This is an example of a **sharp** threshold. In particular, as we go from $0.9 \log n/n$ to $1.1 \log n/n$, the property chagnes very quickly from almost n ever satisfying to almost always satisfying — there's a rapid change in behavior. The physical analogy is when water goes from 1 degree above freezing to one degree above freezing, there's a dramatic changei n the physical nature.

Example 9.12

The threshold for connectedness is the same as for an isolated vertex in an exact way — we have the same formula

$$\mathbb{P}(G(n,p)\text{connected}) \to \begin{cases} 0 & c \to \infty \\ 1 & c \to 0 \\ 1 - e^{-e^{-c}} & c \text{constant} \end{cases}$$

There is a precise way these are linked. The first is quite easy to show using the second moment. For connected, it's not too hard — it already appears in the original Erdős–Renyi paper. But more is true.

Imagine geneating the random graph by adding new random edges, one at a time. SO we start with teh empty graph, adn we consider a process where each time we throw down an edge that hasn't already appeared, uniformly at random. At som epoint, it will not have any isolated vertices.

Theorem 9.13

With probability $1 - o(1)$ (as $n \to \infty$), as soon as there are no isolated vertices, the graph becomes connected.

So if you keep on adding new edges and you stop as soon as there are no isolated vertices, and you freeze the graph at that point, then with probability approaching 1 this grpah will be connected.

This is a precise way in which these two properties are linked to each other.

Likewise, if we replace connected by "having a perfect matching" (where we restrict ourself to n even), everything is exactly teh same. The reason to mention this is the next example:

Example 9.14

What happens for a random hypergraph? Take $G^{(3)}(n, p)$ — n vertices, each *triple* appears with probability p. When does the graph have a perfect matching? Here a matching is actually a *triple* matching — collectino of disjoint triples that cover all vertices.

This turns out to be a really difficult problem. It's not too hard to analyze what happens for when there's no isolated vertices — a similar calculation tells you that has a sharp threshold at $2\log n/n^2$. But then knowing whether it has a perfect matching was a major open problem. It was resolved only in 2008, in a paper by Johanson–Kahn–Vu, who showed that if $p > C\log n/n^2$ then $\mathbb{P}(\text{pm}) \to 1$. This demonstrates a threshold at $C\log n/n^2$. This is a really hard paper — even experts in teh area do not understand the paper.

More recently, there were a couple of developments. In particular, there was a recent breakthrough on the Kahn–Kalai conjecture. That gives a new, much shorter, proof of this theorem.

In 2022 Kahn proved a *hitting time* result (as soon as you get no isolated vertices, you get a perfect matching — similarly to the normal grpah case).

§9.3 Coarse and Sharp Thresholds

Definition 9.15. r_n is a *sharp* threshold if for every $\delta > 0$,

$$\mathbb{P}(\Omega_{p_n} \in \mathcal{F}) \to \begin{cases} 0 & p_n/r_n \le 1-\delta \\ 1 & p_n/r_n \ge 1+\delta \end{cases}.$$

On the other hand:

Definition 9.16. If there exists some $\varepsilon > 0$ and constants c and C so that

$$\mathbb{P}(\Omega_{p_n} \in \mathcal{F}) \in [\varepsilon, 1-\varepsilon]$$

whenever $c \le p_n/r_n \le C$, then we say r_n is **coarse**.

There are finer questions — if it's a sharp threshold, *how* sharp (what is the transition window), what is the shape? There are som einteresting transitions that are sharp from one side and coarse on teh other —- quickly increases to something and then slowly increases to 1. (That would be coarse by this definition, but there are finer questions to ask.)

But determining sharp vs. coarse is already quite difficult. Neverhtelss, there is a very important and strong theorem that givves us a complete characterization of hwen *graph* properties have sharp or coarse thresholds:

We won't state this precisely. On one hand, containing a fixed subgraph is coarse, while being connected, having a perfect matching, having no isolated vertices, are sharp. A qualitative difference between these properties is that having a subgraph is *local* — it depends only on a bounded number of vertices — and the rest are global.

Very roughly speaking:

Theorem 9.17 (Friedgut's Sharp Threshold Theorem, 1999)
All montone properrties with coarse thresholds are close to some local property.

This is a very rough, informal statement of hte real theorme. It has to be stated this roughly because you do need to take into consideration some weird things that can happen — this is about very general graph properties, and you can throw in a lto of irrelevant garbage into the definition of the property (i.e. contains a triangle and at least $\log n$ edges — this is not local but getting $\log n$ edges is almost trivial, so it shouldn't materially affect things). For example, containign one of severl afixed subgraphs would also be coarse; if n is odd then triangle, fi n is even then K_4, would also be coarse.

We saw that containing a fixed H gives you a threshold of the form $n^{-1/m(H)}$ (where this exponent is rational). For several subgraphs it's similar. So Friedgut's theorem has a corollary:

Corollary 9.18
If $r(n)$ is a coarse threshold of some graph property, then there exists a partition of $\mathbb{N}$ into a finite number of sets $\mathbb{N}_1, \dots, \mathbb{N}_k$, and rational s$\alpha_1, \dots, \alpha_k$, so that $r(n) \asymp n^{-\alpha_j}$ for all $n \in \mathbb{N}_j$.

In particular:

Corollary 9.19
If $\log n/n$ is a threshold, then it is automatically sharp.

You cannot have a coarse threshold which has the form $\log n/n$ — that is impossible because of Friedgut's theroem. Likewise, if $n^{-\alpha}$ for irrational α is a threshold, then it's also sharp (although Prof. Zhao does not know any natural properties with such a threshold). That is a surprising result of a quite abstract thing — now from knowing the answer of a threshold $\log n/n$, we know the thing *has* to be sharp.

(The reason for the partition is silly thing slike "if n is even triangle, if n is odd K_4").

There are sometimes 100-page long research papers that show a graph property is sharp, and the entire point is to check that it satisfies the hypothesis of the theorem. So this can be hard.

To conclude, we will see a specific problem for which we still don't know the complete answer.

Question 9.20. W hat is the threshold for 3-colorability (or k-colorability in general)?

We know the threhsold is at $1/n$. But we have a conjecture:

Conjecture 9.21 — for every fixed $k \geq 3$, there exists some constant d_k such that for any fixed $d > 0$,

$$\mathbb{P}(G(n, d/n)\text{is } k\text{-colorable}) \to \begin{cases} 1 & d < d_k \\ 0 & d > d_k \end{cases}.$$

So this says there is a specific sharp threshold of the form d_k/n — slightly larger you are overwhelmingly unlikely to be k-colorable, slightly smaler overwhemlingly likely.

Theorem 9.22 (Achioptas–Friedgut 2000)
FOr every $k \geq 0$, there exists a function $d_k(n)$ such that for every $\epsilon > 0$ and sequence $d(n)$,

$$\mathbb{P}(G(n, d(n)/n)\text{is } k\text{-colorable}) \to \begin{cases} 1 & d(n) < d_k(n) - \varepsilon \\ 0 & d(n) > d_k(n) + \varepsilon \end{cases}.$$

The theorem says there exists some function of n — and this function is bounded between two constants — such that you have a sharp threshold. So this si a sharp threshold statement — we have a sharp threshold around $d(n)/n$. This doesn't prove the conjecture because the theorem allows for the possiblity that $d(n)$ fluctuates between two constants. So this shows you that a sharp threshold exists, but does not pinpoint the location fo the sharp threshold. THis is quite curious, and shows you the nature of that theorem.

At the beginning, we suggested to find the location, and then determine the nature. You'd think you have to understand the location before nature — that's a natural thing to expect. But it's not the case — powerful theorems let us understnad the nature of the transition without knowing where the transition si tkaing place. As if you knew water to ice occurs very quickly, but you don't know hwere the freezing point is.

So we do have some upper and lower bounds on d_k, but we odn't know that it ocnverges.

§10 October 3, 2022

Today we will continue discussing the second moment method.

§10.1 Clique Number

Definition 10.1. The **clique number** $\omega(G)$ is then umber of vertices in the largest clique of G.

Remark 10.2. One way to remember the notation ω is that the independence number is usually denoted α, and α and ω are the first and last letters in the Greek alphabet.

Question 10.3. What is the clique number of $G(n, 1/2)$?

In other words, $\omega(G(n, 1/2))$ is a random variable, which we'd like to understand. This is somewhat similar to how we computed teh threhsold for triangles and K_4's.

Let X be the *number* of k-cliques in $G \sim G(n, 1/2)$, and let

$$f(n, k) = \mathbb{E}X = \binom{n}{k} \cdot 2^{-\binom{k}{2}}.$$

(There's $\binom{n}{k}$ possibilities for the vertices, and each forms a clique with probability $2^{-\binom{k}{2}}$.)

An often useful inequality is

$$\left(\frac{n}{ek}\right)^k \le \binom{n}{k} \le \left(\frac{en}{k}\right)^k.$$

Plugging this in, we get

$$\log_2 f(n,k) = k\left(\log_2 n - \log_2 k - \frac{k}{2} + O(1)\right).$$

We wish to know, for what value of k do we get a transition point (whether this goes to 0 or ∞)? We can check that the transition is at $k \sim 2\log_2 n$ — the two terms $\log_2 n$ and $k/2$ should cancel each other out, as all the other terms are second-order. By "transition," we mean that if $k(n) \ge (2+\delta)\log_2 n$, then $f(n,k) \to 0$, and conversely if $k(n) \le (2-\delta)\log_2 n$, then $f(n,k) \to \infty$.

So then we expect the cliquen number to be $2\log_2 n$. We will establish the following:

Theorem 10.4

Let $k = k(n)$ be a sequence of positive integers.

1. If $f(n,k) \to 0$, then $\omega(G(n,1/2)) < k$ with high probability.
2. If $f(n,k) \to \infty$, then $\omega(G(n,1/2)) \ge k$ with high probability.

This should mirror what we discussed about the presence of a triangle or K_4. The difference is there p changes with n; here p is fixed, and the size of the clique is what changes.

Proof. The first part is an application of Markov's inequality —

$$\mathbb{P}(X \ge 1) \le \mathbb{E}[X] \to 0.$$

For the second, we use the second moment method. Two lectures ago, we had the following setup: for every k-vertex subset S, let A_S be the event that S forms a clique, and recall that we had the quantity

$$\Delta^* = \max_i \sum_{j \sim i} \mathbb{P}(A_j \mid A_i).$$

(Here i and j range over k-vertex subsets, and $j \sim i$ means that j and i are not independent.) In this case, the quantity is symmetric, so we can fix a clique — then we want to know how many cliques intersect with our existing k-clique in at least one edge. So then we are looking at

$$\sum_{T \in \binom{[n]}{k}, 2 \le |S \cap T| \le k-1} \mathbb{P}(A_T \mid A_S)$$

(where S is fixed). So we know S is already a clique, and we are trying to get T to be a clique — and we want to count the additional edges that need to be put in.

We can write out this sum as

$$\sum_{i=2}^{n} \binom{k}{i}$$

since there are $\binom{k}{i}$ ways to pick the intersection, $\binom{n-k}{k-i}$ ways to pick the vertices outside the intersection, and the probability is then $2^{\binom{k}{2}-\binom{i}{2}}$. Some calculation gives that this quantity is $\ll \mathbb{E}X = \binom{n}{k}2^{-\binom{k}{2}}$. So from the setup from two lectures ago, we know that if $\Delta^* = o(\mathbb{E}X)$, then Chebyshev applies asymptotically, and we can conclude that $X > 0$ with high probability. This means there is at least one clique of size at least k, so $\omega(G) \ge k$. □

So very similarly to the appearance of triangles or K_4's, this was completely governed by the first moment.

Question 10.5. What exaclty is k?

If we are more careful with this analysis, we will see that there is a surprising result:

Theorem 10.6 (Two-point concentration of clique number)
There exists some $k = k(n) \sim 2\log_2 n$ such that $\omega(G(n, 1/2)) \in \{k, k+1\}$ with high probability.

In fact, for almost all integers there's 1-point concentration. So our random variable can take different values, but almost always it's concentrated at one piont. This i quite different from anything we've seen before — the *number* of triangles has fluctuation, but here there is *very* sharp concentration (with probability to 1 as $n \to \infty$, it's one of two possible values). So this is a very strong form of a concentration statement.

Proof. The proof is not too difficult. We can check that for $k \sim 2\log_2 n$, we have

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

For $k \sim 2\log_2 n$, this is $n^{-1+o(1)}$.

So around the critical point, the expected number of k-cliques drops quite rapidly — by a factor of n around each k. SO if this is our threshold, it can't fluctuate around not going to 0 or ∞ for long — let $k_0 = k_0(n) \sim 2\log_2 n$ be the value such that $f(n, k_0) \geq n^{-1/2} > f(n, k_0 + 1)$. Then we can see that $f(n, k_0 + 1) \to 0$ and $f(n, k_0 - 1) \to \infty$. By the earlier theorem, we conclude that $\omega(G(n, 1/2))$ is either $k_0 - 1$ or k_0 with high probability. □

So the main point is that you basically only have one chance to stay where you want (not at 0 or ∞) because the thing changes rapidly (it drops like $1/n$).

Even though this result is stated as two-point concentration — and indeed there will be some integers at which we have two-point concentration (both probabilities of k and $k+1$ are bounded away from 0), along "most" n we will actually have $f(n, k_0) \to \infty$ as well, in which case we have one-point concentration.

Note that this also implies $\alpha(G(n, 1/2)) \in \{k_0, k_0 - 1\}$ with high probability — the independence number is the clique number of the complement, and taking the complement of the graph keeps the distribution of $G(n, 1/2)$.

Question 10.7. What about $\chi(G)$?

We know that

$$\chi(G) \geq \frac{n}{\alpha(G)},$$

since each color class has at most $\alpha(G)$ vertices, and we have χ color classes. So this allows us to get some lower bound on χ — we know

$$\chi(G) \geq \frac{(1 - o(1))n}{2\log_2 n}$$

with high probability. It turns out that this result is actually tight, and we'll see this later in the course.

Remark 10.8. At the RSA concentration in Poland, there was a talk that improved this to a wider range — $\alpha(G(n,p))$ is also 2-point concentrated for all $p \geq n^{-2/3+\varepsilon}$.

§10.2 Distinct Prime Factors

The next application is not about random graphs, but about number theory.

Notation 10.9. Let $\nu(n)$ be the number of distinct prime factors of n (not counting multiplicity).

THis is a pretty fundamental quantity in number theory. The next theorem tells us that almost all n have roughly $\log\log n$ prime divisors.

Theorem 10.10 (Hardy–Ramanujan 1917)
For every ε, there exists a constant C such that for all sufficiently large n, all but a fraction ε of $\{1, 2, \ldots, n\}$ satisfy

$$|\nu(n) - \log\log n| \le C\sqrt{\log\log n}.$$

SO for large n, almost all positive integers up to n have roughly $\log\log n$ distinct prime divisors. This is not a result about a random objec t— it is about primes, which are not random. Neverhtelss, the secon moment method will be useful. (This is not how Hardy and Ramanjuan initially proved it — they used much more advanced number-theoretic techniques. The proof we will see is by Turan from 1930s.)

Proof. Choose x uniformly at random from $[n]$. FOr every prime p, let X_p be the random variable which indicates whether $p \mid X$ — it's 1 if $p \mid x$ and 0 if $p \nmid x$. Let $M = n^{1/10}$ (a slowly increasing function of n), and let

$$X = \sum_{p \le M} X_p.$$

(The sum is over primes; this is standard in analytic number theory.) Note that X is not exactly the number of prime divisors — because X might have bigger prime divisors. But it cannot have very many — it can have at most 10 prime divisors bigger than this number. So ν is very close to X — $\nu(x) - 10 \le X \le \nu(x)$.

Now we have converted the problem into one that looks a bit more familiar to us — trying to estimate the sum of a bunch of random variables.

Question 10.11. Are these random variables basically independent?

Not exactly, but almost. First, let's compute $\mathbb{E}X$ — we have

$$\mathbb{E}X_p = \frac{\text{multiples of } p}{n} = \frac{\lfloor n/p \rfloor}{n} = \frac{1}{p} + O\left(\frac{1}{n}\right).$$

This means

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

Theorem 10.12 (Mertens' Theorem)
We have $\sum_{p \le n} 1/p = \log\log n + O(1)$, where $O(1)$ converges to a specific constant.

Plugging in this estimate (taking two logs of M gives us basically the sam ething as n), we get

$$\mathbb{E}X = \log\log x + O(1).$$

Now we want to find $\operatorname{Var} X$. For that, we need to compute the pairwise interactions of these random variable.s

If $p \neq q$, then $X_p X_q = 1$ iff $pq \mid x$. The piont is that these two events are basically independent — they are independent if $pq \mid n$ (by the Chinese Remainder Theorem), while otherwise you're off by a little bit (a smaller order termK).

Then

$$|\mathrm{Cov}(X_p, X_q)| = |\mathbb{E}[X_p X_q] - \mathbb{E}X_p \mathbb{E}X_q| .$$

We can write these down explicitly as

$$\frac{\lfloor n/pq \rfloor}{n} - \frac{\lfloor n/p \rfloor}{n} \frac{\lfloor n/q \rfloor}{n}.$$

This becomes

$$\frac{1}{pq} + O(1/n) - \left(\frac{1}{p} + O(1/n)\right)\left(\frac{1}{q} + O(1/n)\right) = O(1/n).$$

Then we have

$$\sum |\mathrm{Cov}[X_p, X_q]| \lesssim 1/n \cdot M^2 \lesssim n^{-4/5}.$$

So all the covariance contributions are quite small.

So in contrast to our earlier calculations (for the number of triangles), there are a lot of dependencies — every pair of summands could be dependent — but the individual covariances are all quite small, so they don't in aggregate contribute much. Meanwhile earlier we had very few dependencies but they contribute more.

We can also compute

$$\mathrm{Var}\, X_p = \mathbb{E}X_p - (\mathbb{E}X_p)^2 = \frac{1}{p}\left(1 - \frac{1}{p}\right) + O(1/n).$$

Combining these, we get

$$\mathrm{Var}\, X = \sum_{p \le M} \mathrm{Var}\, X_p + \sum \mathrm{Cov}[X_p, X_q] = \sum_{p<M} \frac{1}{p} + O(1) = \log\log n + O(1).$$

This is pretty much the same as the calculation for $\mathbb{E}$.

And now we know that Var is basically teh same as $\mathbb{E}$ — we don't have muc hvariance at all. In particular, by Chebyshev's inequality,

$$\mathbb{P}(|X - \mathbb{E}X| \ge \lambda\sqrt{\log\log n}) \le \frac{(\mathrm{Var}\, X)}{\lambda^2 \cdot \log\log n} = \frac{1}{\lambda^2} + o(1)$$

(as $n \to \infty$) by Chebyshev's inequality.

Now reclal that $|X - \nu(x)| \le 10$. That finishes. □

To recap, we're trying to prove that for a typical integer $x \le n$ (where n is large), teh number of distinct prime factors of x is very close to $\log\log n$. We analyzed this by expressing the number of distinct factors as a sum of random variables X_p; for technical reasons we only sum up to $n^{1/10}$, which is not an issue because we don't really change the number of prime factors. But now we treat this as a sum of random variables and analyze its expectation and variance. In the variance calculation, we note that the pairwise covariances are all quite small — the intuition is for two small primes, whether a random integer is divisible by each of teh two primes is basically an independent event.

We can actually go a bit further. The intuition was that we had a sum of almost independent random variables. We know that the sum of *independent* random variables satisfies the central limit theorem. So we can wonder whether this thing also satisfies a central limit theorem, and it turns out the answer is yes —

Theorem 10.13 (Erdős–Kac 1940)

If $x \sim \operatorname{Unif}[n]$, $\nu(x)$ is asymptotically normal. In other words, $\mathbb{P}_{x\in[n]}((\nu(x) - \log\log n)/\sqrt{\log\log n} \geq \lambda) \to \mathbb{P}(Z \geq \lambda)$ for $Z \sim N(0,1)$.

This is pretty cool — if you plot the integers up to n and scale appropriately, it approaches a normal distribution. You have to go pretty high up to see this though — there are some hidden constants.

The main idea is to compare moments —

Theorem 10.14 (Method of Moments)

If we have a sequence of random variables X_n such that for all $k \geq 0$, $\mathbb{E}X_n^k \to \mathbb{E}Z^k$ where $Z \sim N(0,1)$, then $X_n \to Z$ in distribution.

So to check whether a sequence of random variables is asymptotically normal, it suffices to compare kth moments — a theorem guarnatees this is okay. This isi true not just for the Gaussian, but for lots of nicely behaved distributions (such as the Poisson distribution). There are som especificic conditions you can check — most distributions you run into will. (It is not true for all — there are some fuinky probability distributions not determiend by their moments, meaning you can find two distributions that are not the same but share all moments.)

Proof of E–K. We need some modifications — now set $M = n^{1/s(n)}$ where $s(n) \to \infty$ is a slowly increasing function. (This is because when we take larger powers, we may need to sacrifice more than 2); $\log\log n$ works. Also, instead of directly computing the moments of X, we will construct a model mimicking X — let

$$Y = \sum_{p\leq m} Y_p$$

be the sum fo *actually* independent variables, which are Bernoulli with $1/p$. X_p is *like* that, but we instead construct what we really want it to be (a sum of independent Bernoullis). Then we can compute that $\mu = \mathbb{E}Y \sim \mathbb{E}X$ and $\sigma^2 = \operatorname{Var} Y \sim \operatorname{Var} X$ (by the same computation as before). So if we let $\tilde{X}$ and $\tilde{Y}$ be the normalized versions of X and Y (so $\tilde{X} = (X-\mu)/\sigma$); then we want to show $\mathbb{E}\tilde{X}^k \sim \mathbb{E}\tilde{Y}^k$. Meanwhile, a version fo the central limit theorem says that a sum fo independent Bernoullis with divergent variance gives you the central limit theorem — so in particular $\widetilde{Y} \to N(0,1)$. So all it remains to do is to verify $\mathbb{E}\widetilde{X}^k \sim \mathbb{E}\widetilde{Y}^k$.

This entails expanding the kth powers of some sums of random variables — we expand both $(X-\mu)/\sigma$ and $(Y-\mu)/\sigma$, and compare term-by-term. A similar calculation holds (where we compare k-wise products instead); the comparisons check out, and this remains true. So although X is not the sum of iid's, for any fixed k it looks like one asymptotically. □

Remark 10.15. This proof is much simpler than Erdos and Kac's, who used analytic number theory methods.

§10.3 Distinct Sums

Question 10.16. What is the largest subset of $[n]$ all of whose subsets have distinct sums?

We'd like to choose as many elements as we can so that all our subsets have distinct sums. Equivalently, we can invert the relation between n and size:

Question 10.17. For each k, find the minimum n such that there exists a k-element subset $S \subseteq [n]$ such that all 2^k subset sums of S are distinct.

One construction is the powers of 2 — 1, 2, 2^2, 2^3, …, 2^{k-1}.

It's possible to do somewhat better: we can brute-force some specific better example in front, and then take this construction from there.

On the other hand, we can also get an easy bound by Pigeonhole: all 2^k sums are distinct, and at most nk. So we must have $2^k \le kn$, which means $n \ge 2^k/k$.

So now we see that $n \ge 2^k/k$, and there is an example showing 2^k is enough.

Conjecture 10.18 (Erdős) — $n \gtrsim 2^k$.

(Actually, Erdős offered 300 dollars for a proof or disproof. Erdos offered a lot of momnetary sums for problems he especially liked. The quanitty of money was indicative of how much he cared or how hard he thought it was. Some of the problems have been solved; this one is still open. Erdős signed a bunch of blank checks and gave them to Ron Graham, and whenever someone solved one of these problems they received the cash prize.)

Theorem 10.19

If $S \subset [n]$ is a k-element subset with distinct sums, then $n \gtrsim 2^k/\sqrt{k}$.

This is halfway between the easy bound and the conjecture.

Proof. The main idea is the pigeonhole argument was somewhat wasteful — even though there are 2^k sums and all of tehm are at most kn, most of the subset sums actually lie in a window of length $O(\sigma)$ — we'll see what that means soon.

Let $S = \{x_1, \ldots, x_k\}$. Consider $X = \varepsilon_1 x_1 + \cdots + \varepsilon_k x_k$, where $\varepsilon_i \in \{0, 1\}$ uniformly and independently at random.

Now we can compute some statistics — we have $\mathbb{E}X = (x_1 + \cdots + x_k)/2$, and

$$\operatorname{Var} X = \frac{x_1^2 + \cdots + x_k^2}{4} \le \frac{n^2 k}{4}.$$

By an application of Chebyshev's inequality, we can see that X is quite close to its mean — it shouldn't deviate too far away (where too far is determiend by the standard deviation)

$$\mathbb{P}(|X - \mu| < n\sqrt{k}) \ge \frac{3}{4}.$$

Now since X takes distinct values for each choice of $(\epsilon_1, \ldots, \epsilon_n)$, we have that for every specific value of X, the probability is at most 2^{-k}. So this is a probability distribution, and the individual atoms can have probability at most 2^{-k} because we don't have any collisions.

As a result, since $[\mu - n\sqrt{k}, \mu + n\sqrt{k}]$ contains at most $2n\sqrt{k} + 1$ elements, combining these two statements gives

$$\mathbb{P}(|X - \mu| < n\sqrt{k}) \le \frac{2n\sqrt{k} + 1}{2^k}.$$

We also saw previously that this is at least $\frac{3}{4}$. Combining these gives that

$$n \gtrsim \frac{2^k}{\sqrt{k}}.$$

(We ignore constant factors.) □

Compared to the pigeonhole argument (where we said there's nk possibilities), here we're still using pigeonhole, but restricted to a much narrower window — whose width is basically the standard deviation. Most — at least 75% — of the sums must lie within some interval a standard deviation away from teh mean, and this is a much smaller window on which we're applying pigeonhole.

This is still the best oudn to date; it would be quite exciting to improve this bound asymptotically. Recently there was a short and neat proof which improves the *constant* factor in this bound. The key tool used here is Harper's inequality:

View the Boolean cube $\{0,1\}^n$ as a graph. Then for a subset A, we can consider ∂A as the set of vertices not in A, but adjacent to some vertex in A — so

$$\partial A = \{x \in \{0,1\}^n \setminus A \text{ but adjacent to some element of } A\}.$$

Theorem 10.20 (Harper Vertex-isoperimetric inequality)
Every $A \subseteq \{0,1\}^k$ with exactly $|A| = 2^{k-1}$ then has $|\partial A| \geq \binom{k}{\lfloor k/2 \rfloor}$.

This is basically tight by viewing the cube vertically as a ranked lattice, and taking A to be the bottom half of the cube — then the vertex boundary is the next layer, which has that cardinality.

In fact, the general version provides a complete answer for *every* given size of A — this si an analog of the isoperimetric inequality in standard Euclidean sphere, where balls minimize surface area given volume — the minimizers are essentially iven by Hemming balls.

Now assuming Harper's theorem, here is a quick proof:

Theorem 10.21
If $|S| = k$ has distinct sums, then $\max S \geq \binom{k}{\lfloor k/2 \rfloor} = (\sqrt{2/\pi} + o(1))2^k/\sqrt{k}$.

Proof. As before let $S = \{x_1, \dots, x_k\}$ and let A be the subset of the Boolean cube correspondign to teh ϵ_i, so that $\epsilon_1 x_1 + \cdots + \varepsilon_k x_k < (x_1 + \cdots + x_k)/2$. Due to distinct sums, there do not exist ε for which there is equality here — if there was, you could cancel the two sides and get a collision fo sums. SO it follows by symmetry that $|A| = 2^{k-1}$ (we can always flip to get to the other side).

Now every element of ∂A corresponds to a subset sum that's in the interval $((x_1 + \cdots + x_k)/2, (x_1 + \cdots + x_k)/2 + \max S)$, since we started with something less than this quanitty and then added an elelment of S to cross the midpoint threshold.

But since all sums are distinct, this interval must accomodate at least $|\partial A|$ elements — so then $\max S \geq |\partial A|$. But by Harper's theorem this is at least $\binom{k}{\lfloor k/2 \rfloor}$, which completes the proof. □

Remark 10.22. This has nothing to do with the probabilisti cmethod, but we will see a version of Harper's theorem when we discuss concentration of method. This is an exmaple of a general phenomenon — given a subset of a cube, how large can its boundary be? We will see a slightly weakened version later on, and it'll be important.

§11 October 5, 2022 — The Chernoff Bound

The Chernoff bound is an extremely useful bound — it gives a sub-Gaussian tail bound for the sum of independent random variables.

Theorem 11.1

Let $S_n = X_1 + X_2 + \cdots + X_n$, where $X_i \sim \text{Unif}\{-1, 1\}$ are independent and identically distributed. Then for every λ, we have

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

This is the most basic version of the Chernoff bound; we will see other variations later. First, it's not too hard to check that $\operatorname{Var} S_n = n$, so $\sigma = \sqrt{n}$; also, $\mathbb{E}S_n = 0$. So this gives us a tail bound for S_n being far from its mean by a multiple of the standard deviation, and this bound decays as $e^{-\lambda^2/2}$. Such a bound — one of the form $e^{-\Omega(\lambda^2)}$ — is known as a **sub-Gaussian tail bound**, since it decays at least as fast as that of a normal distribution.

This bound is much better than we could do with the second moment — there we could only obtain $1/\lambda^2$. But here we have stronger assumptions — that S_n is the sum of i.i.d. random variables.

Proof. Let $t \geq 0$, and consider the **moment generating function**

$$\mathbb{E}e^{tS_n} = \mathbb{E}e^{t(X_1+\cdots+X_n)}.$$

Because of independence, we can split the product as

$$(\mathbb{E}e^{tX_1}) \cdots (\mathbb{E}e^{tX_n}).$$

Each factor is identically distributed, so we only need to consider one of them — so

$$\mathbb{E}e^{tS_n} = (\mathbb{E}e^{tX_1})^n = \left(\frac{e^{-t} + e^t}{2}\right)^n.$$

Now by comparing the Taylor series, we have

$$\frac{e^{-t} + e^t}{2} = \sum_{k \geq 0} \frac{t^{2k}}{(2k)!} \leq \sum \frac{t^{2k}}{k!\,2^k} = e^{t^2/2}.$$

(We can check this by term-by-term comparison.) Now by Markov's inequality, we see that

$$\mathbb{P}(S_n \geq \lambda\sqrt{n}) = \mathbb{P}(e^{tS_n} \geq e^{t\lambda\sqrt{n}}) \leq \frac{\mathbb{E}e^{tS_n}}{e^{t\lambda\sqrt{n}}} = \frac{e^{t^2/2 \cdot n}}{e^{t\lambda\sqrt{n}}}.$$

Now we get to choose t to optimize this quantity. We can write it out as

$$e^{-t\lambda\sqrt{n} + t^2\sqrt{n}/2}.$$

Setting $t = \lambda/\sqrt{n}$ gives the desired bound. □

This is an important technique — we use Markov's inequality on the *moment generating function*. This is quite similar to the proof of Chebyshev's inequality — there we also used Markov's inequality, on the square. This is similar, except that now we're taking very high moments — the exponential is kind of like a mixture of various moments. For some applications, you may be able to compute the kth moment (although not the exponential), and this can still be used to obtain bounds.

Even though the Chernoff bound is simple, it's extremely powerful, and it comes up all the time — because we need to analyze sums of i.i.d. Bernoullis all the time.

This bound is so versatile that Prof. Zhao doesn't know what to cite, and everyone is surprised that Chernoff is still alive — Prof. Zhao saw him at a faculty event (he was at MIT as a faculty member some years ago). (It's about as commonplace as Einstein's relativity — we just say it.)

Corollary 11.2

We have $\mathbb{P}(S_n \le -\lambda\sqrt{n}) \le e^{-\lambda^2/2}$, and by combining these we find

$$\mathbb{P}(|S_n| \ge \lambda\sqrt{n}) \le 2e^{-\lambda^2/2}.$$

The same bound is also true if each X_i is an independent mean-0 random variable taking values in the *interval* $[-1, 1]$ — everything stays the same except for the step

$$\frac{e^{-t}+e^t}{2} \le e^{t^2/2},$$

and the same inequality is still true (by a convexity argument). In particular, this allows us to derive the following consequence:

Corollary 11.3

If X is the sum of n independent Bernoulli variables (not necessarily with the same probability), then letting $\mu = \mathbb{E}X$ and $\lambda > 0$, then $\mathbb{P}(X \ge \mu + \lambda\sqrt{n}) \le e^{-\lambda^2/2}$, and likewise for the lower tail.

We could prove this using the same strategy, or by using the above corollary.

This is a correct statement, but there is some subtlety — if the Bernoulli variables have small probability p, this statement is still correct as written, but there is something to be careful of.

Remark 11.4. The Chernoff bound is quite tight, in some sense — we know $S_n/\sqrt{n}$ converges to the standard normal distribution, due to the central limit theorem. If Z is normally distributed, then we have

$$\mathbb{P}(Z \ge \lambda) = \mathbb{P}(e^{tZ} \ge e^{\lambda t}) \le e^{-t\lambda}\mathbb{E}[e^{tZ}] = e^{-t\lambda + t^2/2},$$

which becomes $e^{-\lambda^2/2}$ by setting an appropriate value of t. But this is quite tight — we could calculate the exact value as

$$\mathbb{P}(Z \ge \lambda) = \frac{1}{\sqrt{2\pi}}\int_\lambda^\infty e^{-t^2/2}\,dt \sim \frac{e^{-\lambda^2/2}}{\sqrt{2\pi}\lambda}.$$

When p is quite small, you can hope for better tail bounds — because the variance also is small. One form of such a bound is the following:

Theorem 11.5

If X is the sum of n independent Bernoullis, for every $\varepsilon > 0$,

$$\mathbb{P}(X \ge (1+\varepsilon)\mu) \le e^{-((1+\varepsilon)\log(1+\varepsilon)-\varepsilon)\mu} \le e^{-\frac{\varepsilon^2}{1+\varepsilon}\mu},$$

whereas

$$\mathbb{P}(X \le (1-\varepsilon)\mu) \le e^{-\varepsilon^2\mu/2}.$$

Note that this is asymmetric. In particular, the upper tail is no longer sub-Gaussian for large ε.

In particular, $\text{Binom}(n, c/n) \to \text{Pois}(c)$. The decay of the Poisson distribution is

$$\mathbb{P}(\text{Pois}(c) = k) = \frac{c^k}{k!\,e^c} \approx e^{-\Omega(k\log k)}.$$

For large k, this decays slightly quicker than exponential, but it's not sub-Gaussian.

So on one hand, the message of the Chernoff bound is that for a sum of independent random variables, we can get a sub-Gaussian bound. But this is really only true if we have constant-probability Bernoullis — for sparse probabilities and for the upper tail, it's not necessarily true.

§11.1 Discrepancy

Question 11.6. Given a hypergraph, we wish to color the vertices using two colors — red and blue — so that every edge has a similar number of red and blue vertices (a small discrepancy). Can we always guarantee a small discrepancy?

Theorem 11.7

Given a collection of m subsets of $[n]$, there exists an assignment $[n] \to \{\pm 1\}$ such that for each of the m subsets, the sum over its vertex assignments is $O(\sqrt{n \log m})$ (in absolute value).

We can imagine n and m as being comparable to each other in size. If we have m subsets of a ground set, then we can color the ground set so that there's only square-root discrepancy in each set.

Proof. Color uniformly at random (choosing ± 1 for each element). On each subset,

$$\mathbb{P}(|\text{sum}| > 2\sqrt{n \log m}) \le 2e^{-2\log m} = \frac{2}{m^2}$$

by the Chernoff bound. Now we can union bound over these m subsets to get that the failure probability is less than 1. □

As we can see here, the Chernoff bound is very powerful — it lets us have *very* small probabilities, so that we can apply the union bound. With the earlier concentration bounds from the second moment method, we wouldn't be able to get anything close to this bound.

In fact, we can actually do much better — we can remove the log factor in some circumstances (although we cannot do better than $\sqrt{n}$). There is a paper by Joel Spencer, titled *Six Standard Deviations Suffice*:

Theorem 11.8 (Spencer 1985)

For $m = n$, we can get $|\text{sum}| \le 6\sqrt{n}$ for every sum.

If we assign randomly, then we have a standard deviation of $\sqrt{n}$; and if we allow ourselves six standard deviations, we can get all the discrepancies to be small. The naive argument doesn't work any more; rather, Spencer had to introduce clever techniques that are iterative. He doesn't color everything all at once — instead, he tries to color some subset and then make progress. The idea is to take a *partial* coloring (possibly involving real numbers between -1 and 1), and iteratively make progress on that coloring.

This was a pretty important result — Spencer has said that it's the result he is most proud of. (Spencer will be giving a lecture in two weeks, called 'Three Theses' — he will probably go through some historical perspective on the probabilistic method, and talk about three theses that build the backbone of the probabilistic method. He will also give a combinatorics seminar talk after class, related to deviations.)

Another interesting story is that when Spencer proved this result, it was an existence result — there was a proof that you could get a coloring, but it was not algorithmic. For a long time, it remained open whether you could algorithmically construct a coloring with small discrepancy. Spencer at the time believed you could not. There were some major breakthroughs around 2010, where an efficient algorithm using very clever ideas was found.

Question 11.9. Is it known that the random method doesn't work?

Yes — if you analyze the proof, you can see that we really do need the $\sqrt{\log m}$.

More generally, if $m \geq n$, then the bound Spencer got was $O(\sqrt{n \log(2m/n)})$.

Conjecture 11.10 (Komlós' Conjecture) **—** There exists a constant K such that for any $v_1, \dots, v_m \in \mathbb{R}^n$, all in the unit ball, there exists some choice of signs $\varepsilon_1, \dots, \varepsilon_m$ such that

$$\varepsilon_1 v_1 + \cdots + \varepsilon_m v_m \in [-K, K]^n.$$

Essentially, we want to choose a $+/-$ combination such that the result lies in a box — or in other words, has bounded L^∞ norm.

Spencer's result proves a certain special case.

Note that n doesn't really matter — if it's really large, we can always just restrict ourselves to the subspace the vectors live in, so we can assume $n = m$.

If we take the v's to be the indicator vectors for the sets (normalized to be unit vectors), then the Spencer result corresponds to this statement (if we don't care about the constant).

The best result of this form proves the bound when $K = O(\sqrt{\log n})$ — so here there's a dependence on dimension, but the conjecture is dimension-free.

§11.2 A Geometric Example

Question 11.11. How many vectors in $\mathbb{R}^n$ can we place so that they all have equal angles with each other?

We can take a simplex centered at the origin, which gives an answer of $n + 1$. This is actually tight — there's a short linear-algebraic argument.

Proof. Let $S = \{v_1, \dots, v_m\}$ be such a configuration of unit vectors in $\mathbb{R}_n$; then we must have $\langle v_i, v_j \rangle = \alpha$ when $i \neq j$, where $\alpha \in [-1, 1)$. Now consider the *Gram matrix*

$$\begin{bmatrix} | & \cdots & | \\ v_1 & \cdots & v_m \\ | & \cdots & | \end{bmatrix}^t \begin{bmatrix} | & \cdots & | \\ v_1 & \cdots & v_m \\ | & \cdots & | \end{bmatrix} = \begin{bmatrix} v_1 \cdot v_1 & \cdots \\ \vdots & \ddots \end{bmatrix}.$$

Then this matrix must have 1's on the diagonal, and α's off the diagonal. So we can write it as

$$(1 - \alpha)I + \alpha J,$$

where J is the all-1's matrix. We know the eigenvalues of the all-1's matrix, so then the eigenvalues of this matrix are $(n-1)\alpha + 1$ once, and $1 - \alpha$ $(m-1)$ times. Because the Gram matrix is positive semidefinite, all eigenvalues are nonnegative. In particular, that means $\alpha \geq -1/(m-1)$ — in other words, if α is quite negative, we cannot have many vectors.

If $\alpha \neq -1/(m-1)$, then the matrix is nonsingular. Then its rank is equal to m. But its rank is at most n, since we started with vectors in $\mathbb{R}^n$. So then $m \leq n$.

Meanwhile, if $\alpha = -1/(m-1)$, then the rank is $m - 1 \leq n$, so we get $m \leq n = 1$. □

So no matter what the angle is, we have linear growth in the number of vectors with pairwise equal angles to each other.

Question 11.12. What happens if instead of asking for *exactly* equal angles, we ask for *approximately* equal angles?

Surprisingly, with just a little room for error, we can get exponentially many vectors:

Theorem 11.13

For every $\alpha \in (0,1)$ and $\varepsilon > 0$, there exists $c > 0$ such that for every n, there exists a collection of 2^{cn} unit vectors in $\mathbb{R}^n$ whose pairwise inner products all lie within ε of α.

So even though if we want a bunch of vectors 30° from each other we only get *linearly* many, if we now want between 30° and 31° we can get *exponentially* many.

Doing this by hand would be quite tricky. So instead, we construct these vectors randomly.

How do we construct vectors randomly? We don't really have the tools to analyze picking a random point on a sphere, so we will do something more discrete — we pick random vectors on a cube.

Proof. Let $x_1, \dots, x_m \in \{0,1\}^n$ be points on a cube, for some m to be decided. However, we don't want to choose these points *uniformly* at random — instead, we'll let each coordinate be 1 with probability α (we'll see why this is a good choice later), and everything is independent.

Now we have a large collection of vectors. Not all of them will have the desired inner products, but it turns out many of them will. In particular, for every i we find that

$$x_i \cdot x_i \sim \text{Binom}(n, \alpha),$$

since it's the sum of Bernoullis with probability α. So this is strongly concentrated aroud $n\alpha$, and in particular

$$\mathbb{P}(|x_I \cdot x_i - n\alpha| \geq nt) \leq 2e^{-nt^2/2}$$

(for some t we will pick later). Likewise, when $i \neq j$, their inner product is

$$x_i \cdot x_j \sim \text{Binom}(n, \alpha^2),$$

and we have the Chernoff bound estimate

$$\mathbb{P}(\left|x_i \cdot x_j - n\alpha^2\right| \geq nt) \leq 2e^{-nt^2/2}.$$

Now we know we have a bunch of points, and they very likely have the length we're looking for, and the inner products we're looking for.

But they could fail with some probability. To deal with that, we do alteration — we can get rid of the bad cases. Let $Y = \{i \in [m] \mid |x_i \cdot x_i - n\alpha| \geq nt\}$ be the set of vectors with undesirable lengths, and $Z = \{(i,j) \in [m]^2 \mid i \neq j \text{ and } \left|x_i \cdot x_j - n\alpha^2\right| \geq nt\}$ be the set of vectors with undesirable product. Then we want to throw out the bad vectors — first, let $Z' = \{i \in [m] \mid (i,j) \in Z\}$, and let $X = [m] \setminus (Y \cup Z')$ (the indices that *don't* get thrown out).

Then for distinct $i, j \in X$, we have

$$]frac\alpha^2 - t\alpha + t \leq \frac{x_i \cdot x_j}{\sqrt{x_i \cdot x_i} \cdot \sqrt{x_m \cdot x_j}} \leq \frac{\alpha^2 + t}{\alpha - t}.$$

We can pick t such that these upper and lower bounds are within ε of α (so t only depends on α and ε).

Then X works, and we want to know how many elements we have left. We know

$$\mathbb{E}X \geq m - \mathbb{E}\,|Y| - \mathbb{E}\,|Z|\,.$$

Meanwhile, by the Chernoff bounds we have the estimates

$$\mathbb{E}\,|X| \geq m - 2m^2 e^{-nt^2/2}.$$

By choosing $m = e^{nt^2/2}/4$, we get that $\mathbb{E}\,|X| \geq m/2$. This gives exponentially many points. Finally, we can normalize the vectors by their lengths and take $x_i/|x_i|$ for all $i \in X$, which works. □

Remark 11.14. This is somewhat surprising — there's a night-and-day difference between exactly equal and approximately equal.

One of the reasons Prof. Zhao is interested in this problem is that there's a related problem — what if you instead ask for equiangular *lines*? They solved a version of this problem where you fix the angle, and let the dimension go to ∞ — then the growth rate in the dimension is linear.

Student Question. What happens if $\alpha < 0$?

This proof doesn't work anymore. But the situation is actually very different:

- If we hvae m vectors in $\mathbb{R}^n$, all of whose pairwise inner products are at most $-\beta < 0$, then $m \leq 1 + 1/\beta$. So we'd have at most a constant number of vectors (and this is tight).
- If we have m vectors in $\mathbb{R}^n$, and all pairwise inner products are at most 0, then $m \leq 2n$, by taking $\pm$ of the unit vectors. (This is trickier to prove, but it can also be proven.)

So the requirement that $\alpha > 0$ is quite important.

Student Question. Is there combinatorial intuition behind why we get an exponential bound?

The intuition is sort of that the Chernoff bounds are really good — so the inner products are extremely concentrated, and you can have a lot of things that look roughly the same in $\mathbb{R}^n$.

§11.3 An Example from Graph Theory

Earlier, we touched upon some of this object before, when talking about the crossing number inequality.

Theorem 11.15 (Four-Color Theorem)

Every planar graph is 4-colorable.

In graph theory, this is a very important result, and there's a lot of effort to understand for what reasons this is really true, and how it generalizes.

One way is to generalize planarity.

Theorem 11.16 (Kuratowski, 1930)

A graph is planar if and only if there are no $K_{3,3}$ or K_5 subdivisions.

Definition 11.17. A **subdivision** is like a subgraph, but we're allowed one further operation — if we have a path, we can convert it to an edge.

Example 11.18

A K_5 with an extra vertex inserted into one of its edges is a subdivision of K_5.

Theorem 11.19 (Wagner, 1930)
A graph is planar if and only if it has no $K_{3,3}$ or K_5 as a minor.

A minor allows even more flexibility than in a subdivision:

Definition 11.20. In a **minor**, we are allowed *contractions*.

Example 11.21
Take a hexagon, with a K_4 at the bottom and a trapezoid with vertical down-edges from the top two vertices. Then we get to squeeze the top edge to form a K_5.

In particular, the 4-color theorem can be written in the following way:

Theorem 11.22
Every graph without a $K_{3,3}$ or K_5 minor is 4-colorable.

Between $K_{3,3}$ and K_5, which should matter more? Probably K_5 — it seems that K_5 is the true obstruction to 4-colorability.

Question 11.23. Is it true that not having a K_5 minor implies 4-colorability?

The answer is yes.

Conjecture 11.24 (Hadurger's Conjecture) — For every $t > 0$, every K_{t+1}-minor-free graph is t-colorable.

This is a pretty central conjecture in graph theory.

The case $t = 1$ is trivial — not having a K_2 minor means we have an empty graph.

The case $t = 2$ is nearly trivial — if we do not have a K_3 minor, then you can't have a cycle. So the graph is a tree or forest, which is 2-colorable.

The case $t = 3$ is not too difficult, but it is already interesting.

The case $t = 4$ is already equivalent to the 4-color theorem (this follows from Wagner's theorem — we can prove that we can remove teh $K_{3,3}$ hypothesis from the restatement of the 4-color theorem).

We only have computer-assisted proofs of the four color theorem, so we'd think that $t = 5$ should be much harder, or even false. But it turns out to *also* be equivalent to the four-color theorem! (A paper from the 1990s reduced it to the 4-color theorem.)

Meanwhile, every $t \geq 6$ is open, and this is a major conjecture in graph theory.

Wagner's theorem describes planarity in terms of minors, but we also have a definition of planarity in avoiding subdivisions.

Conjecture 11.25 (Hajos' Conjecture) — The same conjecture, but with minor replaced by subdivision.

It turns out this is too ambitious — this is false. (It's a weaker condition.) It is true for $t \leq 3$. Initially, it was disproved by a concrete construction for $t \geq 6$. (In fact, it's still open for $t = 4$ and 5.)

Today, we will see a result of Erdős and someone else that shows that not only is Hajós's conjecture false, but it's extremely false — it's false for $G(n, 1/2)$. A good lesson here is that it's good to check random graphs as examples for potential statements.

Theorem 11.26

With high probability, $G(n, 1/2)$ has no K_5-subdivision with $t = \lceil 10\sqrt{n} \rceil$.

This leads to a counterexample because we already saw that the independence number of $G(n, 1/2)$ is $\alpha(G(n, 1/2)) \sim 2\log_2 n$, and so $\chi(G(n, 1/2)) \geq n/\alpha(G) \gtrsim n/\log n$. So $G(n, 1/2)$ has very large chromatic number, but it has no small subdivisions.

Proof. If G had a K_t-subdivision, let's think about how it would look. We have a set S of t vertices. Then between every pair of vertices in S, one of two things can happen — either they're joined by an edge, or they're joined by some path that goes out of S (we'd need additional vertices to join paths outside).

How many vertices can we use outside of S? There's only n vertices in total. So the number of non-edges in S is at most n — if it's a non-edge, we have to eat up an extra vertex to form the red path. So we only have that much room to form these red paths outside S.

But that means the number of edges in S is at least $\binom{t}{2} - n$.

Now for a fixed t-element S, the number of edges in S should be about half of $\binom{t}{2}$, and the Chernoff bound says we shouldn't have much more than that —

$$\mathbb{P}(e(S) \geq \binom{t}{2} - n) \leq \mathbb{P}(e(s) \geq \frac{3}{4}\binom{t}{2}) \leq e^{-t^2/10}.$$

So the number of non-edges in S is at most $1/4$ of the available edges, so the edge density has to be at least 75%, and we woudl not expect that in a random graph (where we should have strong concentration). Taking a union bound over all t-vertex sets S, of which there are $\binom{n}{t}$, we have

$$e^{-t^2/10}\binom{n}{t} \leq e^{-t^2/10} n^t.$$

The first factor is significantly smaller than the second, so this goes to 0 as $n \to \infty$. In other words, no such S exist in $G(n, 1/2)$ with high probability. □

To recap the main idea, the claim is that $G(n, 1/2)$ has no K_t-subdivision for $t \approx 10\sqrt{n}$. By the Chernoff bound, S typically has no more than 75% edge density (it's really around 50%). This means for all the non-edges, we need to find extra vertices outside S to form the paths. But we don't have much room outside — we only have n vertices. So a typical random graph fails Hajós's conjecture, suggesting that you should check your conjectures on random graphs (which is standard now, but was not at the time).

§12 October 12, 2022 — Lovasz Local Lemma

The Lovasz Local Lemma is an important tool in the probabilistic method. Often, we want to avoid some collection of bad events. There are a couple of cases where things are particularly easy —

(a) If we have complete independence, then the success probabilities multiply, which is easy to analyze.

(b) If all the probabilities add up to less than 1, we can take a union bound.

The local lemma lets us analyze *intermediate* situations. In an application to the local lemma, each event has a few or a small amount of dependencies, and these dependencies are fairly local in some sense — like looking at a graph with small maximum degree.

§12.1 The Setup

First, we need to define a somewhat subtle notion.

Definition 12.1. We say that an event A_0 is **independent** from a collection $A_1, \ldots, A_m$ of events if A_0 is independent of every event of the form $B_1 \cap B_2 \cap \cdots \cap B_m$, where each B_i is either A_i or $\overline{A_i}$ (this denotes a complement).

If we have two events A and B, recall that A and B are independent if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B).$$

Equivalently, we can write this in terms of conditional probabilities, as

$$\mathbb{P}(A \mid B) = \mathbb{P}(A).$$

So when we say A_0 is independent from a *collection* of events, we're saying it's independent from every logical combination of events. That *doesn't* just mean it's independent from every A_i individually — that would be much weaker.

Given a collection of events, we can record their dependencies via a dependency graph:

Definition 12.2. Let $A_1, \ldots, A_n$ be events (these are the bad events we wish to avoid). We let G be a (directed) graph with vertex set $[n]$, and edges which roughly correspond to dependencies. More precisely, we say that G is a **dependency (di)graph** for these events $A_1, \ldots, A_n$ if: for all i, A_i is independent from all $\{A_j \mid j \notin \{i\} \cup N(i)\}$, where $N(i)$ is the set of *out*-neighbors.

In practice, we usually don't need to worry about direction; but in complete generality we would a digraph.

So we have a graph with a bunch of events, and a bunch of edges connecting the events. THese edges correspond to dependenceis, in teh sense that each vertex is independent from all logical combinations of its non-neighbors.

Remark 12.3. This is *not* built by joining A_i and A_j whenever they are not dependent, meaning $\mathbb{P}(A_iA_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$. In second moment calculations we only worried about pairwise dependencies, but here we need to do more.

As a concrete example:

Example 12.4

Consider $X_1, X_2, X_3 \in \mathbb{Z}/2\mathbb{Z}$ chosen uniformly at random. Let A_1 be the event that $X_2 + X_3 \equiv 0 \pmod 2$, and define A_2 and A_3 similarly (as the events that $X_1 + X_3$, and $X_1 + X_2$, are zero).

Then A_1 and A_2 are independent — $\mathbb{P}(A_1A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$. (We can see that if we hold x_3 constant, then the rest is independent.) Likewise, any other pair is also independent. But A_1, A_2, and A_3 are *not* independent — for example, if A_1 and A_2 occur, this implies A_3 (since their sum is 0).

So a valid dependency graph would be: first, for any set of events, having all the edges is a valid dependency graph. In this case, just having two edges is a valid dependency graph.

In particular, the notion of dependency graph is nto unique. That mkaes it a bit cumbersome, but in practice it's usually not so bad — we are almost always working with a *random variable model.*

§12.1.1 Random Variable Model

The random variable model, or hypergraph coloring, is what we saw in the first lecture.

The setup is that we have a collection of independent random variables x_i for $i \in I$, and we have a collectino fo events $E_1, \ldots, E_n$ where each event E_i depends only on some subset of the variables — only by variables indexed by some subset $B_i \subseteq I$.

In this case, we can construct a canonical dependency graph — where the vertices are indexed by $[n]$, and we have an edge between i adn j if and only if $B_i \cap B_j \neq \emptyset$.

We can check that in this case, this is indeed a valid dependency graph — A_i is independent from all events it's not connected to, becaues A_i doesn't use any of hte variables involved in its non-neighbors. So no matter what logical combination you take, you have independence.

This is the main model we'll be discussing in this chapter — even in this model, there's a lot of interesting things to say.

§12.2 Boolean Satisfiability

One reason LLL is so significant is because of its applicatiosn to computer science, in particular to the class of problems called **Boolean satisfiability** (SAT).

Definition 12.5. A **CNF** formula (conjugate to normal form, or 'and of ors') is a formula

$$(x_1 \cup x_2 \cup x_3) \cap (\overline{x_1} \cup x_2 \cup x_3) \cap \cdots.$$

(Use triangular things for or/and instead of circular.)

Here A_i is the event that the first clause fails. Then A_1 and A_2 share variables, so we draw an edge between them.

This is an important class of problems, and it's one of the reasons the local emma plays an important role in computer science.

§12.3 The Local Lemma

The point of the local lemma is that if we have a small amount of dependencies, we can avoid all the bad events. The most common form is the symmetric form:

Theorem 12.6 (LLL, symmetric form)

Suppose we have events $A_1, \ldots, A_n$ such that $\mathbb{P}(A_i) \leq p$ for all A_i, and suppose each A_i is independent from a set of all other A_j's except at most d of them. (In other words, there exists a dependency graph with maximum degree at most d.) If $ep(d+1) \leq 1$, then with positive probability, none of the bad events $A_1, \ldots, A_n$ occur.

So we have a set of events with a sparse dependency graph (with maximum degree d). Then as long as $ep(d+1)$ is at most 1 (where e is the constant), we can avoid all the bad events. It turns out e is actually the best constant you can put here, but in practice we don't generally care about the exact value of the constant — in the original paper they put 3.

The point here is that this is almost like a union bound locally — $p \cdot d$ is like a union bound.

We will deduce the symmetric form from a more general statement.

Theorem 12.7 (LLL, general)

Suppose we have events $A_1, \dots, A_n$ that we wish to avoid. Suppose we have some dependency graph where i has out-neighbors $N(i)$. If $X_1, \dots, X_n \in [0, 1)$ satisfy the inequality that for every i,

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

then $\mathbb{P}(\text{none of the events occur}) \ge (1 - x_1)(1 - x_2) \cdots (1 - x_n)$.

In particular, the probability is positive. Note that this is still a local condition.

First let's prove the symmetric form from the general form.

Proof that gen implies sym. In the general form, to apply it we have to specify the funny variables x_i (these are some real numbers, not random variables). In the setting of the symmetric form, we want to set $x_i = \frac{1}{d+1} < 1$. Then we can check that

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

since the degree of every node is at most d. We can check that the second term is bigger than e, so then

$$x_i \prod (1 - x_j) > \frac{1}{(d+1)e} \ge p$$

by the hypothesis. □

The symmetric form is useful when there's symmetry — all the variables are similar — but in some cases you'll have to figure out by hand what the appropriate stuff to set in are.

Here is another corollary of the general form:

Corollary 12.8

In the above setup, if all the events have $\mathbb{P}(A_i) < 1/2$, and for every i we have $\sum_{j \in N(i)} \mathbb{P}(A_j) \le 1/4$, then with positive probability none of the events occur.

Here we can again look at the general form and figure out some appropriate values to set. We want to put in some x_i's so that this inequality is satisfied. A general rule is to set the x_i's to be somewhat similar to the event probabilities. So here we can set $x_i = 2\mathbb{P}(A_i)$ (usually there is a bit of room we can work with). Then our expression on the RHS of the LLL hypothesis is

$$\ge x_i(1 - \sum_{j \in N(i)} 2\mathbb{P}(A_j)) \ge 2\mathbb{P}(A_i)(1 - 2\sum \mathbb{P}(A_j)).$$

The second thing is at most $1/2$, so this gives the desired bound.

This isn't just about having few neighbors — maybe we have lots of neighbors but all of them have small probabilities, adn that's okay too. The other is that contrary to this calculation, we might not want to set $x_i = 2\mathbb{P}(A_i)$ all teh time — we might want to set x_i to be different — and we will explore this in teh next problem set, where we'll need to figure out the appropriate x_i to set.

Now let's prove the general form of the LLL. It's not long, but it's very clever.

The LLL first appeared in a paper due to Erdős and Lovasz in teh 1970s.

Proof of LLL. We will prove the following statement by induction:

$$\mathbb{P}(A_i \mid \text{big cap triangle denoting and}\overline{A_j}) \le x_j$$

for any set S of indices j not containing i.

So we will rpove that having already avoided some arbitrary collection fo bad events, the probability of A_i is at most x_i. Once the above is proven, then we can deduce that

$$\mathbb{P}(\overline{A_1} \cdots \overline{A_n}) = \mathbb{P}(\overline{A_1})\mathbb{P}(\overline{A_2} \mid \overline{A_1})\mathbb{P}(\overline{A_3} \mid \overline{A_1 A_2}) \cdots \mathbb{P}(\overline{A_n} \mid \overline{A_1} \cdots \overline{A_n})$$

by Bayes' Rule. Now we can bound the first factor by $1 - x_1$, the second by $1 - x_2$, and so on. SO

$$\mathbb{P}(\overline{A_1} \cdots \overline{A_n}) \ge (1 - x_1)(1 - x_2) \cdots (1 - x_n).$$

We will prove this by induction on $|S|$.

First, the base case is when $|S| = 0$. When S is empty, then we already know that $\mathbb{P}(A_i) \le x_i \prod_{j \in N(i)}(1 - x_j) \le x_i$, so there is nothing to prove.

Now let $i \notin S$. We partition S into two sets: S_1 is the neighbors of i, and S_2 are the non-neighbors of i. (So $S_1 = S \cap N(i)$ and $S_2 = S \setminus S_1$).

Now let's do a calculation. We want to explore

$$\mathbb{P}(A_i \mid \bigcap_{j \in S} \overline{A_j}).$$

Using Bayes' Rule, we pull out the events in S_1 — we can write this as

$$\frac{\mathbb{P}(A_i \cap j \in S_1 \mid j \in S_2)}{\mathbb{P}(\cap_{j \in S_1} \overline{A_j} \mid \cap_{j \in S_2} \overline{A_j})}.$$

In teh numerator, we can first drop the first part — this is upper-bounded by $\mathbb{P}(A_i \mid \bigcap_{j \in S_2} \overline{A_j})$. But now since A_i is non-adjacent to everything in S_2, it's independent from all logical combinations — so this conditional thing does nothing, and we can replace this expression by $\mathbb{P}(A_i)$. We know from the hypothesis that

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

Now let's go to the denominator. This is some probability of a combination fo events, so let's use Bayes' rule in chain to rewrite the denominator. Writing teh elemnts of S_1 as $j_1, \ldots, j_r$, we can decompose this as the chain fo conditional probabilities

$$\mathbb{P}(\overline{A_{j_1}} \mid E)\mathbb{P}(\overline{A_{j_2}} \mid \overline{A_{j_1}} E) \cdots \mathbb{P}(\overline{A_{j_r}} \mid \overline{A_{j_1}} \cdots A_{j_{r-1}} E)$$

where E is the and of complements of stuff in S_2. Now we can use the induction hypothesis. ALl the events conditioned here are of the form as in the induction hypothesis — they're ands of complements of bad events, and the number of events we're conditioninign on is a strict subset of S. So we can apply the induction hypothesis to conclude that the first probabiltiiy as at least $(1 - x_{j_1})$, the second is at least $(1 - x_{j_2})$, and so on. In particular, this is at least $\prod_{j \in N(i)}(1 - x_j)$ (because we may need to throw in some other neighbors, but the inequality goes in the right direction).

Now we are basically done — putting the numerator and denominator inequalities together, we see that the expression is at most x_i, which is what we were trying to show in the induction step.

So this finishes the proof by induction fo the claim, and once we have this claim we can chain together the Bayes' rule calculation to derive LLL. □

This isn't long — everything fits on two blackboards — but it's tricky, and Prof. Zhao doesn't know how they came up with it. In the rest of this chapter, we'll focus on how to apply LLL — we'll see lots of applications, and some twists and turns with what to even call the events (as well as other variations).

Remark 12.9. Note that we're only using independence in a certain direction. Later we'll see the *lopsided LLL*, where we only need dependence not affecting you in one direction.

Remark 12.10. The "in other words" referring to the dependency graph (in the symm version statement) is not exactly true, but in practice it doesn't matter.

§12.4 Coloring Hypergraphs

Previously, we saw the following simple fact:

Theorem 12.11

Every k-uniform hypergraph on fewer than 2^{k-1} edges is 2-colorable.

You can get this by union bound.

Theorem 12.12

A k-uniform hypergraph is 2-colorable if every edge intersects at most $2^{k-1}/e - 1$ other edges.

So we can have a lot of edges, but as long as there's very few *interactions* — local interactions, so every edge intersects a small number of edges (exponential in k) — then we can two-color teh vertices to have no monochromatic edges.

Proof. First let's see how to set up. Here we can use the random variable model — for every edge f, let A_f be the event that f is monochromatic (under a uniform random coloring). This follows the random variable model, and we can have the dependency graph where each A_f is joined to other $A_{f'}$'s that intersect A_f. In the canonical dependency graph for the random variable model, we have $A_f \sim A_{f'}$ if f and f' intersect.So the maximum degree is then at most $2^{k-1}/e - 1 = d$.

Then we can check that $ep(d+1) \leq 1$, so the LLL in its symmetric version applies. □

In particular:

Corollary 12.13

For all $k \geq 9$, every k-uniform k-regular hypergraph is 2-colorable.

Here regular means that every vertex is in exactly k edges.

Proof. Every edge intersects at most $k(k-1)$ edges — there are k vertices to consider, each is in $k-1$ other edges (this may overcount but that's fine), and we have

$$e(k(k-1)+1)2^{-k+1} < 1$$

for $k > 9$. □

Remark 12.14. As a side-story, what about smaller values of k? For $k = 2$, a triangle is not 2-colorable, so the answer is no.

For $k = 3$, is a 3-uniform 3-regular hypergraph 3-colorable? We've actually seen an example before — the Fano plane is a 3-uniform 3-regular hypergraph that is not 2-colorable.

In fact, though, it's true whenever $k \geq 4$ — this was proved by Thomason in the 90's, using sophisticated tools from graph theory.

Here is a slightly contrived example, where the symetric version of LLL is not sufficient.

Theorem 12.15

Let $\mathcal{H}$ be a non-uniform hypergraph (edges can have different sizes), where every edge has size at least 3 and

$$\sum_{f \text{intersecting } e} 2^{-|f|} \leq \frac{1}{8}$$

for all edges e. Then H is 2-colorable.

The point of this example is that firs,t it follows from the corollary stated earlier (if all neighbors have small totla probability, then we can apply the asymmetric form of LLL). But the main point is that the symmetric version is insufficieint — if we only knew the symmetric version, we could not prove this statement. That's because you could have some vertices with very high degrees, and some with fairly high probabilities.

An indictaion of when you should paply the asymmetric form is when you have events of very different nature (or very different types) — the event that an edge is monochromatic depends quite a lot on the size of that edge. Then we should try assigning different x_i's for each event.

§12.5 Dealing with Infinity

Now we'll see a compactness argument that allows you to apply this even to settings with an infinite number of events and variables.

To apply LLL, we only need to knwo locally what happens — the condition does not depend on teh totla number of vertices. So we can imagine the total number of vertices being unbounded or infintie, and it makes sense that we should be able to apply it in that situation as well .

Theorem 12.16

Let H be a non-uniform hypergraph on a possibly infinite set, such that each edge is finite and has at least k vertices, and intersects at most d other edges. If $e2^{-k+1}(d+1) \leq 1$, then H has a proper 2-coloring.

The point is that H now gets to have an infinite number of vertices. Note that if H is finite, then this already follows from the symmetric version of LLL. We color uniformly at random, and each edge is monochromatic with probability 2^{-k+1}; the maximum degree in the dependency graph is at most d. So if these stuff multiply to at most 1 then LLL applies.

The point of what we will do next is how to extend from finite to infinite.

The point is that in the random variable model, if we can color any finite set of constraints, then we can satisfy everything. There's a compactness argument that lets us do that.

Proof. The set of colorings is given by $[2]^V$, where V is any (countable or uncountable) set. We know that if we restrict V to any finite subset, then we can find a good coloring.

But $[2]^V$ is more than just a set — it's also a topological space. And by Tikhonov's Theorem (which says that a product of compact spaces, including an infinite product, is compact in the product topology — where open an dlcosed sets are defined by restricting to finite sets of coordinates).

For each finite subset $X \subseteq V$, let $C_X \subset [2]^V$ be the subset of colorings where no edge contained in X is monochromatic. So this is a finite restriction fo the problem. Because we could use LLL for finite hypergraphs, we can conclude that $C_X \neq \varnothing$ for all X. (C_X is a set of colorings of all vertices, but we only need to care about the edges contained in X.)

Note that if $Y \subseteq X$, then $C_Y \supset C_X$ — if we restrict X to a smaller subset, then the constraints we need to satisfy are a subset of the original one, and we can get potentially more (but not fewer) colorings.

In particular, if we have a bunch of finite subsets $C_{X_1} \cap C_{X_2} \cap \cdots \cap C_{X_n} \supset C_{X_1 \cup \cdots \cup X_n}$ is nonempty.

So $\{C_X \mid X \subseteq V, |X| < \infty\}$ is a collection of closed subsets of 2^V — each is closed under the product topology because we're only specifying a finite set of coordinates, and these have the **finite intersection property** (a property of a collection of sets, saying that the intersection of any finite subcollection is nonempty).

One of the defining facts about compactness (from point-set topology) is (one definition is equivalent to the following):

Definition 12.17. A space is compact if and only if every family of closed subsets having the finite intersection property (FIP) has a nonempty intersection.

You can view this as a definition fo a compact space. (We probably saw other definitions, such as finite open covers; this is equivalent, by taking the complement.)

Then goign back to our argument, by the compactness of $[2]^V$, the intersection of C_X over all finite subsets $X \subseteq V$ is nonempty. and any element in this intersection corresponds to a desired coloring. □

That finishes the argument, extending from finite coloring to infintie coloring via compactness. This is a pretty general argument, and it works pretty easily in a lot of situations. For example:

Example 12.18

IF we have a random variable model where each variable has finitely many choices, but there can be infinitely many events, then if it is possible to avoid every finite collection of events, then it's possible to avoid all events.

Note that the *probability* of avoiding all the events is 0 — but it's still possible. In a much simpler model, we're asking whether it's possible to color $\mathbb{Z}$ so that all elements are red. Obviously you can do this, but if you do it randomly the probability is 0. We should not confuse positive probability with existence in the infinite case.

§13 October 17, 2022

We've been talking about the Lovasz Local Lemma.

In teh LLL, we have a collection of *bad* events A_i that we wish to avoid. We saw two forms: first, we have a dependency graph, where each event A_i is independent from all its non-neighbors (meaning any logical combination of its non-neighbors). Usually we use the random variable model — so we draw an edge if they share variables.

The symmetric vrsion says that if the maximum degree is d, and $ep(d+1) \leq 1$, and all probabilities are at most p, then it is possible to avoid all the events.

The asymmetric form, which is more generla, says that if we can select variables x_i such that the probabilities are at most some expression in the x_i, then it's possible to avoid all the A_i's.

We can prove the symmetric form from the asymmetric one. But sometimes i'tll be useful to go to the asymmetric form and explicitly choose the x_i.

We also saw a compactness argument — it suffices to check local conditions. SO ieven if we have an infinite situation, as long as there's finitely many choices for each variable and each event depends on a finite set of variables, if we allow infinitely many events, then if it's possible to avoid any finite subset of elements then it's possible to avoid all events. It's important that we're working in the random variable, and that there's only finitely many choices for each variable — if these assumptions aren't satisfied, then you can come up with counterexamples.

§13.1 Hypergraph Colorings

Our first example comes from teh paper where LLL was introduced — they looked at many examples related to hypergraph colorings.

> **Question 13.1.** For each k, is there some m so that for all subsets $S \subseteq \mathbb{R}$ with $|S| = m$, one can k-color $\mathbb{R}$ so that every translate of S is multi-colored (meaning that all k colors appear)?

This is a question about coloring an *infinite* hypergraph. But via the compactness argument, it suffices to consider any *finite* collection of translates of S. Then via the argument we discussed last time, we can extend this to the entirety of the real line.

In any finite collection fo translates, there's only finitely many points. So then we can perform a uniform random coloring; we'd like to use LLL to check that this uniform random coloring has a positive chance of succeeding.

Theorem 13.2

The answer is yes, if

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

Proof. Uniformly and independently pick a color for each point.

Now given a fixed translate with m points, the probabiltiy that this translate is not multicolored is at most

$$k \cdot \left(1-\frac{1}{k}\right)^m$$

(since $(1-1/k)^m$ is the probability that we're missing a given color), by the union bound (over the probability that we're not using the ith color). Call this quantity p.

Also note that every translate of S intersects at most $m(m-1)$ other translates of S: for example, suppose $S = \{0,1,3\}$, which we can draw as 3 points. Then how many translates of S can intersect $\{0,1,3\}$? We need to figure out which two positions collide — so if we want to collide with 0, we need to figure out which other point to shift to there.

So now we have a canonical dependency graph, where the vertices are translates of S and the edges are overlapping relations. Then each vertex has degree at most $m(m-1)$. The inequality $e(d+1)p \leq 1$ i ssatisfied by the hypothesis, so LLL tells us it is posible to color our finite set of points to make sure that everything is multicolored. □

In particular, for every k, if we take m large enough then teh left-hand side goes to 0. If we work out the asymptotics, we will find that this proof gives that if $m > (3+\varepsilon)k \log k$ for all large enough k, then this proof works.

It might take some getting used to setting up this dependency graph and figuring out the edges and relations; we will be doing that a lot today.

§13.2 Coloring Integers to Avoid Monochromatic APs

Theorem 13.3

For every $\varepsilon > 0$, there exists some k_0 and a 2-coloring of $\mathbb{Z}$ so that there are no monochromatic k-term arithmetic progressions with $k \geq k_0$ and common difference less than $2^{(1-\varepsilon)k}$.

So we wish to color $\mathbb{Z}$ to avoid any monochromatic k-term arithmetic progressions with k large enough and common difference less than some number which depends on k.

The point of this example is to see how to apply the asymmetric form of LLL.

This is an example similar to earlier of a coloring model. The edges of the hypergraph we're trying to color correspond to k-term arithmetic progressiosn.

So let's pick a uniform random coloring — uniformly and independently at random 2-color $\mathbb{Z}$. For each k-term arithmetic progression with the desired properties, consider the event that this AP is monochromatic. Then $\mathbb{P}() = 2^{-k+1}$, since these k points should all have the same color.

We need to consider *alL* AP's satisfying (*). Of course there's infiintely many, but that we know how to deal with. The bigger issue is that there's very different *types* of events — one corresponding to each k, which can go from k_0 to any large integer, and the probability of satisfying is wildly different. And there are infinitely many other events that intersect a specific event.

So this is not a graph of finite degree, which means we are not going to be able to use the symmetric form. Instead, we will use the asymmetric form — h aving events of very different probabilities is indicative thatw e hsould.

Recall that we need to pick the x_i's. There is a bit of an art to picking the right x_i's, especially when the situation gets complicated. But generally you should pick the x_i's to be somewhat close to the actual event probabiliities.

So in this case, if $\mathbb{P}(A_i) = 2^{-k+1}$, then we will pick $x_i = 2^{-(1-\varepsilon/2)k}$. (We will see why this choice later, but the point is that it's roughly similar to the event probability, with a little bit of room.) Another way to see this is that it's roughly twice the event probability, raised to a power that's close to 1.

With this choice, let's now think about what we need to check. Fix a k-AP P. Then the number of ℓ-APs satisfying (*) that intersects P is — we need to pcik a position in the k-AP and a position in the ℓ-AP, and then we need to figure out teh common difference of the ℓ-AP, so we get at most $k\ell 2^{(1-\varepsilon)\ell}$.

Now we want to verify the thing in the asymmetric LLL. To apply LLL, we want to show that

$$2^{-k+1} \leq 2^{-(1-\varepsilon/2)k} \prod_{\ell \geq k_0} \left(1 - 2^{-(1-\varepsilon/2)\ell}\right)^{k\ell \cdot 2^{(1-\varepsilon)\ell}}.$$

We want this to be true for all $k \geq k_0$. Once we can show this is true, then the Lovasz local lemma applies to any finite collection of APs, and by compactness we can extend it to all the APs.

Once you write down what you need to show, it's fairly straightforward to verify. We can see here why we picked x_i to be close to $\mathbb{P}(A_i)$ — wehn we bring it over we see that we end up with wanting

$$2^{-\varepsilon k/2+1}.$$

For the right-hand side we use $1 - x \geq e^{-2x}$ for $x \in [0, 1/2]$, so we see that

$$\prod (1-2) \geq \exp\left(- \sum_{\ell \geq k_0} 2^{1-(1-\varepsilon/2)\ell} k \ell 2^{(1-\varepsilon)\ell} \right).$$

Stuff cancel out and we get

$$\exp\left(-k \sum_{\ell \geq k_0} \ell 2^{1-\varepsilon\ell/2} \right).$$

Now this is an exponential decreasing series for fixed ε, so we can choose k_0 large enough to make this thing smaller than say $\varepsilon/4$, and that allows us to deduce that this entire expression is at least the left-hand side, which is what we want.

The message to take away here is that the sign to use asymmetric LLL is when you have ver different types of events, with very different probabilities. And often you want to assign x_i's similar to the probabilities; this often works out well if it's indeed possible to use the local lemma.

§13.3 Decomposing Coverings

Definition 13.4. We say that a collection of disks (balls) in $\mathbb{R}^d$ is a **covering** if their union is $\mathbb{R}^d$.

So we try to cover the space using disks.

Definition 13.5. We say that a collection fo disks is a k-fold covering if every pointin $\mathbb{R}^d$ is contained in at least k disks.

So a 1-fold covering is the same as a covering — in general we're trying to cover each point k times using some geometric configuraiton of disks.

Definition 13.6. A k-fold coloring is **decomposable** if it can be partitioned into two coverings.

So we have a k-fold covering, where every point is covered k times. It's decomposable if I can split the covering by putting each disk in either the first set or second set, so that both sets individually form a covering.

Now a basic question is:

Question 13.7. Is it always possible to decompose a covering?

Exercise 13.8. In $\mathbb{R}^1$, every k-fold covering by intervals can be decomposed into k individual coverings.

This is quite special about $\mathbb{R}^1$ — if you have a collection fo intervals such that ever ypoint is covered at least k times, then we can split it into k different colroigns of the line.

Theorem 13.9 (M...lewska and Pack 86)
Every 33-fold covering of $\mathbb{R}^2$ by disks is decomposable.

Why 33? We can construct a 2-fold coering that is not decomposable. But this is not what we want to discuss — what we want to discuss is in higher dimensions.

Surprisingly, the same people also showed that;

Theorem 13.10

For every k, there exists a k-fold indecomposable covering of $\mathbb{R}^3$ using disks.

This is also true in higher dimensions.

So there exists a 1000-fold covering of $\mathbb{R}^3$ but we cannot partition teh disks into two points that individually form colorings. (Disks and balls are the same; they can have any size.)

This is maybe somewhat surprising. It is not what we will show; we will talk about the following somewhat-related fact, also shown by them:

Theorem 13.11

Every k-fold indecomposable covering of $\mathbb{R}^3$ by open unit balls must cover some point at least $\gtrsim 2^{k/3}$ times.

You might think that it might be possible to cover the space quite evenly. It turns out that this is not the case — if it's indecomposable, then there must be some point which is covered an enormous amount of times.

What in the world does this have to do with LLL?

Let's try to analyze this problem from the perspective of hypergraph colorings. Really, what we're trying to show is a contrapositive — we want to show that if we have some k-fold covering of $\mathbb{R}^3$ which covers every point not too many times, then we can decompose it. A decomposition means coloring every ball red or blue, so that no point is *only* covered by red balls, and no point is *only* covered by blue balls.

So let's consider the contrapositive — we want to show that if we have a k-fold covering, where every point is covered less than $c2^{k/3}$ times, then there exists a 2-coloring of the balls so that no point in $\mathbb{R}^3$ is covered only by red balls or only by blue balls (balls of one color).

So this is why the problem is an instance of the hypergraph coloring problem, once we've set up the appropriate hypergraph.

Before giving the proof, we'll need ap urely geometric lemma:

Lemma 13.12

A set of $n \geq 2$ spheres in $\mathbb{R}^3$ cuts $\mathbb{R}^3$ into at most $O(n^3)$ connected componens.

We certainly have an upper bound of 2^n, because every point is in or out of some sphere.

Can we get a much better bound? Let's suppose we start to put a sphere, and now we put another sphere in. The second sphere creates at most 2 extra regions in this case. Now let's put one more sphere. What if we put one more sphere?

Eahch new region is described by the nubmber of previous regions. Interesting stuff happens on teh existing spheres.

Let's zoom in on the first (yellow) sphere. Then we start putting spheres left and right, and we only look at what's hpapneing on the surface of this sphere. THey start cutting up this sphere into some number of regions.

So we're drawing n circles on the sphere, which divide the sphere into how many regions?

Essentially, we're counting crossings – each time we cross, there's a new region (ignoring a factor of 2). So we get $O(1 + 1 + 2 + 3 + \ldots) = O(n^2)$ regions — each circle we put in is going to create an extra $O(n)$ regions, so together we have $O(n^2)$.

Now let's think about adding spheres. The first sphere gives you $O(1)$, the second $O(2^2)$, and so on, so we get $O(1^2 + \cdots + n^2)$ —- once you introduce the nth sphere in, it cuts up each surface into n^2, and altogheter you have n^3.

This is quite handwavy but you can make it rigorous. It is important that this is much smaller than the naive combinatorial estimate of 2^n.

Now let's go back to our theorem.

Proof. Suppose for contradiction that every point in $\mathbb{R}^3$ is covered by at most $t = c \cdot 2^{k/3}$ unit balls.

Now we construct an infinite graph — we have an infinite hypergraph $\mathcal{H}$. The vertex set corresponds to the set of balls, and the edges have the form E_x, where E_x is the set of balls containing x. In a two-dimensional picture, with 4 balls (disks), we'd have four vertices, and for each point in space we'd have an edge through the balls that are in it. We see that if we move x around a little bit, as long as it stays in the same connected region, that edge doesn't change; we don't include repeate edges. So we only have one edge for each region.

Claim 13.13 — Every edge intersects at most $O(t^3)$ other edges.

This is not hard, but it's a bit confusing to think about, because we are swithcing between points and edges.

Proof. Let's fix $x \in \mathbb{R}^3$, and think about $E_x \cap E_y \neq \emptyset$. This means that they share a ball — so we have x, and we have y, and if E_x and E_y have nonempty intersection then there is some ball containing both x and y. In particular, $|x - y| \leq 2$ (since we have unit balls).

Now let's think about all the balls in E_y. These all lie in the radius-4 ball centered at x. Since every point lies in at most t balls, and the total voluem of this space is 4^3 times t he volume of the unit ball, this means there are at most $4^3 \cdot t$ balls appearing among those E_y intersecting x.

And these balls cut the radius-2 ball centered at x into $O(t^3)$ connected regions, by the earlier lemma. Two different y's lying int he same region give identical E_y's. Therefore E_x intersects $O(t^3)$ other edges.

Now with the parameters we've chosen — $t = c \cdot 2^{k/3}$ and $d = O(t^3)$ — we have $e \cdot 2^{-k+1}(d+1) \leq 1$. So we can apply the symmetric form of the local lemma, and together with a compactness argument we see that this hypergraph is 2-colorable. This 2-coloring of the hypergraph produces exactly what we wan t— a decompositiion fo the covering into two separate coverings (because the fact that no edge is monochromatic means that both sets are coverings). □

To recap, we wanted to show that every k-fold indecomposable covering of $\mathbb{R}^3$ by open unit balls must cover some point at least $2^{k/3}$ times. The point si to rephrase this problem into one about hypergraph colorings — vertices correspond to unit balls, and we want to avoid some point that's only covered by red balls, so we're coloring balls red and blue and we don't want any point only colored by red or blue balls.

And we can convert that into hypergraph, add we want to check the maimum degree. That boils down to analyzing some combinatorial geometry. □

For hypergraph colorings, it is usually straightforward to figure out hwat are the bad events — it may take some work to analyze the dependency graph, but teh setup is fairly straightforward (there is one fore ach edge).

But in many interesting applications, it is not *a priori* clear what should be used as ba events. The next example is fairly simple, but it illustrates there is some room to figure out what we should use as bad events.

§13.4 Large Independent Sets

Fact 13.14 — Every n-vertex graph with maximum degree δ has an independent set of size at least $n/(\delta+1)$.

For example, we can choose greedily — take a vertex, put it in the independent set, remove all its neighbors, ad repeat. The point of the next statement is to show that with some additional information, if you don't care about constant factors, you can guarantee additional structure:

Theorem 13.15

Let $G = (V, E)$ with maximum degree Δ, and let $V = V_1 \cup \cdots \cup V_r$ where each part V_i has size at least $2e\Delta$. Then there exists an independent set in G containing one vertex from each V_i.

So the graph has some additional structure — we have this partition — and we want to choose one vertex from each vertex-set.

A natural source of randomness is to pick a vertex from each set. The slightly tricky step is that we are going to first shrink all the V_i's to have size *exactly* something — we may assume that *every* V_i has size exactly $k = \lceil 2e\Delta \rceil$, because we can otherwise remove extraneous vertices. (This doesn't hurt, so we might as well do this. If we don't do it then we might run into trouble later.)

Now the randomness is that we pick each vertex $v_i \in V_i$ uniformly and independently for each i. This is an instance of the random variable model — these guys are the random variables, and we would like to design a set of bad events so that avoiding all these bad events implies that the v_i's together form an independent set.

The point to explain here is that it should not be obvious what we want to pick as the bad events — we may have to try a few different attempts before something works. It turns out that there is a correct answer, as we will see soon. But first let's explore several possibilities, including some incorrect ones, to see that some choices of bad events work better than others.

What are some things we coudl choose as bad events, so that avoiding all the bad events implies an independent set?

- A_{ij} is the event that there is an edge between v_i and v_j, for all i and j such that there is at least one edge between V_i and V_j. (Otherwise we don'tn neven need to consider it.)
- For each edge, let A_e be the event that both endpoints of e are selected, ranging over $e \in E$.

Both are valid choices of bad events, in teh sense that avoiding the bad events implies an independent set. So they are both worth trying, and we will try both of them; but it turns out htat exactly one will be good in terms of applygin LLL. Unless you can see very far ahead, that's not easy to tell from teh beginning; ont eh pset there will be many examples where you will have to try some different choices.

Call these (1) and (2).

In (1), we can check that $\mathbb{P}(A_{ij})$ — if we pick v_i, then there are potentially Δ edges going to v_j, and we are picking v_j over here, so we have an upper bound of Δ/k. The canonical dependency graph has A_{ij} adjacent to $A_{i'j'}$ if and only if $\{i, j\}$ shares elements with $\{i', j'\}$. Then we can get a maximum degree on this dependency graph — fix i and j. Then we see that if we want to overlap with i we should think about how many other i''s or j's we can go from here. Starting form i, each choice leads to Δ many stuff, so we get an upper bound of $2k\Delta$.

So then we have $p = \Delta/k$ and $d = 2k\Delta$, which means $ep(d+1)$ is too large — it is $e \cdot 2\Delta^2$.

Now let's turn to the other design. What is $\mathbb{P}(A_e)$, so we are fixing an edge e and trying to find the chance that both endpoints of e get chosen. That is $1/k^2$. In the dependency graph coming from the RVM, $A_e \sim A_f$

if e and f share some V_i's (i.e. they are sticking out of the same blob — if there exists some V_i that intersects both e and f).

Now we want to look at e and find the number of edges we can emit from V_i or V_j. We pick one of the sides, then one of the vertices, and then an emitting edge; this gives $2k\Delta$. So we now have $p = 1/k^2$ and $d = 2k\Delta$, and we can check that indeed $ep(d+1) \le 1$. So this works out — we have

$$e \cdot \frac{1}{k^2} \cdot 2k\Delta \le 1$$

if k is a sufficiently large multiple of Δ.

The moral here is that it's not always clear from teh beginning what we shoul dpick as the bad events — there are some choices that are better than others, and it may take a few tries to figure out what works.

§14 October 24, 2022 — Lovasz Local Lemma

§14.1 Directed Cycles

First, we'll use the Lovász Local Lemma to find directed cycles of length divisible by k.

> **Definition 14.1.** A directed graph is k-regular if every vertex has k edges in and k edges out — in other words, $\operatorname{indeg}(v) = \operatorname{outdeg}(v) = k$ for all v.

We will prove the following result:

> **Theorem 14.2** (Alon–Liniel)
>
> For every k, there exists d such that every d-regular directed graph has a directed cycle of length divisible by k.

So as long as we have some local conditions — being d-regular — we can find some potentially large structure.

This also gives a result about *undirected* graphs:

> **Corollary 14.3**
>
> For every k, there exists d such that every $2d$-regular (undirected) graph has a cycle of length divisible by k.

First let's prove the corollary from the theorem:

Proof. We'd like to convert a $2d$-regular graph into a d-regular directed graph. So at every vertex, we want to choose some orientation so that there's d edges in and d edges out; we'd like to do this consistently, so that every vertex has d edges in and out.

One way to do this is to find an Eulerian cycle — every connected graph where all vertices have even degree has an Eulerian circuit. This means every $2d$-regular graph has an Eulerian tour on each of its connected components (going through every edge once and returning to the vertex where we started), and this tour gives us an orientation of the edges (by following the direction of the tour). In particular, now every vertex has an equal number of edges in and out. This produces a d-regular digraph, and applying the previous theorem gives us the corollary. □

The real content we'll discuss is how to prove the theorem. Looking at the problem statement, it's really unclear what we should use as our source of randomness. What we'll do is the following: we're going to assign every vertex an element of $\mathbb{Z}/k\mathbb{Z}$, and only look at the edges $i \to i+1$. If we can keep following such edges around, then any cycle must visit labels 1, 2, 3, 4, …, k (possibly multiple times), and therefore must have a multiple of k number of vertices.

We'll prove a slightly more general statement:

Theorem 14.4

Every digraph with minimum out-degree δ and maximum in-degree Δ contains a cycle of length divisible by k, as long as

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

In particular, for a given k, when we take δ and Δ to be d, as long as d is large enough we have the desired conclusion.

Proof. First we'll do a bit of cleaning that doesn't cost us at all, but will be important later on — by deleting some edges, we can assume that every vertex has out-degree *exactly* δ. (Initially we assumed that every vertex has out-degree at least δ, but if it's strictly greater we can simply delete the extraneous edges; this does not violate the 'maximum degree' condition).

Now assign every vertex v an element $x_v \in \mathbb{Z}/k\mathbb{Z}$, independently and uniformly at random. Now we'll look for cycles where we're only using edges of the form $i \to i+1$ for some i, so that any directed cycle will necessarily have length divisible by k.

We can start with some vertex, and keep tracing — we keep looking for an out-edge, and if we eventually loop back to complete a cycle, then we're done. So we're worried about vertices which end up without an emitting edge of this form — since we might get stuck at such vertices. If we don't have any vertices with this issue, then we're done — we can start at an arbitrary vertex and follow our edges until we complete a cycle, which will have the desired property.

With that in mind, we can consider some bad events to encapsulate this condition. As in last time, there are some design choices as to what we take as bad events. For each vertex v, we can let A_v be the event that there's nowhere to go from v, i.e., that there does not exist an out-neighbor of v with label $x_v + 1$.

This is an example of the random variable model — the independent random variables are the x_v's, and each event only involves v and the labels of its out-neighbors. So now we have a canonical dependency graph, and we'd like to understand its maximum degree.

In the canonical dependency graph, A_v depends on teh variables corresponding to $\{v\} \cup N^+(v)$, where $N^+(v)$ denotes the set of out-neighbors of v. So the canonical dependency graph has $A_v \sim A_w$ if and only if $\{v\} \cup N^+(v)$ intersects $\{w\} \cup N^+(w)$. Then given a fixed v, we'd like to estimate the number of w's which could interact with v.

Suppose we fix v. Then there's a few ways that the closed neighborhood of w can intersect that of v:

(a) w could be an out-neighbor of v;

(b) v and w could share an out-neighbor;

(c) v could be an out-neighbor of w.

These are the only possibilities — so the dependency graph only has these three types of edges.

Now we can estimate the maximum degree. Starting at v, the number of choices of type (c) is Δ, the number of type (a) is δ, and the number of type (b) is $\delta(\Delta-1)$ (we can walk out and return, but we don't want to return to the same vertex). So then the maximum degree is at most $\Delta + \delta\Delta$. (This is why we needed to delete edges in the beginning.)

Now let's estimate the probability $\mathbb{P}(A_v)$. A bad event occurs if none of the neighbors of v are of the label we're looking for, so

$$\mathbb{P}(A_v) = \left(1 - \frac{1}{k}\right)^{\operatorname{outdeg}(v)} \le \left(1 - \frac{1}{k}\right)^{\delta}.$$

As long as

$$e\left(1 - \frac{1}{k}\right)^{\delta}(1 + \Delta + \delta\Delta) \le 1,$$

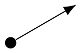
we can avoid all the bad events by the Lovász Local Lemma. Using the bound $(1 - 1/k)^\delta \le e^{-\delta/k}$, we get that we can avoid all the bad events if

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

If we just wanted to prove the original statement, this would be good enough; but the theorem we wrote is a bit stronger — in the theorem, we don't have the extra Δ.

There's a small subtlety that allows us to get rid of Δ. So far, we've only been considering the random variable model; but the general form of the Lovász Local Lemma doesn't require the random variable model, and works for any dependency graph. We may actually have a smaller dependency graph than the one given by the canonical one, and that applies here:

> **Claim —** A_v is independent of all A_w where $N^+(v)$ is disjoint from $N^+(w) \cup \{w\}$.



The reason for this is that even if we condition on some events involving vertices leading into v (shown in yellow), as long as we don't intersect the out-neighborhood of v, it's still true that

$$\mathbb{P}(A_v \mid *) = \left(1 - \frac{1}{k}\right)^{\operatorname{outdeg}(v)}.$$

This is because even if we put a value on v, the conditional probability is still the same.

We can use this to get a dependency graph with even fewer edges, and that observation lets us remove Δ. In the grand scheme of things, this isn't important; but it serves to illustrate that it can be good to return to the original theorem statement and think about what the independence condition really means — there are times when we can get a little more out of it by moving away from the random variable model. □

§14.2 Lopsided Local Lemma

In the second part of the lecture, we'll *completely* move away from the random variable model — we'll look at an extension of the Lovász Local Lemma where instead of asking that a lot of events are *independent*, we want *positive correlation*.

Suppose we want to avoid a collection of bad events, and the events are correlated, but in a way that avoiding A_1 and A_2 only makes it *easier* to avoid A_3 (rather than A_1 and A_2 being independent of A_3). Intuitively, this should make it even *easier* to avoid all the bad events — if avoiding some bad events makes it easier to avoid another, then this only helps us (it shouldn't hurt us to not have independence in this direction).

To make this precise, we can look at the proof of the local lemma, and check where we used the hypothesis of independence. There was a specific step in the proof where we argued that

$$\text{numerator} \le \mathbb{P}(A_i \mid \bigwedge_{j\in S_2} \overline{A_j}),$$

where all A_j for $j \in S_2$ are non-neighbors of A_i. Then using the definition of the dependency graph, we had that this expression was exactly $\mathbb{P}(A_i)$.

But if instead of equality here, we had a $\le$, this would still be fine — the proof would still work. So we'd be okay if instead of requiring independence, we had an *inequality* in conditional probabilities.

This allows us to revise the local lemma to a slightly more general statement, with exactly the same proof (other than this small modification):

Theorem 14.5 (Lopsided Local Lemma)

Let $A_1, \ldots, A_n$ be events such that for each i, there is a subset $N(i) \subset [t]$ such that for all $S \subseteq [n] \setminus N(i)$ (i.e., where S consists of non-neighbors of i), we have

$$\mathbb{P}(A_i \mid \bigwedge_{j\in S} \overline{A_j}) \le \mathbb{P}(A_i).$$

Also suppose there exist $x_1, \ldots, x_n$ in $(0,1)$ such that $\mathbb{P}(A_i) \le x_i \prod_{j\in N(i)}(1-x_j)$ for all i. Then

$$\mathbb{P}(\text{none of the } A_i \text{ occur}) \ge \prod_i (1-x_i).$$

Everything is exactly the same, except that we changed the equality to a $\le$ sign; and the proof is exactly the same as well.

As earlier, we also have a symmetric version:

Corollary 14.6 (Symmetric Version)

If $|N(i)| \le d$ and $\mathbb{P}(A_i) \le p$ for all i, and $ep(d+1) \le 1$, then $\mathbb{P}(\text{none of the } A_i\text{'s occur})$ is positive.

§14.3 Random Injection Model

As before, we have a directed graph structure. Instead of calling this a dependency graph, it's customary to call it a **negative dependency graph** — because intuitively, the edges record negative correlations (although this isn't the definition — the definition is as stated in the theorem).

This seems a bit annoying to check — we have to design these events, and then check how to construct a graph such that this condition holds. In the usual Lovasz Local Lemma, we generally used the random variable model, which gave us an easy-to-construct canonical dependency graph (where edges are simply events that share variables).

Now we'll set up another model that applies to a lot of interesting applications where we can apply the lopsided local lemma, and where the canonical dependency graph is also fairly easy to construct.

Suppose we have two finite sets X and Y, with $|X| \le |Y|$. Then suppose we have an *injection* $f\colon X \to Y$ chosen uniformly at random.

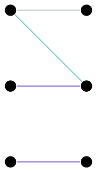
It'll be helpful to regard this injection as a graph — we can draw the elements of X and the elements of Y, and represent the injection as a matching — so we have a matching M that's complete from X to Y. (Here f, and therefore M, are chosen uniformly at random.)

Given any matching F (not necessarily complete) in $K_{X,Y}$, let A_F be the event that $F \subseteq M$ — so someone gives us a partial event, and the bad events are that our random matching contains F.

Let F_1, F_2, …, F_n be matchings (not necessarily complete) in $K_{X,Y}$. The *canonical negative dependency graph* for the events A_{F_1}, …, A_{F_n} has:

- One vertex for each event;
- For two distinct events F_i and F_j, we have an edge $A_{F_i} \sim A_{F_j}$ if and only if F_i and F_j are *not* vertex-disjoint.

As an example, suppose we have three matchings:



Here if 1 is the matching in red, 2 the matching in blue, and 3 the matching in green, then 1 and 2 don't share any vertices, but 1 and 3 do and 2 and 3 do, so our canonical dependency graph has edges between 1 and 3, and between 2 and 3.

Here we're still choosing an assignment for every element of X, but unlike the random variable model, the assignments are not independent.

Now we need to show that this is a valid negative dependency graph — i.e., it satisfies the correlation inequality from the lopsided local lemma. The proof comes down to the following claim:

Theorem 14.7

Let F_0 be a matching in $K_{X,Y}$ such that F_0 is vertex-disjoint from $F_1 \cup \cdots \cup F_k$. Then

$$\mathbb{P}(A_{F_0} \mid \overline{A_{F_1}} \cdots \overline{A_{F_k}}) \le \mathbb{P}(A_{F_0}).$$

Once we prove this claim, it verifies that our graph is a valid negative dependency graph.

Proof. Suppose we have some matching F_0. Let X_0 and Y_0 be the m vertices of F_0.

Then for each matching T, let $\mathcal{M}_T$ be the set of complete matchings (i.e. injections $X \hookleftarrow\!\!\!\to Y$) containing T, but not containing any of F_1, …, F_k.

We can write our conditional probability as a quotient of two cardinalities — our desired conditional probability is

$$\frac{|\mathcal{M}_{F_0}|}{|\mathcal{M}_\emptyset|}.$$

We can rewrite the denominator by summing over all partial matchings $T\colon X_0 \hookleftarrow\!\!\!\to Y$ — we break up our information based on where X_0 is sent to.

Meanwhile, the right-hand side is the probabiltiy that the uniform random matching contains F_0. This is simply

$$\frac{1}{\#\{T\colon X_0 \hookleftarrow\!\!\!\to Y\}},$$

since all injections $X_0 \hookleftarrow\!\!\!\to Y$ have equal probability of happening.

So it suffices to show that

$$|\mathcal{M}_{F_0}| \le |\mathcal{M}_T|$$

for all $T\colon X_0 \hookleftarrow\hookrightarrow Y$. We'll prove this by constructing an injection $\mathcal{M}_{F_0} \hookleftarrow\hookrightarrow \mathcal{M}_T$.

Suppose we start with some F_0, and T (both of which match X_0 to some set). Call the right-hand sides of the matchings Y_0 and Y_1.

Then $\mathcal{M}_{F_0}$ is the set of all matchings containing our yellow edges and some other edges, but avoiding some edges that we wish to avoid. We'd like to transform this into a matching with the purple edges, and some additional emanating edges.

To do this, intuitively we should try to move these edges to where we want them to be. Mathematically, this can be described via a permutation on the right-hand side — fix a permutation σ (depending on F_0 and T) such that:

- σ fixes all elements of Y outside of $Y_0 \cup Y_1$ — we don't want to change anything that we don't need to change.
- σ sends F_0 to T.

In other words, applying our permutation allows us to send the yellow vertices to the purple ones.

There are some choices here — there could be more than one valid permutation, and it'll also specify some other things (for example, a purple but not yellow vertex is sent to oen fo the empty vertices).

Then σ induces a permutation on the set of complete matchings $X \hookleftarrow\hookrightarrow Y$.

By construction, we know that this sends $F_0 \to T$. But now we want to check the conditions about avoiding $F_1, \ldots, F_k$.

We have a bunch of edges that don't involve any vertices inside X_0 and Y_0, and we need to not contain any of these $F_1, \ldots, F_k$. So we want to check that if a complete matching avoids $F_1, \ldots, F_k$, then applying σ, we still avoid them.

This is because there's no room to introduce problems — anything that could happen is happening in places that aren't involved.

This proves the theorem. □

So now we have a canonical negative dependency graph. Once we have this, everything else should be the same as before.

§14.4 A Simple Application

Corollary 14.8

The probability that a uniform permutation $[n] \to [n]$ has no fixed points is at least $(1 - 1/n)^n$.

Proof. We already have a random permutation $[n] \to [n]$. For each i, we have a bad event A_i that $f_i = i$, which in our language corresponds to a horizontal edge in our graph (so we want to avoid random matchings containing that edge). Then the canonical negative dependency graph is checked by checking vertex overlaps. This means it's empty — because we have no vertex overlaps. So we can apply the lopsided local lemma with $x_i = 1 - 1/n$ (the probability of the event A_i) to get that this probability is at least $(1 - 1/n)^n$ — i.e. the product of probabilities of avoiding individual bad events.

So the probability we have no fixed points is at least the product of probabilities of 1 not being a fixed point, 2 not being a fixed point, and so on. □

It turns out that this is actually pretty good — both the actual formula (from inclusion-inclusion) and this lower bound converge to $1/e$.

§14.5 Latin Transversals

Definition 14.9. A **Latin square** is a $n \times n$ array with n symbols, such that every row and column has exactly one symbol of each type.

Example 14.10

One Latin square of order 3 is

1	2	3
3	1	2
2	3	1

.

Remark 14.11. These are called Latin squares because Euler studied them, and used the Latin alphabet to fill the squares.

Question 14.12. Do these Latin squares contain transversals?

Definition 14.13. In any $n \times n$ array, a **transversal** is a set of n entries, one of each row and column. A **Latin transversal** is a transversal with all distinct entries.

Example 14.14

In our example Latin square, one Latin transversal would be

1	2	3
3	1	2
2	3	1

.

A famous open problem is whether we always *have* a Latin transversal.

Conjecture 14.15 (Ryser's Conjecture, 1967) — Every odd Latin square has a transversal.

This is false for the case of an even Latin square, but there's an extension:

Conjecture 14.16 (Ryser–Brualdi–Stein Conjecture) — If n is even, then every order n Latin square contains a transversal with all but one symbol.

There is a recent related result — if instead of avoiding one symbol we're allowed to avoid $O(\log n/\log\log n)$ symbols, then the result is true. This improves on an earlier result that got this for $(\log n)^2$.

Erdős and Spencer first introduced the lopsided lemma to prove the following:

Theorem 14.17 (Erdős–Spencer 1991)

Every $n \times n$ array where every symbol appears at most $n/4e$ times necessarily has a Latin transversal.

Note that these are not Latin squares — they're objects which give you more flexibility, in that every symbol only appears $n/4e$ times.

In hindsight, this is a perfect problem for an application of the random injection model.

Proof. Pick a transversal uniformly at random — this is equivalent to picking a permutation from the rows to the columns, or equivalently a perfect matching between rows and columns.

We're trying to obtain a Latin transversal; the only requirement for a transversal to be Latin is that it has no repeat entries.

So for each pair of equal entries in the array, not both lying in the same column or row, we can create a bad event that our random transversal contains both entries.

This corresponds to some event described by two edges, in the random injection model.

Suppose we have a pair of equal entries. Then to find an event with an edge in the negative dependency graph, we need another pair sharing some row or column.

To bound this, we can first choose some entry in one of the four relevant rows or columns; this gives at most $4n$ choices. Then we need to pick another entry equal to this orange entry; that can be done in at most $n/4e - 1$ ways (by the hypothesis). This gives that the maximum degree is at most $4n(n/4e - 1)$.

Meanwhile, the probability that a bad event occurs is $p = 1/n(n-1)$. We can check that

$$ep(d+1) = e \cdot \frac{1}{n(n-1)} \cdot \left(4n\left(\frac{n}{4e} - 1\right)\right) \leq 1,$$

so we can apply the lopsided local lemma and conclude that with positive probability none of the bad events occur, and therefore there exists a Latin transversal. □

§15 October 26, 2022

§15.1 Algorithmic Local Lemma

The local lemma tells us that under some circumstances, we can avoid a collection of bad events. A natural question, especially from computer science, is whether we can do this computationally — whether we can find an instance that avoids all the bad events (rather than just asserting mathematically that such an instance exists).

Question 15.1. Is there some way to find a solution avoiding all the bad events?

In the most general setting, this is too hard — we can encode a problem believed to be computationally difficult (the discrete logarithm problem) in the language of the local lemma.

Example 15.2

Let $q = 2^k$, and $f\colon [q] \to [q]$ some fixed bijection. Our goal is to computationally invert f — given y, we want to find x such that $f(x) = y$.

If $x \in [q]$ is chosen uniformly at random, then $f(x)$ is also chosen uniformly at random (its distribution is also uniform, since we have a bijection). So for each $i \in [k]$, we can let A_i be the bad event that $f(x)_i \neq y_i$ (i.e. we're guessing x at random, and the bad events are that we are wrong on the ith coordinate).

Then we can check that the events $A_1, \ldots, A_k$ are all independent. So the Lovász Local Lemma guarantees the existence of x (i.e. we can avoid all the bad events — of course, we already knew that, since we have a bijection). But computationally, it's believed that there are some hard functions to invert — unless P = NP (which we don't believe is the case), it's believed that there are some *one-way functions* that are easy to compute but hard to invert. An example of such a function is the discrete log function — we can pick a generator from the multiplicative group, and the function takes the exponent. (Such functions are the bedrock of cryptograpy — many cryptographic algorithms depend on the believed difficulty of such inversion.)

So the most general formulation of the Lovász Local Lemma cannot be made algorithmic. But perhaps if we restrict the lemma to some specific settings, such as the random variable model, then we can say more. (The situation here is *not* an instance of the random variable model.)

For a long time, this was a major open problem — if we're given a problem where the Lovász Local Lemma guarantees a solution exists, can we find it algorithmically (via a polynomial time algorithm)? Moser in 2009 found the first algorithm to do so; this work was subsequently extended and improved by Moser–Tardos 2010 (in the random variable model, if we have existence then we have an algorithm).

Here we are in the random variable model.

Question 15.3. We would like an algorithm to do the following:

Input: A set of variables, and a set of bad events.

Output: An assignment of variables avoiding all bad events.

Many algorithmic tasks are quite complicated. But it turns out that there is a very simple algorithm here:

Algorithm 15.4 (Moser–Tardos 2010) — First, initialize all variables arbitrarily.

Then while there exists some violated event, we simply resample all the variables that this event depends on (uniformly at random).

The initial assumption may not be satisfactory — there may be some bad events that occur (we would perhaps expect this — if we assign uniformly at random, we are usually going to expect bad events to come up, since the probability of success in the local lemma is exponentially small). But we can simply try to fix the violated bad events by resampling — if we see a bad event, we pick one and fix it. And if we keep going, hopefully we're eventually done. For this reason, this is called the 'fix-it algorithm.'

This is so simple that you'd never think it works — but it turns out that it does.

Theorem 15.5

Let $A_1, \ldots, A_n$ be the bad events. Suppose that there exist $x_1, \ldots, x_n$ in $(0,1)$ such that $\mathbb{P}(A_i) \leq x_i \prod_{j \in N(i)}(1 - x_j)$ for all $i \in [n]$. Then for every $i \in [n]$, the expected number of times that A_i is resampled is at most $x_i/(1 - x_i)$.

In particular, the total number of times that any event is resampled is polynomially bounded — so this algorithm shouldn't take very long.

Remark 15.6. This is a randomized *Las Vegas* algorithm. We may have heard of a more familiar type of randomized algorithms, called *Monte Carlo* algorithms — for example, we could estimate some quantity by sampling until we have enough samples to do our estimate.

A *Monte Carlo* algorithm is one where we run in a bounded amount of time, and we have some (hopefully large) success probability. Meanwhile, a *Las Vegas* algorithm always succeeds, but it might run for a long time (perhaps an infinite amount of time) — but for a good algorithm, we want some bound on the *expected* runtime.

If we have a Las Vegas algorithm where the expected runtime is polynomial in the parameters of the problem, then we can convert it into a Monte Carlo algorithm by cutting it off at some point — if we have expected runtime A, we can terminate after time $100A$ (and declare failure). This then gives us a Monte Carlo algorithm — it runs in time at most $100A$, and it succeeds with probability at least 0.99 by Markov's inequality.

In general, there may not be a way to convert a Monte Carlo algorithm to a Las Vegas algorithm, because even though we succeed with some probability, this might not be checkable — for example, finding a 2-edge-coloring in K_n without a monochromatic clique of size $100 \log_2 n$. This can be done — we can do it randomly, and we succeed with exponentially high probability. But we have no way of checking whether we've succeeded, so there's no way to convert this into a Las Vegas algorithm.

This is really amazing — the algorithm gives something that's simple enough we can actually implement it on a computer (unlike many algorithms). And the paper is extremely beautifully written — Prof. Zhao highly recommends we read it.

We won't prove the general form, but we'll prove a special case: Moser's original theorem involved the k-SAT problem:

Question 15.7. We have some variables, and some *clauses* where each clause is an *or* of some variables (and their complements), for example,

$$(x_1 \vee x_2 \vee \overline{x_3}) \wedge (x_2 \vee \overline{x_4} \vee x_5) \wedge \cdots.$$

Here each clause has k variables. We want to find a satisfying assignment.

The Lovász Local Lemma tells us that if we sample randomly, then if each clause shares variables with at most d clauses, then as long as $ep(d+1) \leq 1$ there exists a solution. Here each clause fails with probability $p = 2^{-k}$ (for its k variables, there's one way to fail), so the Lovász Local Lemma works if $e \cdot 2^{-k}(d+1) \leq 1$.

Algorithm 15.8 (Moser's Fix-It Algorithm) — The input is a k-CNF formula, and the output is a satisfying assignment.

We first initialize the variables arbitrarily. Then while there exists a violated clause C, we run a sub-routine $\textsc{fix}(C)$.

The subroutine $\textsc{fix}(C)$ does the following:

- Resample all the variables in C uniformly at random.
- While there exists some violated clause D that shares a variable with C, we recursively call $\textsc{fix}(D)$.

Note that we are allowed to have $D = C$.

This is similar to what we just saw, but we're separating the recursive fix-it into an outer loop, and an inner recursive function that calls itself whenever it sees some violated clause.

Theorem 15.9

In a k-CNF where every clause shares variables with at most 2^{k-3} other clauses, Moser's algorithm succeeds in expected polynomial time (in the number of variables and clauses).

This condition is quite similar to what's guaranteed by the local lemma — the local lemma tells us that if we replace 2^{k-8} with $2^k/e$, then there *exists* a satisfying assignment. And Moser's algorithm tells us that if we relax the constant a bit, then we can *find* it in polynomial time.

We'll prove this in two parts.

Lemma 15.10

Each clause appears at most once in the outer loop.

So if our outer loop calls C_1, then C_1 can never come up again.

Proof. When we call $\textsc{fix}(C)$, every clause that was previously satisfied *remains* satisfied after $\textsc{fix}(C)$ completes. The fixing might destroy some clauses that were previously satisfied, but because of all the recursion, when we *finish*, anything that was previously satisfied must remain satisfied.

To see this, suppose we have some clause that's initially violated; and we resample, and that might end up violating the second one. But then the function call is going to fix whatever problems we've messed up. The nature of the recursion is that if in the process of calling FIX we've messed up some other clause, then the routine is going to fix that — so we're not going to create any new violations.

Now the lemma is clear — when we call $\textsc{fix}(C)$ it becomes fixed, and then in the outer loop, it won't ever become violated again. □

Lemma 15.11

Fix a k-CNF on n variables, where each clause shares variables with at most 2^{k-3} other clauses. Also fix a clause C_0, and some assignment of variables (not necessarily satisfying). Then in an execution of $\textsc{fix}(C_0)$, for every integer ℓ,

$$\mathbb{P}(\geq \ell \text{ recursive calls to } \textsc{fix}) \leq 2^{-\ell+n+1}$$

(Here we do not count the initial call).

In particular, this means we shouldn't have much more than n recursive calls — in expectation, there are $n + O(1)$ recursive calls, and therefore in total there are $nm + O(m)$ expected calls to FIX.

Proof. We first formalize our randomness in the following sense — instead of generating k random bits each time, we first set some $x \in \{0,1\}^{k\ell}$ (chosen uniformly at random), and this is all the randomness that we have — every time we sample, we read the next k bits of x. This way, we can keep track of what randomness is involved, which will be important in the argument. Furthermore, suppose that FIX is recursively called at the ℓth time; then we halt the algorithm before this call, and declare that it has failed. (This means we never run more than ℓ resamplings — this is important because eventually we'd run out of random bits.)

We're going to keep track of how the algorithm runs by keeping track of an *execution trace* — we have some fixed CNF, and for illustrative purposes we can write down a few clauses involving some variables.

Then the execution trace keeps a log: for example, this may look like:

```
        fix(C7) called
        fix(C4) called
        fix(C7) called
        while loop ended
        fix(C2) called
        while loop ended

```

This tells us which clause got resampled, and in what order. We don't keep track of what we're using for the resampling — that'll come in later on (we already know this, because we know what x is — so if we see the execution trace, then we can by hand follow everything and figure out what happened to each of our variables). This can also be thought of as a stack — a new call adds something to our stack, and a 'while loop ended' pops it.

We'd like to understand the information content of this string, and we'll do so by converting this trace to a bit string. There's an obvious way to do this — seeing FIX(C_7) called and converting 7. But this could be quite expensive, since there's a lot of clauses.

But we know that every clause intersects at most 2^{k-3} other clauses. So we can instead record *relative* position — when we call FIX(C), there's only at most 2^{k-3} strings that can be called inside. So we can represent a call to FIX as a 0-bit along with (exactly) $k-3$ bits representing the relative position of our clause.

Then when the while loop ends, we place a single bit 1.

This lets us convert this execution trace to a bit string, and if we read the string, then we can recover the log — every sentence begins with 0 or 1, so there's no ambiguity as to which type it is.

At this point, we can see that right before the ℓth call, we've already used up all our random bits (since we've resampled ℓ times, and therefore exhausted x). The key claim is the following:

> **Claim** — Right before the ℓth recursive callto FIX, we can *completely recover* x from the execution trace and the final assignment of variables.

Proof. The execution trace tells us a history of which clause got resampled when, and in what order — but it doesn't tell us *how* it was resampled (which values were put in). But we'll try to recover this information by rewinding history.

Note that if FIX(D) is called, then D must have been previously violated before the call. This means there's a *unique possibility* for how the variables in D were assigned.

So let's say that D was the last clause we resampled. Then we put the new values of D back into the end of x. Then rewinding, we know *exactly* what we should put in D from the previous step — we put the variable assignments in the one way that violates D.

So for example, in the execution trace, if we know that $x_2 \vee \overline{x_4} \vee x_5$ was the last clause called, then we know that before it was called we must have had $x_2 = 0$, $x_4 = 1$, and $x_5 = 0$; so we revert D back to those values. And meanwhile we see the final assignment of variables, so we remove those from D and build those as the last elements of x. ■

Now how long can the execution trace be? For each of the ℓ times something was called, we gain at most $1 + (k-3) + 1 = k - 1$ bits, so then the trace has length at most $\ell(k-1)$. In particular, the number of possibilities for the trace is at most $2^{\ell(k-1)+1}$ (where the extra 1 is because the execution trace may not take the full length).

Now the key claim means that each x that leads to a failed execution can be completely recovered from this data. So the number of strings that lead to a failed execution is *at most* the number of possibilities for the execution trace, times the number of possibilities for the final assignment — so

$$\#\{x \in \{0,1\}^k \text{ that lead to } \textsc{fail}\} \leq 2^{\ell(k-1)+1}2^n.$$

But the left-hand side is exactly $\mathbb{P}(\textsc{fail}) \cdot 2^{k\ell}$, and rearranging gives that $\mathbb{P}(\textsc{fail}) \leq 2^{-\ell+n+1}$. □

Remark 15.12. The more general Moser–Tardos theorem gives exactly the same guarantee as in the Lovász Local Lemma. The idea is very similar, but instead of a trace, they keep a *tree* — our recursive sampling has a tree structure (we can think of a parent call and its children call), and they keep track of the tree structure and analyze it more carefully. This allows them to get the right constant, and a more general statement.

Student Question. *Why do we need to record 'while loop ended'?*

Answer. We need to keep track of which line we're on — whether we've gone back to the previous execution (whether our recursive call has gone back to its parent call). If we didn't have the while loop ended, then we would have ambiguity.

Remark 15.13. This work was really influential, and inspired a lot of further research for other variants of the random variable model (such as random permutations). There are other interesting results there, using similar ideas — once you have this idea, it's not so surprising you can do other things.

§15.1.1 Some Intuition

The history of this work is kind of funny. Before Moser, no one knew how to do much — there were some partial results, but nowhere close to this (previous results were of a different order of magnitude). When Moser came up with this, it was a big shock.

The original paper was much more complicated than the proof we saw. Then in a talk, he gave essentially this beautiful argument; this was never recorded, but it was subsequently popularized (by various blogs, including Terry Tao's). Terry Tao gave this the term **entropy compression**.

We'll discuss entropy later in the course, but intuitively, it's an amount of information.

Here this refers to the following: at each iteration, we are drawing k bits of randomness, and outputting $k-1$ bits to the trace. And this process is recoverable. So it *losslessly* compresses information from k random bits into $k-1$ bits of information. That is not possible — you cannot compress k bits into $k-1$ bits. Of course there is a bit of overhead — the final variable assignment — but that is a constant. So each time we're compressing this randomness, it cannot go on forever — it can't go on for very long before we violate some principle of information theory. (The fact that this is completely recoverable is from the arguments we saw that we can rewind through the execution trace to get back our source of randomness.)

So despite all the underlying combinatorics, the underlying thing that makes it work is this — that we're drawing k random bits, and we can keep track of what happened using $k-1$ random bits — and this lossless compression cannot go on forever.

Remark 15.14. So far, we've seen two different proofs of the Lovász Local Lemma — the original proof (the clever and mysterious-looking manipulation of conditional probability inequalities), and a completely different proof that's also really clever. Are these somehow the same proof, or are they completely different?

They *look* nothing like each other. But understanding whether there's some way the two proofs are related toe ach other is an important question — often with a deep theorem, it's worth trying to understand the relationships between these proofs. And in this case, that's really mysterious.

§16 October 31, 2022 — Correlation Inequalities

The main theorem of today, the *Harris inequality*, gives us a simple situation where we can say that events are positively correlated. Informally, the main theorem tells us the following: increasing events of independent boolean random variables are positively correlated.

Now we'll state the actual theorem.

Definition 16.1. Suppose $A \subseteq \{0,1\}^n$ is a subset of the boolean cube. We say that A is a **up-set** if A is upwards closed — in other words, if $x \in A$ and $y \geq x$ coordinate-wise, then $y \in A$ as well.

Up-sets may also be called **increasing event**, or **increasing property** — we've seen these names before in the discussion on thresholds, where we saw that every increasing event has a threshold.

Theorem 16.2 (Harris Inequality)
If A and B are increasing events of *independent* Boolean random variables, then

$$\mathbb{P}(AB) \geq \mathbb{P}(A)\mathbb{P}(B).$$

So as long as we're looking at increasing events of independent boolean random variables, these events are positively correlated.

In many applications, our boolean variables will be identically distributed, but they don't *have* to be — we don't need each to have the same probability.

Remark 16.3. The theorem can also be stated in terms of conditional probability, as $\mathbb{P}(A \mid B) \geq \mathbb{P}(A)$.

Example 16.4
Suppose that we're in $G(n,p)$, where every edge is chosen independently with probability p; and suppose we'd like to understand the probability that $G(n,p)$ contains a Hamilton cycle. Suppose we want to compare the probabilities that our graph has a Hamiltonian cycle given that it has average degree at least 2. Then Harris inequality tells us

$$\mathbb{P}(\text{Hamiltonian} \mid \text{average degree} \geq 2) \geq \mathbb{P}(\text{Hamiltonian}),$$

since both events are increasing. This matches our intuition — we're conditioning on something that only helps us.

We can insert any increasing events here, many of which for this would be difficult to prove directly.

Student Question. *What are the boolean random variables?*

Answer. There is one for each edge; so we have $\binom{n}{2}$ independent Boolean random variables.

This is really a prototypical example.

Student Question. *Can we relax the requirement of independence to assuming the Boolean variables are also positively correlated?*

Answer. There are situations where you'd want to relax independence.

The Harris inequality was initially found in the context of the study of *percolations*, and more generally in mathematical problems in statistical physics. An important problem is that if we start with a grid, and we randomly keep some of the edges but not others – we keep every edge with probability p (so it's like $G(n, p)$, but we start with an infinite grid instead of the complete graph), then we get some subgrid (which gives a beautiful picture). The basic question is, for which p do we have an infinite component?

These kinds of problems are called percolation; there is a whole area of mathematics called *percolation theory* related to this. For the most part, we don't know the answer to these problems, but this is one of the special cases where we do.

If we plot p versus the probability that 0 (the origin) belongs to an infinite component, then it turns out there's a threshold at $1/2$ — the probability is 0 up to $1/2$, and then increases slowly after that. This is a deep result; the easier direction, that percolation doesn't occur at smaller probabilities, was shown by Harris — and this is where the inequality was introduced. The opposite direction was an important and famous result of Keshton.

This is basically the only case in percolation theory where we know the exact answer — such results are extremely rare, and it's a subject of much ongoing research. (For example, we have no idea what's going on in three dimensions.)

People working in this area have also used other models, that surprisingly have connections to percolation. For example, in statistical phisics you may hear about the Ising model, and things coming from the Gibbs distribution and Potts model. These models do not correspond to independent random variables — they correspond to random variables with some inherent correlation. (The Ising model is some model of magnetism, and if you have two different sites with ± signs of magnets, they have some correlation.) There are extensions of the Harris inequality developed to handle such situations. More generally, there's an inequality called the FKG inequality, which is an extension of Harris's inequality for *distributed lattices*; that applies to important statistical mechanical settings that arise from the Ising model. (In this class, though, we will just focus on the independent case.)

Student Question. *Is it obvious whether the question of whether there's an infinite component and whether* 0 *is in the infinite component is related?*

Answer. In probability theory, there's a 0-1 law that states that the answer to the firtst question is always 0 or 1. Meanwhile, the answer to the second is some probability between 0 and 1 (you can always cut off 0); it turns out that there's an elementary argument that shows that the first is 1 if and only if the second is nonzero.

§16.1 Proof of Harris Inequality

We'll use induction, and we'll actually prove a slightly stronger version of Harris inequality:

Definition 16.5. Consider functions $f\colon \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n \to \mathbb{R}$, where each Ω_i is some linearly ordered set (such as $\{0,1\}$). We say that f is **monotone** (or *monotone increasing*) if whenever $x \le y$ coordinate-wise, then $f(x) \le f(y)$.

So an increase in some coordinate only weakly increases the function.

Student Question. *What does 'linearly ordered' mean?*

Answer. It just means that all the elements are ordered; this can also be called 'totally ordered.' (This is as opposed to 'partially ordered,' where only some pairs of elements are comparable.)

We will prove the following:

Theorem 16.6

If f and g are monotone increasing functions of independent random variables, then

$$\mathbb{E}[fg] \ge \mathbb{E}[f] \cdot \mathbb{E}[g].$$

This property is sometimes called *positive association.*

To see how we deduce the originally stated version, we can take $f = \mathbf{1}_A$ to be the indicator function of A, and $g = \mathbf{1}_B$.

Proof. We'll use induction on n, the number of variables.

In the case $n = 1$, we have a function of a single variable, and we need to show that if f and g are increasing functions, then this inequality holds.

To prove this, we have the inequalty

$$0 \le (f(x) - f(y))(g(x) - g(y))$$

for *all* x and y (in all the cases $x < y$, $x = y$, and $x > y$, f is similarly ordered to g, so the terms are both nonnegative or both nonpositive). Now taking the expectation over all x and y (which vary independently over Ω_1), we get

$$0 \le \mathbb{E}[(f(x) - f(y))(g(x) - g(y))].$$

Now we can expand using linearity of expectations. This gives us

$$\mathbb{E}[f(x)g(x)] + \mathbb{E}[f(y)g(y)] - \mathbb{E}[f(x)g(y)] - \mathbb{E}[f(y)g(x)] \ge 0.$$

We can forget about labels; this then gives us

$$0 \le 2\left(\mathbb{E}[fg] - \mathbb{E}f\mathbb{E}g\right).$$

(This is an instance of the rearrangement inequality.)

Now assume $n \ge 2$, and that we've proven this for all smaller n.

Now let $h = fg\colon \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$. Now we define **marginals** — functions f_1, g_1, and h_1. These are all functions of a *single* input variable, $\Omega_1 \to \mathbb{R}$. We define f_1 as

$$f_1(y_1) = \mathbb{E}[f \mid x_1 = y_1].$$

In other words, we take $x_1 = y_1$ and average out all the other coordinates, and see what happens — more explicitly,

$$f_1(y_1) = \mathbb{E}_{x_2,\ldots,x_n} f(y_1, x_2, \ldots, x_n),$$

where y_1 is non-random and the other variables are random. Likewise, we can define g_1 and h_1 in the same way, where $g_1 = \mathbb{E}[g \mid x_1 = y_1]$ and $h_1 = \mathbb{E}[h \mid x_1 = y_1]$.

Now we can see that f_1, g_1, and h_1 are one-variable functions. In particular, f_1 and g_1 are both monotone increasing one-variable functions — looking at our expression $\mathbb{E}f(y_1, x_2, \dots, x_n)$, if we increase y_1 this expression can never decrease.

Now for a fixed $y_1 \in \Omega_1$, note that the function $(x_1, \dots, x_n) \to f(y_1, x_2, \dots, x_n)$ is monotone increasing. (If we increase one of the coordinates, this value cannot decrease.) This is likewise true for g and for h. So applying the induction hypothesis for $n-1$, we find that

$$h_1(y_1) \ge f_1(y_1)g_1(y_1).$$

(This comes from fixing y_1, comparing $\mathbb{E}f(y_1, x_2, \dots, x_n)$, $\mathbb{E}g(y_1, x_2, \dots, x_n)$, and $\mathbb{E}h(y_1, x_2, \dots, x_n)$, and applying the induction hypothesis since we have one fewer variable.)

Now we let the first variable vary. By definition, we know

$$\mathbb{E}[fg] = \mathbb{E}[h] = \mathbb{E}h_1.$$

(Now h_1 is a function of one variable, but it already averaged out the remaining variables; so $\mathbb{E}h_1 = \mathbb{E}h$.)

Now applying our previous observation, we saw that $h_1 \ge f_1 g_1$ pointwise (i.e., for all y_1), so we have

$$\mathbb{E}[h_1] \ge \mathbb{E}[f_1 g_1].$$

But we can see that f_1 and g_1 are both monotone increasing. So applying the Harris inequality for one-variable functions, we get that $\mathbb{E}[f_1 g_1] \ge \mathbb{E}f_1 \mathbb{E}g_1$ as well. Finally, we know that $\mathbb{E}f_1 = \mathbb{E}f$, and $\mathbb{E}g_1 = \mathbb{E}g$. So then $\mathbb{E}[fg] \ge \mathbb{E}[f]\mathbb{E}[g]$, which finishes the proof. □

There's a few things to observe in this proof. We really did need to do the base case separately, because in this chain of inequalities, there were two applications of Harris inequality — one step from the $n-1$ case, and one from the 1 case. The other observation is that we did need the version that allows *functions*, rather than just $\{0,1\}$-valued functions — because when we took these marginals, even if f and g were $\{0,1\}$-valued functions, taking the marginals would give us real-valued functions.

Remark 16.7. This idea of doing induction variable-by-variable is a pretty important tool — it comes up in the *hypercontractivity inequality* as well, for instance.

Student Question. *Is this also true for continuous random variables?*

Answer. Yes — the proof doesn't depend on the random variables being discrete. The conditioning as written may not make sense (what conditional expectation is becomes a very subtle issue), but it can be made to work in other settings.

§16.2 Some Quick Consequences

We'll now take the statement and generalize it to *decreasing* events, as well as *multiple* events.

Corollary 16.8

If A and B are events on independent Boolean random variables, then:

(a) If A and B are both increasing, then $\mathbb{P}(AB) \geq \mathbb{P}(A)\mathbb{P}(B)$.

(b) If A is increasing and B is decreasing, then $\mathbb{P}(AB) \leq \mathbb{P}(A)\mathbb{P}(B)$.

(c) If A and B are both decreasing, then $\mathbb{P}(AB) \geq \mathbb{P}(A)\mathbb{P}(B)$.

Furthermore, if we have *several* increasing events $A_1, \ldots, A_k$, then

$$\mathbb{P}(A_1 \cdots A_k) \geq \mathbb{P}(A_1) \cdots \mathbb{P}(A_i).$$

Proof. For (b) and (c), note that if A is decreasing, then $\overline{A}$ is increasing. Then instead of working with decreasing events, we can plug in the complements into Harris Inequality; the inequality we get out is precisely what is claimed.

For (d), we can do induction — we can show this step-by-step (adding one event at a time), since all products of these events are increasing. □

§16.3 Some Applications to Random Graphs

Question 16.9. What is the probability that $G(n,p)$ is triangle-free?

We'd like to have some reasonable estimate that $G(n,p)$ is triangle-free, for different regimes of p — typically it'll be a very small probability, and we'd like to estimate it.

The correlation inequality will allow us to get some lower bound.

For each triple ijk of distinct vertices, the event that ijk is not a triangle is a decreasing event — the random variables are the edge indicator random variables (this is always true when we're referring to $G(n,p)$), and making some edges from 1 to 0 only helps us.

So then we can apply Harris Inequality — we have

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) = \mathbb{P}(ijk \text{ is not a triangle for all } ijk).$$

This is the logical *and* of a bunch of decreasing events, so we can use Harris inequality to obtain that this is at least

$$\prod_{ijk} \mathbb{P}(ijk \text{ is not a triangle}) = (1-p^3)^{\binom{n}{3}}.$$

This gives us the following result:

Theorem 16.10

The probability that $G(n,p)$ is triangle-free is at least $(1-p)^{\binom{n}{3}}$.

Question 16.11. How good is this bound?

For the most part, we'll assume $p \leq 0.99$ (the problem looks weird when p is close to 1). In this case, we have $1 - p^3 \asymp e^{-\Theta(p^3)}$ (this is a good approximation when $p = o(1)$, but it's always true). So the above estimate gives us a lower bound of

$$(1-p)^{\binom{n}{3}} = e^{-\Theta(n^3p^3)}.$$

We can think about whether there's easier ways to obtain lower bounds. Another lower bound is that the probability $G(n,p)$ is triangle-free is that $G(n,p)$ is empty (since if $G(n,p)$ is empty then it is certainly triangle-free). This probability is $(1-p)^{\binom{n}{2}} = e^{-\Theta(n^2 p)}$.

When p is constant, our bound from Harris does horribly — we have e^{-n^3} whereas the silly bound is e^{-n^2}. But as p gets closer to 0, our bound becomes better — there's a transition roughly at $p = 1/\sqrt{n}$. But putting the two estimates together, we get that

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \geq \begin{cases} e^{-\Theta(n^3p^3)} & \text{if } p \lesssim 1/\sqrt{n} \\ e^{-\Theta(n^2p)} & \text{if } 1/\sqrt{n} \lesssim p \leq 0.99 \end{cases}.$$

It turns out that these bounds are tight — when we see Janson's inequality, we'll see how to *upper-bound* this inequality.

This is kind of surprising, given that it seems that we've done very simple things to get these lower bounds; we'll do trickier things to prove the upper bounds, but it turns out that this is the answer.

§16.4 Maximum Degree

Question 16.12. What's the probability that the maximum degree of $G(n,1/2)$ is less than $n/2$?

Individual vertices satisfy this with probability $1/2$. Knowing this, by Harris's inequality, this is at least

$$\prod_v \mathbb{P}(\deg v \leq n/2) \geq 2^{-n},$$

since these are all decreasing events.

This is a true bound; we can now ask how good it is. It turns out that it is not very good, and this is where the surprise is.

Theorem 16.13 (Riordan–Selby 2000)
$\mathbb{P}(\max\deg G(n,1/2) \leq n/2) = (0.6102\cdots + o(1))^n$.

In particular, the base is not $1/2$; it's greater.

We won't prove this, but we'll see some intuition for why this is true. Instead of considering *boolean* random variables, we can consider a slightly different model where we assign an independent standard *normal* random variable $N(0,1)$ to each edge of K_n; call Z_{uv} the variable assigned to the edge uv. Then we can define $W_u = \sum_v Z_{uv}$ (as v ranges over all the other vertices).

This is supposed to be a model for the maximum degree. The intuition is that they should have similar behavior; at the least, it is intuitively a *reasonable* model.

Assuming this, we can then think about $\mathbb{P}(W_u \leq 0 \text{ for all } u)$.

Theorem 16.14
$\mathbb{P}(W_u \leq 0 \text{ for all } u) = (0.6102\cdots + o(1))^n$.

Harris's inequality also gives a lower bound of 2^{-n} in this case; so we still have an exponentially larger inequality.

Proof sketch. The main point here is that unlike $G(n,p)$, where analyzing exactly is difficult, in random normal variables we can write down a pretty closed formula for this expression because of properties like rotational invariance. More explicitly, we can think about the tuple $(W_1, \dots, W_n)$. This is a n-tuple of random variables, and each one is a sum of i.i.d. normal variables, so is normal. But they're not independent — two of them will share some component. Then $(W_1, \dots, W_n)$ is a *join normal*, and it's completely specified by its covariance matrix.

We have $\operatorname{Var} W_i = n - 1$, since it's the sum of $n-1$ standard Gaussian normals. Meanwhile,

$$\operatorname{Cov}[W_i, W_j] = 1,$$

since W_i and W_j share exactly one summand (we can expand and calculate this explicitly).

As a result, we see that $(W_1, \dots, W_n)$ can be rewritten — it has the same distribution as taking i.i.d. random normals $(Z_1', \dots, Z_n')$, normalizing by multiplying by $\sqrt{n-2}$, and adding an additional component $Z_0'(1, \dots, 1)$ (where we take another standard normal multiplied by the all-1 vector).

We can see that the covariances are all the same, so they're the same distribution. But we can analyze our new situation quite exactly. Let Φ be the cdf of $N(0,1)$, so that

$$\mathbb{P}(W_v \le 0 \text{ for all } v) = \mathbb{P}\left(Z_i' \le -\frac{Z_0'}{\sqrt{n-2}} \text{ for all } i\right).$$

To calculate the right-hand side, we can first condition on the value of Z_0, and then integrate over the distribution of Z_0. This gives

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \Phi\left(-\frac{z}{\sqrt{n-2}}\right)^n dz$$

(where z is the value of Z_0'). This is a pretty closed formula, and the rest is some semi-routine asymptotic analysis to figure out how this expression behaves as n gets large. This is orthogonal to the point, so we won't get into the details; but if we look at an expression like this, it's an integral where what's inside the integral is raised to the power of n, and we can ask ourselves where the dominant contribution to the integral is. This should happen where the integrand has the maximum value. So roughly what happens is — suppose we do a substitution where we set $z = y\sqrt{n}$. Then we find that the integral can be rewritten as

$$\sqrt{\frac{n}{2\pi}} \int_{-\infty}^{\infty} e^{nf(y)}\, dy,$$

where $f(y)$ is the expression

$$f(y) = -\frac{y^2}{2} + \log \Phi\left(y\sqrt{\frac{n}{n-2}}\right).$$

This dependence on n is very small, so we will ignore it; then

$$f(y) \approx \frac{y^2}{2} + \log \Phi(y).$$

Then when we let n get large, basically the only point that matters is where f achieves its maximum (everything else dies away exponentially quickly; this can be justified by Taylor-expanding f around its maximum point).

So then this expression is essentially just $(\max e^f + o(1))^n$. Then we can figure out what value of y maximizes the function (which is a very concrete problem that can be solved analytically); this is where the number $0.6102\cdots$ comes from. □

If we want to know how to prove the result about random graphs, we can look at their paper; the proof comes down to a way to rigorously model the random graph setting by the normal approximation, which is quite technical. But the other thing to remember is that we did get a crude estimate using Harris's inequality, but it turns out that it's way off.

§17 November 2, 2022 — Janson Inequalities

We'll spend the next few lectures discussing a set of inequalities collectively known as Janson inequalities, which will provide exponentially good upper bounds on tail events.

When Spencer was here, he mentioned that at a random graphs workshop, he posed a question about the probability of being triangle-free, and Janson went to his room and came up with the solution; this led to the Janson inequalities.

Question 17.1. What is the probability that $G(n, p)$ is triangle-free?

Last time, we gave a lower bound using Harris inequality, as well as a different lower bound using the observation that an empty graph is triangle-free. Today we will see the matching upper bound inequalities — that the bounds we saw last lecture are tight.

§17.1 The General Setup

We're considering a random subset $R \subseteq [N]$ of a ground set $\{1, 2, \dots, N\}$, where each element is included independently. Usually our elements will be included with the same probability, but they don't have to be.

We fix a collection $S_1, \dots, S_k$ of subsets of $[N]$, and we let A_i be the event that $S_i \subseteq R$.

Our goal is to count $X = \sum \mathbf{1}_{A_i}$, the number of containments — for example, we can think of X as the number of triangles in the random graph (the ground set consists of all the edges, we have a list of all the triangles S_i, and we're counting how many triangles are contained in our graph).

We let $\mu = \mathbb{E}X$, and we form a dependency graph — we write $i \sim j$ if $i \neq j$ and S_i and S_j are not disjoint. As in the second moment method, we write

$$\Delta = \sum_{i \sim j} \mathbb{P}(A_i A_j).$$

(where the sum is over all ordered pairs (i, j) with overlapping S-sets).

Remark 17.2. Note that in this notation, each pair is counted twice.

Here we have a dependency graph based on pairwise dependence, but this is fine because we're in the random variable model, and we're only counting containments.

Theorem 17.3 (Janson Inequality I)
In this setup, $\mathbb{P}(X = 0) \leq e^{-\mu + \Delta/2}$.

This gives an upper bound on the probability of nonexistence. This inequality is only useful if Δ is small compared to μ.

This is a familiar setting of the second moment method — Δ is also something we used to bound the covariance in the second moment method. We knew in that situation that $X = 0$ has *small* probability. But Janson's inequality tells us it has *exponentially* small probability — which is much better than we were able to obtain using the second moment method.

Remark 17.4. When all the events A_i have probabilitiy $o(1)$, which is typically the case, an application of Harris inequality gives us a lower bound of the form

$$\mathbb{P}(X = 0) = \mathbb{P}(\overline{A_1} \cdots \overline{A_j}) \geq \mathbb{P}(\overline{A_1}) \cdots \mathbb{P}(\overline{A_k}) = \prod (1 - \mathbb{P}(A_i))$$

(all the A_i are increasing events, so their complements are decreasing events). And because these event probabilities are all assumed to be quite small, we have a good approximation of $1 - \mathbb{P}(A_i) \approx e^{-\mathbb{P}(A_i)}$ — so then

$$\mathbb{P}(X = 0) \geq \exp\left(-(1 + o(1))\sum_{i=1}^{k} \mathbb{P}(A_i)\right) = e^{-(1+o(1))\mu}.$$

If furthermore $\Delta = o(\mu)$, then this is a matching lower bound — so the probability of nonexistence is simply $e^{-(1+o(1))\mu}$.

This should look familiar — it's Poisson-like (in the sense that the same is true for the Poisson distribution). In fact, if μ is a constant, then in many cases it's actually true that X converges to a Poisson distribution.

Student Question. *Does this generalize to infinitely many A_i?*

Answer. In some cases you can take the limit — if the probability rapidly decays, so that our expressions are finite.

Let's now prove Janson's inequality. The proof we will see is in some sense reminiscent of the proof of the Lovász Local Lemma.

Proof. Let $r_i = \mathbb{P}(A_i \mid \overline{A_1} \cdots \overline{A_{i-1}})$. Observe that

$$\mathbb{P}(X = 0) = \mathbb{P}(\overline{A_1} \cdots \overline{A_k}) = \mathbb{P}(\overline{A_1})\mathbb{P}(\overline{A_2} \mid \overline{A_1}) \cdots \mathbb{P}(\overline{A_3} \mid \overline{A_1 A_2}) \cdots \mathbb{P}(\overline{A_k} \mid \overline{A_1} \cdots \overline{A_{k-1}}).$$

We can write each factor in terms of our r_i, so we have

$$\mathbb{P}(X = 0) = (1 - r_1)(1 - r_2) \cdots (1 - r_k).$$

It suffices to prove the following claim:

Claim — For every i, $r_i \geq \mathbb{P}(A_i) - \sum_{j<i, j\sim i} \mathbb{P}(A_i A_j)$.

This claim gives a lower bound on the r_i — it says that we can forget about the conditioning with a small bit of cost.

First let's see why the claim implies the theorem. By summing over all i, we obtain that

$$\sum r_i \geq \sum \mathbb{P}(A_i) - \sum_i \sum_{j<i, j\sim i} \mathbb{P}(A_i A_j) = \mu - \frac{\Delta}{2}.$$

Thus the probabilitiy that $X = 0$ is

$$\mathbb{P}(X = 0) = \prod (1 - r_i) \leq \exp\left(-\sum r_i\right) \leq e^{-\mu + \Delta/2}.$$

So it suffices to prove this claim; now we will prove the claim.

Proof. To prove this claim, we will do somewhat tricky manipulations with conditional probabilities. (This is the part that will look somewhat reminiscent to the proof of the Lovász Local Lemma.) First fix i for the rest of this proof.

We define two events D_0 and D_1, such that

$$D_0 = \bigwedge_{j<i, j\not\sim i} \overline{A_j}$$

whereas similarly

$$D_1 = \bigwedge_{j<i, j\sim i} \overline{A_j}.$$

(This is similar to the separation used in the proof of the Lovász Local Lemma — the point is to separate the conditioned event into D_0 and D_1.)

Then by definition,

$$r_i = \mathbb{P}(A_i \mid \overline{A_1} \cdots \overline{A_{i-1}}) = \mathbb{P}(A_i \mid D_0 D_1).$$

By Bayes' Rule, we can rewrite this as

$$r_i = \frac{\mathbb{P}(A_i D_0 D_1)}{\mathbb{P}(D_0 D_1)}.$$

Now let's do some inequalities. First we can drop an event in the denominator — we have

$$r_i = \frac{\mathbb{P}(A_i D_0 D_1)}{\mathbb{P}(D_0 D_1)} \geq \frac{\mathbb{P}(A_i D_0 D_1)}{\mathbb{P}(D_0)} = \mathbb{P}(A_i D_1 \mid D_0).$$

(Effectively, we've moved one of our events from the condition side to the actual event side.)

We want to have some lower bound on this, and we want to use Harris inequality. But right now it's a bit complicated because A_i is an increasing event (containment), D_0 is a decreasing event (it's non-containment), and D_1 is also decreasing. And this doesn't allow us to use Harris.

The next part looks simple, but it is a tricky step: we rewrite this as

$$\mathbb{P}(A_i \mid D_0) - \mathbb{P}(A_i \overline{D_1} \mid D_0).$$

And now we are in a position to do a few things. First, A_i and D_0 are independent by the way D_0 was defined, so $\mathbb{P}(A_i \mid D_0) = \mathbb{P}(A_i)$.

In the second term, A_i is increasing, $\overline{D_1}$ is increasing, and D_0 is decreasing. We know increasing events are negatively correlated with decreasing events, so then we can drop the conditioning using Harris — this gets

$$\mathbb{P}(A_i D_1 \mid D_0) \geq \mathbb{P}(A_i) - \mathbb{P}(A_i \overline{D_1}).$$

It remains to analyze the term $\mathbb{P}(A_i \overline{D_1})$. And we see that $\overline{D_1}$ is the *or* of its constituent A_j, so

$$\mathbb{P}(A_i \overline{D_1}) = \mathbb{P}\left(A_i \wedge \bigvee\nolimits_{j<i, j\sim i} A_j\right).$$

Using the union bound, this is at most

$$\sum_{j<i, j\sim i} \mathbb{P}(A_i A_j),$$

as desired. ■

So this finishes the proof of the first Janson inequality. □

Question 17.5. What did we actually use about the set, that we didn't use in the Lovász Local Lemma?

The relevant step was when we used Harris. This requires that our events are *increasing* — it doesn't work in general.

It turns out that if we're already working with increasing events, then having pairwise independence is equivalent to depending on disjoint proofs of variables — so it turns out that this is not actually an issue. So everything here does work for general increasing events.

This is a bit of a tricky question because the original Janson inequality does *not* work for increasing events in general.

So the name 'Janson inequality' is reserved for *principle upsets* (counting containments). It was a recent result that generalized it to general increasing events, and the proof shown here is very different from Janson's original method (which involved analytic interpolation — we start with one event, and do analysis to interpolate it to a different event).

This proof builds on a proof by Spencer, but with a few additional modifications that simplify the proof quite a bit — the proof in the textbook looks different.

So we'll only use it for counting containments, but it turns out that it works for increasing events in general.

§17.2 Triangle-Free Graphs

Question 17.6. What is the probability that $G(n, p)$ is triangle-free?

Here the ground set $[N]$ is the set of all possible edges, $\binom{[n]}{2}$, and $S_1, \ldots, S_{\binom{n}{2}}$ is the set of triangles in K_n; and X is the number of triangles. We'd now like to estimate μ and Δ.

We have actually done this before — we have

$$\mu = \binom{n}{3}p^3 \asymp n^3p^3.$$

Meanwhile Δ counts pairs of triangles that overlap; there are n^4 ways to choose the vertices, and the probability of seeing this configuration is p^5. This gives $\Delta \asymp n^4p^5$.

To apply the Janson inequality, we should be working in the regime where $\Delta = o(\mu)$. When $p \ll 1/\sqrt{n}$ this is true, and applying Janson gives us the following:

Theorem 17.7
If $p = o(n^{-1/2})$, then

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \asymp e^{-(1+o(1))\mu} \asymp e^{(1+o(1))n^3p^3/6}.$$

Janson's inequality gives us the upper bound; but in fact, because of what we said earlier about using Harris inequality to get a lower bound, this is not just an upper bound, but an asymptotic equality.

This gives a satisfactory answer when p is small — when $p \ll 1/\sqrt{n}$. In particular, when $p = c/n$, then the answer is $e^{(1+o(1))c^3/6}$. When c is a constant, this gives the result we found in the homework by checking moments. One of the motivations for Janson's inequality (as mentioned by Spencer) was what happens when we allow c to grow. Then the moment method no longer works; but Janson's inequality still works to give us this estimate.

Question 17.8. What happens when p is large?

We saw that this is no longer the correct answer — we have a much better answer just by considering the probability that the graph is empty.

When p is large, Δ is much larger than μ, so Janson is not going to work (it will give us no information). So the next thing we will do is bootstrap Janson's inequality to give us a second theorem:

Theorem 17.9 (Janson inequality II)
If $\Delta \geq \mu$, then

$$\mathbb{P}(X = 0) \leq e^{-\mu^2/2\Delta}.$$

The second Janson inequality will be quite effective when Δ is larger than μ.

Surprisingly, we're going to prove the second Janson inequality by applying the first. This might not seem possible because we're looking at very different settings — in the first Δ is small, and in the second Δ is large. But it turns out that this is possible.

This should be somewhat reminiscent to the technique of sampling and boosting that we saw in the linearity of expectations chapter. There we saw the *crossing number inequality* — we showed that the number of crossings in a graph with a lot of edges is quite high. The way we did this was by first coming up with a weak bound — in a *planar* graph, $|E| \leq 3|V|$, which allows us to get a weak bound that $\operatorname{cr}(G) \geq |E| - 3|V|$ (since we can count each excess one at a time, and take it away). And then we used sampling to boost this bound to a much better bound — by starting with a graph, and sampling down to a smaller subgraph.

It's funny because this is a theorem about probabilities, but we will prove it by applying the probabilistic method.

Proof. For each subset $T \subseteq [k]$ (of event indices), let $X_T = \sum_{i \in T} \mathbf{1}_{A_i}$ be the number of events in T that occur. Then if no event occurs, certainly no event in T occurs, which means

$$\mathbb{P}(X = 0) \leq \mathbb{P}(X_T = 0) \leq e^{-\mu_T + \Delta_T/2},$$

where $\mu_T = \mathbb{E}X_T = \sum_{i \in T} \mathbb{P}(A_i)$ and Δ_T is the same as Δ, but restricted only to events in T — so

$$\Delta_T = \sum_{(i,j) \in T, i \sim j} \mathbb{P}(A_i A_j).$$

Now we're going to choose $T \subseteq [k]$ randomly, where every element is included with probability q.

We have an upper bound, and we'd like to know if we can get a better upper bound by just looking at a subset of events. If we do this randomly, then

$$\mathbb{E}\mu_t = q\mu,$$

since every event is kept with probability q. Furthermore,

$$\mathbb{E}\Delta_T = q^2\Delta,$$

since every pair of events is kept with probability q^2. And so our exponent from Janson has expectation

$$\mathbb{E}[-\mu_T + \Delta_T/2] = -q\mu + \frac{q^2}{2}\Delta.$$

So by linearity of expectations, there exists some choice of T so that

$$-\mu_T + \frac{\Delta_T}{2} \leq -q\mu + \frac{q^2}{2}\Delta.$$

Now we just have to make sure that we can choose q so that the right-hand side is small. We can set $q = \mu/\Delta$, which is in $(0, 1)$ because we assumed $\mu \le \Delta$. With this choice, we have that

$$-q\mu + \frac{q^2}{2}\Delta = -\frac{\mu^2}{2\Delta},$$

which gives the desired upper bound on $X = 0$. □

This is a sampling-and-boosting proof — ew started with a relatively weak bound, and saw that if we apply the weak bound to the entire subset, then Δ blows up. But by restricting ourselves to a subset of events, we can get a better bound — and the way we pick that subset of events is by doing it at random.

So we now have two Janson inequalities, which work in complementary regimes; they both give exponential tails, which is important in some applications.

Now let's revisit our question:

Question 17.10. What is the probability that $G(n, p)$ is triangle-free?

In the regime $p \gg 1/\sqrt{n}$, we see that $\mu \asymp n^3p^3$ and $\Delta \asymp n^4p^5$; we can see that $\Delta \gg \mu$. So Janson II tells us that

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \le e^{-\mu^2/2\Delta} = e^{-\Theta(n^2p)}.$$

Meanwhile, we also saw that

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \ge \mathbb{P}(G(n,p) \text{ is empty}) = e^{-\Theta(n^2p)},$$

so this is tight in the exponent.

We can combine these results:

Proof. Suppose that $p \le 0.99$. Then

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) = \begin{cases} \exp\left(-\Theta(n^2p)\right) & \text{if } p \gtrsim 1/\sqrt{n} \\ \exp\left(-\Theta(n^3p^3)\right) & \text{if } p \lesssim 1/\sqrt{n}. \end{cases}$$

□

Technically we have not addressed $p \asymp 1/\sqrt{n}$, but the same proof works.

Remark 17.11. Some related results we know, giving much more precise information about triangle-freeness:

1. In the very dense case, when $p = 1/2$ (so we're just counting graphs), we know the number of triangle-free graphs on n vertices is $2^{(1+o(1))n^2/4}$. One way to get lots of triangle-free graphs is to fix some complete bipartite graph, and look at all its subgraphs; and this essentially gives the right answer. In fact, it's also true that
$$\mathbb{P}(\text{bipartite} \mid \text{triangle-free}) \to 1.$$
2. Suppose $m \geq Cn^{3/2}\sqrt{\log n}$ for some constant $C > \sqrt{3}/4$. Then a $(1 - o(1))$-fraction of n-vertex m-edge triangle-free graphs are bipartite. Furthermore, this constant C is the best possible — so it's a sharp threshold for the phenomenon that a particular triangle-free graph is bipartite. (It's a sparser version of what happens in the first setting.)
3. For $1/\sqrt{n} \ll p \ll 1$,
$$-\log \mathbb{P}(G(n,p) \text{ is triangle-free}) \sim -\log \mathbb{P}(G(n,p) \text{ is bipartite}) \sim \frac{n^2 p}{4}.$$

One technique involved in this line of work is the hypergraph container result that we'll discuss towards the end of the course.

§17.3 Lower Tails

The third Janson inequality has to do with *lower* tails.

Question 17.12. If X is the number of triangles in $G(n,p)$, what is $\mathbb{P}(X \leq (1-\delta)\mathbb{E}X)$, for some positive real number δ?

This should be a fairly rare event, and we'd like to bound its probability.

Theorem 17.13 (Janson III)
For $0 \leq t \leq \mu$,
$$\mathbb{P}(X \leq \mu - t) \leq \exp\left(-\frac{t^2}{2(\mu + \Delta)}\right).$$

Remark 17.14. In a way, this encapsulates the first two versions we've seen so far — by setting $t = \mu$, we get that
$$\mathbb{P}(X = 0) \leq e^{-\mu^2/2(\mu+\Delta)},$$
and if we don't care about constant factors in the exponent, we see that this is the combination of those two statements. So this is convenient even if we just care about nonexistence.

So far, the tail bounds we've done in this class are Chebyshev's inequality (which gives polynomially slow decay), and Chernoff. This is exponentially quick decay, so it should remind us of Chernoff. The method of proof of Chernoff was to look at the *exponential* moment generating function and use Markov on it.

We will do the same here.

Proof. First, we'll rewrite this in terms of the exponential moment generating functions, as

$$\mathbb{P}(X \le \mu - t) = \mathbb{P}(e^{-\lambda X} \ge e^{-\lambda(\mu - t)}).$$

Now we will do something that looks a bit crazy — let $1 - q = e^{-\lambda}$ (with $\lambda > 0$ and $q \in [0, 1]$). Then using Markov's inequality, we see that

$$\mathbb{P}(e^{-\lambda X} \ge e^{-\lambda(\mu-t)}) \le e^{\lambda(\mu-t)}\mathbb{E}e^{-\lambda x} = (1-q)^{-\mu+t}\mathbb{E}[(1-q)^X].$$

So far, this is just an exponential moment calculation. But now we'll look at the right-hand side and see if we can interpret it combinatorially. Since X counts the number of containments, $(1-q)^X$ should have some interpretation.

Let $T \subseteq [k]$ where each element is included with probability q independently. Similarly to before, let X_T be the number of events in T that occur.

> **Claim** — $\mathbb{P}(X_T = 0 \mid X) = (1-q)^X$.

In other words, if I tell you that three events have occurred, and then ask for the probability that none of the events in T occur, the answer is $(1-q)^3$.

Now taking the expectation gives us

$$\mathbb{E}[(1-q)^X] = \mathbb{P}(X_T = 0)$$

(by expectation chaining). But to bound the probabilty that $X_T = 0$, we can use Janson I (conditioning on T, not X). Similarly to what happened in the proof of Janson II, we condition on T that's optimally chosen; this then gives a bound of

$$\mathbb{E}[(1-q)^X] \le \mathbb{E}_T e^{-\mu_T + \Delta_T/2}.$$

This is not quite okay for us — what we want is to acutally have T in the exponent. But by convexity, we can do this — so then we get that this is at most

$$\mathbb{E}_T e^{-\mu_T + \Delta_T/2} \le \exp\left(E_T\left[-\mu_T + \frac{\Delta_T}{2}\right]\right) = \exp\left(-qu + \frac{q^2}{2}\Delta\right).$$

Finally, we want to pick a good choice of μ. (Note: this is not correct — we want to already have chosen the best T, but how does that work with this conditioning?)

Because $1 - q = e^{-\lambda}$ and $\lambda \ge 0$, we have $\lambda - \lambda^2/2 \le q \le \lambda$ (by Taylor approximation). So in our calculation from earlier, we get an upper bound of

$$(1-q)^{-\mu+t}e^{-q\mu+q^2\Delta/2} \le \exp\left(\lambda(\mu - t) - \left(\lambda - \frac{\lambda^2}{2}\right)\mu + \lambda^2\frac{\Delta}{2}\right).$$

Now we can pick the optimal value of λ; we can first clean up this expression to

$$\exp\left(\lambda t - \frac{\lambda^2}{2}(\mu + \Delta)\right).$$

Then if we pick $\lambda = 1/(\mu + \Delta)$, then we get an upper bound of

$$\exp\left(-\frac{t^2}{2(\mu+\Delta)}\right). \qquad \square$$

The exponential calculations are familiar from the Chernoff bound. The part that's new and clever is reinterpreting the exponential moment as a probabilistic combinatorial situation — we are interpreting this as the probability that the restricted set of events has no occurrences, and then using the first Janson inequality to do an upper bound.

Remark 17.15. This proof was never published, and is very different to the proof originally due to Janson. But it is very nice.

We'll conclude with an immediate application to triangle counts: if X is the number of triangles in $G(n, p)$, and $\delta \in (0, 1]$ is fixed, then

$$\mathbb{P}(X \le (1-\delta)\mathbb{E}X) \le \begin{cases} \exp\left(-\Theta(n^2 p)\right) & p \gtrsim n^{-1/2} \\ \exp\left(-\Theta(n^3 p^3)\right) & p \lesssim n^{-1/2}. \end{cases}$$

(Here the constants depend on δ.)

Note that we also have the lower bounds from our lower bounds for $X = 0$.

Remark 17.16. The constants are an open question. It is known that if p is quite close to 1, and δ is quite small (say, less than 0.01), we can say what the constant is — it's what happens if we sample in $G(n, q)$ with a slightly smaller q. But if δ is very close to 1, this isn't true — there is a phase transition where we go from 'replica symmetry' to 'symmetry breaking'.

§18 November 7, 2022

§18.1 Janson Inequalities

In the Janson inequalities, we have a random set $R \subseteq [N]$, with each element included with some probability (which doesn't have to be the same over all elements). We fix a collection of sets $S_1, \ldots, S_k$, and we define A_i to be the event that $S_i \subseteq R$. We want to count how many of these sets R contains, so we let $X = \sum_i \mathbf{1}_{A_i}$.

In random graphs, we think of the ground set as being the set of edges, and the sets (for example) as being the collection of all triangles.

We use μ to denote $\mathbb{E}X$, and we write $i \sim j$ if $i \ne j$ and S_i and S_j intersect; we also write

$$\Delta = \sum_{i \sim j} \mathbb{P}(A_i A_j).$$

(Note that in this sum, by convention we sum each pair twice — once for (i, j) and once for (j, i).)

Last time, we stated several Janson inequalities.

Theorem 18.1 (Janson I)
We have

$$\mathbb{P}(X = 0) \le e^{-\mu + \Delta/2}.$$

Theorem 18.2 (Janson II)
If $\Delta \ge \mu$, then

$$\mathbb{P}(X = 0) \le e^{-\mu^2/2\Delta}.$$

These two theorems are about non-existence; the third is about lower tails.

Theorem 18.3 (Janson III)

For any $0 \leq t \leq \mu$,

$$\mathbb{P}(X \leq \mu - t) \leq \exp\left(-\frac{t^2}{2(\mu+\Delta)}\right).$$

Remark 18.4. By plugging in $t = \mu$, we can recover the case $X = 0$; this expression basically encompasses the first two, up to a constant factor in the exponent.

Janson I is useful when Δ is small compared to μ; Janson II is useful when Δ is large compared to μ.

In the second moment method, we saw that if we had control on the second moment, then we could understand the probability of nonexistence through Chebyshev — but those bounds decay quite slowly. Meanwhile Janson gives exponential decay — but we're only allowed to apply them in the more limited setting of counting containments.

Last time, we proved I and II and tried to prove III, but we made a mistake. We will begin this lecture by correcting this mistake.

Proof. As in the proof of the Chernoff bound, we first write this in terms of the exponential generating function and apply Markov's inequality. As a result, we want to upper bound

$$\mathbb{E}[(1-q)^X].$$

We do this by coming up with a new random process, and then applying Janson I to the new random process.

The random process is the following: for each $i \in [k]$ (indexing an event), let $W_i \sim \text{Bernoulli}(q)$ be an independent Bernoulli variable. Now we consider the new random variable

$$Y = \sum_{i=1}^{k} \mathbf{1}_{A_i} W_i.$$

(So we only count our containments with probability q — depending on whether our Bernoullis are on and off.)

Note that Y also falls in the framework of Janson, with an extended ground set — with the ground set extended by k elements. So we can apply Janson's inequality to Y. (We essentially create Y by starting with our same set and adding k random variables; then each S_i is appended a new element, corresponding to this Bernoulli.)

Now let's apply Janson's inequality to Y. To do this, we need to figure out the relevant parameters.

First, we have

$$\mu_Y = \mathbb{E}Y = q\mu,$$

since W_i has expectation q and is independent from everything else. Meanwhile,

$$\Delta_Y = \sum_{i \sim j} \mathbb{P}(A_i A_j W_i W_j).$$

Again, because W_i and W_j are independent from everything else, we have

$$\Delta_Y = q^2 \Delta.$$

Now we apply Janson I to this situation, which implies that

$$\mathbb{P}(Y = 0) \leq e^{-\mu_Y + \Delta_Y/2} = e^{-q\mu + q^2\Delta/2}.$$

So far, we've created a new random variable and applied Janson I to it, for any $q \in (0,1)$.

To relate this to our quantity $\mathbb{E}[(1-q)^X]$, note that $\mathbb{P}(Y = 0 \mid X) = (1-q)^X$. (If I tell you the value of X — the number of original events that occurred — then conditioning on this, for Y to be 0 we must have all X Bernoullis be 0). Now taking the expectation over X of both sides, we get that

$$\mathbb{E}[\mathbb{P}(Y = 0 \mid X)] = \mathbb{E}[(1-q)^X].$$

On the left, the expectation eliminates the condition, so

$$\mathbb{E}[(1-q)^X] = \mathbb{P}(Y = 0),$$

which is the quantity that we just upper-bounded.

That completes the estimate that we feed into the rest of the argument with exponential generating functions and Markov, and that completes the proof. (The point overlooked last time is that we need to create new random variables so that we can apply Janson I.) □

§18.2 Upper Tails

The third Janson inequality automatically gives us a consequence about the lower tails of triangle counts — if X is the number of triangles in $G(n,p)$, then

$$\mathbb{P}(X \le (1-\delta)\mathbb{E}X) = \begin{cases} \exp(-\Theta(n^2 p)) & \text{if } p \gtrsim 1/\sqrt{n} \\ \exp(-\Theta(n^3 p^3)) & \text{if } p \lesssim 1/\sqrt{n} \end{cases}.$$

This is tight — for the lower bound we can use Harris's inequality, and we can use Janson to get an upper bound. In the first case, the lower bound comes from the probability that the graph is *empty.*

But this example leads us to the next discussion, that there's an asymmetry between upper and lower tails. In the discussion of Janson's inequality, we only talked about a *lower* tail probability estimate. It's natural to ask whether the same inequality would hold if we replaced our condition with $X \ge \mu + t$. In the second moment method, there was no difference. But for Janson, it really does matter — a statement like this becomes false if we replace lower tails with upper tails.

If we again consider the problem of triangle counts — the probability that $X \ge (1+\delta)\mathbb{E}X$ — then the behavior is quite different.

One way to obtain a lower bound on this probability is to construct an event that generates lots of triangles, and estimate the probability of that event.

One way to get a lot of triangles is to uniformly lift the edge density — if instead of having edges generated with probability p, we have them generated with higher probability. This would give some lower bound.

An even simpler way would be if we force some edges to be present — then we can get lots of triangles.

Suppose we force edges in a set of cnp vertices to be present. Then the number of triangles is at least $\binom{cnp}{3}$, which for an appropriate value of c is already significantly larger than the expected number of triangles (because the expected number of triangles is on the order of n^3p^3, so as long as we choose c large enough, forcing these edges to be present already gives lots of triangles). So in particular, we have a lower bound of the form

$$\mathbb{P}(X \ge (1+\delta)\mathbb{E}X) \ge p^{\binom{cnp}{2}} = e^{-\Theta(n^2p^2 \log 1/p)}$$

(where the hidden constant depends on δ). This is much bigger than our previous answer — we already have a lower bound on our upper tail probabilities that exceeds what happened for lower tails.

This is perhaps a bit surprising; it's an example that shows that Janson's inequality is really about what happens near the bottom, and there's a completely different story for upper tails.

Remark 18.5. The question of $\mathbb{P}(X \geq (1+\delta)\mathbb{E}X)$ was a subject of much research. This is a difficult question that was at the center of probabilistic combinatorics — it was seen as a litmus test for various concentration inequalities (and new methods have been developed specifically for this problem).

For a long time, there was not much progress — this problem was called the *infamous upper tail.*

In 2012, two papers independently showed that

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = p^{\Theta(n^2p^2)} \text{ if } p \gtrsim \frac{\log n}{n}.$$

(The behavior when p is smaller is somewhat different, but we won't discuss that.)

So this lower bound, up to a constant factor in the exponent, is the correct one. Then there was effort to determine what the right constant factor in the exponent is. The final answer is the following:

Theorem 18.6 (Harel–Mousset–Samotij 2022)
We have

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = p^{(1+o(1))\min\left\{\frac{\delta}{3}, \frac{\delta^{2/3}}{2}\right\}n^2p^2} \text{ if } p \gg \frac{1}{\sqrt{n}}.$$

Meanwhile, this becomes

$$p^{(1+o(1))\frac{\delta^{2/3}}{2}n^2p^2} \text{ if } \frac{\log n}{n} \ll p \ll \frac{1}{\sqrt{n}}.$$

We'll explain how to prove a *lower* bound. As earlier, we just need to find some event that generates lots of triangles cheaply. The way to do this is to figure out how to plant some edges that already generates enough triangles for you to be done — we need to essentially generate $\delta n^3p^3/6$ new triangles by planting edges.

One way to do this is to plant a clique. To do this, we'd need to plant a clique of size cnp for some c depending on δ.

There's also a different way to do this — to plant a *hub*. This means we select a set of bnp^2 vertices, and we force all the edges with one endpoint in this set to be present in the graph.

These edges alone do not generate lots of triangles. But we originally already had $G(n,p)$, and some edges of $G(n,p)$, together with these additional vertices, will give you lots of new triangles; and that will be enough to give the amount you want.

Whichever one uses fewer planted edges is cheaper probabilistically, and that gives the lower bound matching with this expression.

Student Question. *Why do we want to plant this number of triangles?*

Answer. The expected number of triangles is $n^3p^3/6$; we want to plant a factor of δ extra triangles.

Remark 18.7. Prof. Zhao worked on this problem, and his paper with Lubetsky was the first place to find a variation of the first expression.

This belongs to a whole area of mathematics called large deviation theory — if we have a random variable, what's the probability it really deviates from its mean? Often, this problem reduces down to some extremal or variational problem, as in this case.

The constant in the *lower* tail also has an interesting story that isn't completely understood. The story in some ways is similar to the upper tail — you can get lower bounds by coming up with a way to *reduce* the number of triangles. One way to reduce the number of triangles is to force some edges *not* to be present, but that doesn't give the optimal bound (unless you're looking for zero triangles).

But another way is to uniformly compress the density. Suppose instead of generating $G(n,p)$ we generated $G(n,q)$ for some $q < p$, and calculated the relative entropy of $G(n,q)$ with respect to $G(n,p)$ — the probabilistic cost of forcing edges to appear with lower probability. This gets something on the right order of magnitude. Prof. Zhao showed that if $\delta < 0.01$ is quite small, then this gives the right bound. Furthermore, if $\delta > 0.99$ is quite large (close to 1), then this does *not* give the right bound. So there is likely a phase transition somewhere in between, and we do not understand what happens exactly. Somehow, as δ goes from 0.01 to 0.99, it's likely that somewhere there's a behavioral change.

We can phrase this in a different way:

Question 18.8. If we draw a $G(n,p)$ conditioned on having few triangles, what does this conditioned graph look like? In particular, does it look like $G(n,q)$ for some q?

The answer is yes if δ is quite small, and no when δ is quite large. When $\delta = 1$ we know what the graph looks like — a complete bipartite graph with edges sampled with some probability. With δ very large, we think that it may look very close to a complete bipartite graph with p-density, and with some extra edges sprinkled to get some triangles but not very many — but we have no way of proving this. It's an open problem to understand what actually happens.

Student Question. *Do we know what q looks like?*

Answer. The number of triangles in $G(n,q)$ is $q^3n^3/6$, so we want

$$\frac{q^3n^3}{6} = (1-\delta)\cdot\frac{p^3n^3}{6}.$$

We can solve this equation to get the q that we want.

Student Question. *Can we understand $G(n,p)$ conditioned on something computationally?*

Answer. The naive thing to try is to sample $G(n,q)$ and test. This doesn't work very well. It turns out there's a framework reducing this to an optimization problem involving entropy. One can try to do numerics on it, but in practice it is difficult.

Remark 18.9. This is somewhat handwavy, but there's two reasons we saw it. One is to emphasize the point that Janson is for lower tails. Also it's a problem that Prof. Zhao has worked on and likes, and he wants to share with us some of the current research on upper and lower tails for triangle counts.

§18.3 Chromatic Number of a Random Graph

Question 18.10. What is the chromatic number of $G(n,1/2)$?

We're going to use χ to denote the chromatic number.

In teh chapter on the second moment method, we proved that the *clique* number of $G(n,1/2)$, denoted ω, is concentrated around two values — with high probability

$$\omega(G(n,1/2)) \sim 2\log_2 n.$$

We used the second moment to show this — roughly speaking, we computed the expected number of k-cliques for various values of k, and this is the point where the expected number goes quickly from big to close to 0, and we used the second moment method to understand what happens.

The clique number of a graph is the same as the independence number of its complement, so this also tells us that $\alpha(G(n, 1/2)) \sim 2\log_2 n$ with high probability. (The **independence number**, denoted α, is the size of the largest independent set.)

We also know that given any graph G, we have

$$\chi(G) \geq \frac{|V|}{\alpha(G)},$$

since in a proper coloring, all the color classes are independent sets. Combining all of this, we already obtain the following lower bound on $\chi(G)$ — that we must have

$$\chi(G(n, 1/2)) \geq (1 + o(1))\frac{n}{2\log_2 n}$$

with high probability.

Bollobás proved a matching upper bound:

Theorem 18.11 (Bollobás)

With high probability,

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

To prove an upper bound on χ, we need to show that we can color the graph — and each color is some independent set. So we would like to better understand the problem of independent sets in random graphs. For that, we'll use the following lemma:

Lemma 18.12

Let k_0 (which is a function of n) be the largest possible integer k such that $\binom{n}{k}2^{-k/2} \geq 1$. Then

$$\mathbb{P}(\omega(G(n, 1/2)) < k_0 - 3) \leq e^{-n^{2-o(1)}}.$$

Note that $\binom{n}{k}2^{-k/2}$ is the expected number of k-cliques.

Recall that in the second moment method, when k is such that this number is roughly around 1, that's where the clique number should concentrate. This lemma says that if we lower this by a small constant, then the probability drops superexponentially quickly. So it shows us a very strong concentration of the clique number.

Proof. Let $\mu_k = \binom{n}{k}2^{-\binom{k}{2}}$ be the expected number of k-cliques in $G(n, 1/2)$. For $k \sim k_0 \sim 2\log_2 n$, we saw in the second moment method that

$$\frac{\mu k + 1}{\mu_k} = \frac{\binom{n}{k+1}}{\binom{n}{k}}2^{-k} \sim \frac{n}{k} \cdot 2^{-(1+o(1))2\log_2 n}.$$

Here k is quite small, so we can put it into an $o(1)$ in the exponent, giving

$$\frac{\mu_{k+1}}{\mu_k} \sim n^{1-o(1)} \cdot \frac{1}{n^{2-o(1)}} = \frac{1}{n^{1-o(1)}}.$$

So when we change k in this region, μ_k changes by a factor of approximately $1/n$.

Now let $k = k_0 - 3$. We are going to apply Janson inequality, where X is the number of k-cliques. We have $\mu = \mu_k$, and because at k_0 we know $\mu_k \geq 1$, when we decrease k by 3, we increase μ_k by around n^3. This means

$$\mu = \mu_k > n^{3-o(1)}.$$

Meanwhile, it turns out that $\Delta = n^{4-o(1)}$. (We will not do this computation on the board because it is somewhat hairy.) In particular, for sufficiently large n, $\Delta \geq \mu$, so Janson II tells us that

$$\mathbb{P}(X = 0) \leq e^{-n^{2-o(1)}}.$$

But $X = 0$ is precisely the event that we're looking for — that $\omega < k_0 - 3$ (because X is the event of *not* having a clique of size $k = k_0 - 3$). □

So we used Janson to get a bound that decays superexponentially quickly, which will be important for the upcoming steps.

> **Remark 18.13.** Note that up to $o(1)$, this bound is tight. If you increase 3 to a bigger number, you might think that you could get a bigger number, but this doesn't happen. The reason is that if you take an empty graph, it has $\omega(G(n, 1/2)) < k_0 - 3$, and an empty graph has a probability of 2^{-n^2}. So the 2 in the exponent cannot be improved.

Now let's use this to obtain the upper bound on chromatic number. We want to show that there is a strategy to properly color $G(n, 1/2)$ with not too many colors.

But color classes are independent sets. So the general idea is to greedily take out large independent sets and then see what happens; hopefully we don't get stuck.

Note that in the lemma we can replace 'clique' by 'independent set' and everything stays the same, by symmetry.

Proof of Bollobás. The strategy is to take out independent sets of size around $2 \log_2 n$ iteratively, until very few vertices remain — more precisely, until $o(n/\log n)$ vertices remain. At this point, we can just use a new color for every vertex; asymptotically this contributes very few colors, so we're still okay.

The first step is that we already know that $G(n, 1/2)$ has an independent set of this size, so we can take it out — and that's our first color.

An idea that *doesn't* work: in what is remaining, it's still $G(n, 1/2)$ (since n hasn't changed very much), so we can use the same bound to take out another independent set, and keep going.

The problem is that once we remove the vertices in the first independent set, what remains is no longer $G(n, 1/2)$ — we cannot resample the edges (once we've taken out these vertices, we cannot flip all the coins again). So this is why the naive strategy does not work.

Instead, we will use better bounds — in particular, what we just showed — to continue with this strategy.

The way it will work si that we will see $G(n, 1/2)$ once and for all, and we will prove that some event occurs with high probability; and then we will proceed assuming that event from that point onwards.

Suppose $G \sim G(n, 1/2)$, and $m = \lfloor n/(\log n)^2 \rfloor$ (we just need something a bit less than $n/\log n$). For every subset $S \subset [n]$ of exactly m vertices, the induced subgraph $G[S]$ has distribution $G(m, 1/2)$ — the edges are still random and independent.

So then we can apply the lemma to $G[S]$, to get that with $k = k_0(m) - 3 \sim 2 \log_2 m \sim 2 \log_2 n$,

$$\mathbb{P}(\omega(G[S]) < k) = e^{-m^{2-o(1)}}.$$

But m is basically n for all these asymptotics, so we can rewrite this as

$$e^{-n^{2-o(1)}}$$

(since we have a bit of room in the exponent).

This is for a fixed m-vertex subset. Now by taking a union bound, we find that

$$\mathbb{P}(\text{exists a } m\text{-vertex subset } S \text{ with } \alpha(G[S]) < k) < e^{-n^{2-o(1)}} \cdot 2^n$$

by a union bound (there are at most 2^n possible subsets). In particular, this event is extremely unlikely — it has probability $o(1)$.

Now we're going to look at when this event does *not* occur. Let (*) be the property that every m-vertex subset contains a k-element independent set — so with high probability, (*) is true.

Now we're going to assume this is true, and work only with graphs satisfying (*); we will show that every n-vertex graph satisfying the property (*) can be colored with around $n/2\log_2 n$ colors.

The strategy for coloring is as discussed earlier — while there's at least m vertices remaining, by (*) we can find a k-vertex independent set, assign these vertices a color, and remove their vertices. We keep doing this as long as there are at least m vertices in the graph.

Once there are fewer than m vertices, we then color the remaining vertices each with its own color. This results in a proper coloring, and the number of colors used is at most

$$\frac{n}{k} + m \sim \frac{n}{2\log_2 n}.$$

(The term m is $o(n/\log n)$ so it doesn't appear in the asymptotics.) □

So the strategy is to first show that in $G(n, 1/2)$, with high probability every m-vertex subset (where $m = n/\log^2 n$) contains a k-element independent set — this can be shown by the union bound and Janson. Once we have this, we can start a greedy coloring — we keep taking out k-vertex independent sets until we have too few vertices left, at which point it doesn't matter what we do. This finishes the proof.

> **Student Question.** *What if $1/2$ is replaced by some other probability?*
>
> **Answer.** If we replace it with a constant probability, not much changes. If p decreases with n, that's an interesting question. There has been a lot of work on this and Prof. Zhao doesn't know precisely what happens.
>
> In the next chapter, we will discuss some things also related to the chromatic number of $G(n, 1/2)$.
>
> This proof is actually not Bollobás's original proof — it's via martingale concentration inequalities, that we will see soon. (Janson's inequality is a difficult tool — it wasn't easy to develop.)
>
> All the concentration inequalities we've seen before had the property that we have to know where the mean is — to prove that some random variable is concentrated, we first need to find the mean and the variance, and then do some stuff. But it turns out that historically, it was first shown that the chromatic number of $G(n, 1/2)$ is concentrated in a window of width $\sqrt{n}$ — you can determine the concentration window *without* knowing anything about the mean of the random variable. That's the next thing we'll see — we'll be able to determine various quantities are concentrated even without knowing where.
>
> This has a simple proof, but it was a big deal. It left people wondering what is the actual amount of concentration of $G(n, 1/2)$. This was a major open problem that was resolved fairly recently, in the last few years, and there's a surprising answer — this will be discussed more next time.

This finishes the chapter on Janson inequalities. The next thing we'll see is concentration of measure — which is an important set of tools for understanding when a random variable is very close to its mean. We've seen results of this form in the second moment chapter, and in Janson we saw som eexponential *lower* bounds. Next time, we'll develop much more systematically tools for understanding the concentration of random variables.

§19 Concentration of Measure

For the next several lectures, we'll talk about the concentration of measure. Right now, this word may not mean anything to us; hopefully at the end of several lectures, it'll mean something. It's a big subject that we'll only scratch the surface of; and it's important in probability, analysis, and computer science as well.

We've already seen some concentration inequalities earlier in the course. In particular, the Chernoff bound tells us that if z is a sum of n independent Bernoullis (which may have different probabilities), then for all $t > 0$,

$$\mathbb{P}(|z - \mathbb{E}z| > t\sqrt{n}) \le 2e^{-t^2/2}.$$

This tells us that z is very concentrated around its mean, to a window of width $\sqrt{n}$. Beyond that, the property drops off as a Gaussian — so this behavior is sometimes called *sub–Gaussian.*

This is an example of concentration, and we'd like to establish these types of tail bounds in other settings where we may not have a sum of independent Bernoullis.

§19.1 Bounded Differences Inequality

The first theorem we'll see, informally speaking, tells us that a Lipschitz function of many independent random variables is concentrated. In particular, we will prove the *bounded differences inequality*:

Theorem 19.1

Suppose $X_1 \in \Omega_1$, ..., $X_n \in \Omega_n$ are independent random variables, each drawn from some probability space. Suppose that $f \colon \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ is a function satisfying the property that

$$|f(x_1,\ ldots, x_n) - f(y_1, \dots, y_n)| \le 1$$

if $(x_1, \dots, x_n)$ and $(y_n, \dots, y_n)$ differ in exactly one coordinate.

Then the random variable $z = f(X_1, \dots, X_n)$ is very concentrated: for all $t > 0$,

$$\mathbb{P}(z - \mathbb{E}z \ge t) \le e^{-2t^2/n} \text{ and } \mathbb{P}(z - \mathbb{E}z \le -t) \le e^{-2t^2/n}.$$

(This is a discrete version of continuity — if we change the input at just one coordinate, then the output doesn't change very much.)

So the bounded differences inequality says that if we have a function taking n input variables, such that if we flip one coordinate then the output never changes by more than 1, then if we input independent random variables into the function, the output is concentrated according to these inequalities.

In particular, the bound $\mathbb{P}(z - \mathbb{E}z \ge t) \le e^{-2t^2/n}$ is very similar to that of Chernoff.

> **Remark 19.2.** For the function $f(x_1, \dots, x_n) = x_1 + \cdots +_n$, we get Chernoff.

This is much more powerful, because we can apply it to arbitrary functions as long as they're not affected much by changing individual coordinates — even if the function is hard to understand in general.

Example 19.3 (Coupon Collector)

Suppose we have coupons from $[n]$, and we draw s_1, ..., s_n from $[n]$ uniformly and independently (with replacement — some coupons we draw may be repeated). Let the number of *missing* elements be $Z = [n] \setminus \{s_1, \dots, s_n\}$.

Then Z is a function fo $s_1, \ldots, s_n$, and changing one coordinate changes Z by at most 1 in absolute value. If one of our coupons look different, we might have one more coupon missing, or one fewer missing, or the same number; but we can't change by more than 1.

So as a result, we can apply the bounded differences inequality to deduce that

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \lambda) \leq 2e^{-2\lambda^2/n}.$$

In other words, Z has sub-Gaussian concentration with a window of $\sqrt{n}$.

In this case, it's not hard to figure out what $\mathbb{E}Z$ is — there are n possible elements, and each is missing with probability $(1 - 1/n)^n$, so

$$\mathbb{E}Z = n\left(1 - \frac{1}{n}\right)^n \approx \frac{n}{e}.$$

But in many cases, it might actually be very difficult to determine $\mathbb{E}Z$ — but with this tool, we can determine its concentration even *without* knowing the expectation.

In all the previous examples we saw in this class, we first found the mean and then did variance calculations to determine concentration. But here we've managed to decouple them — and in some cases, finding the concentration will be easy using this method while finding the mean may be difficult.

We will prove a somewhat more general version of the bounded differences inequality:

Lemma 19.4

In the above hypothesis, suppose $f\colon \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ satisfies the slightly more generous condition that

$$|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)| \leq c_i$$

when x and y differ only in the ith coordinate (where c_i is some constant). Then $Z = f(X_1, \ldots, X_n)$ satisfies the similar inequality

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \exp\left(-\frac{2\lambda^2}{c_1^2 + \cdots + c_n^2}\right),$$

and likewise

$$\mathbb{P}(Z - \mathbb{E}Z \leq t) \leq \exp\left(-\frac{2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

The proofs of these inequalities will go through *martingales*.

§19.2 Martingales

Martingales are important and basic objects in probability. We will only look at *discrete*-time martingales in this class.

Definition 19.5. A **martingale** is a random real sequence $Z_0, Z_1, \ldots$ such that:

- $\mathbb{E}\,|Z_n| < \infty$ (for technical reasons);
- $\mathbb{E}[Z_{n+1} \mid Z_0, \ldots Z_n] = Z_n$.

Example 19.6

Random walks with independent mean-0 steps are martingales. (Independence is not necessary; but at each step we should have mean 0 conditioned on what we've done before.) Here a *step* is a difference between consecutive terms.

An example more relevant to us is the *Doob martingale*:

Example 19.7

Suppose we have a random variable, and over time we reveal some partial information about this random variable. Then the expected value of the random variable based on the revealed information forms a martingale.

How this will come up for us is the following: given random variables $X_1, \ldots, X_n$ (which do not have to be independent), we have a function $f(X_1, \ldots, X_n)$. Let $Z_i = \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i]$ (so we're taking the expectation of our function, given the information of what the first i variables are). Then Z_0, Z_1, …, Z_n is a martingale.

Here Z_0 is the expectation of f without knowing *any* information — in particular, it's non-random. Meanwhile, Z_n is $f(X_1, \ldots, X_n)$ — so it is a random variable (and in fact, it's the one we wish to understand).

The Doob martingale then allows us to interpolate between a non-random object (the mean) and the random variable we wish to understand via a sequence, defined by slowly revealing information about the inputs until we see the final answer.

A more concrete example that will be relevant to us:

Example 19.8

Consider the random graph $G(n, p)$, and suppose we have some function, the chromatic number. Then $\chi(G(n, p))$ is a random variable. We can try to understand this random variable via a martingale — by revealing the graph in some way. Initially I have the random graph, and I haven't shown it to you. Then you ask me for some information and I show it to you, and this allows you to update your expectation for what the chromatic number should be.

Let's do this specifically in the case $n = 3$ and $p = 1/2$. Then there are 8 possible graphs on 3 vertices, with various values for the chromatic number 3, 2, 2, 2, 2, 2, 2, 1.

In the **edge-exposure martingale**, I reveal the graph to you edge by edge. Initially you don't know anything, other than this is $G(n, p)$. So if we ask the expectation of the chromatic number of $G(n, p)$, we'd average all these numbers and find that it's 2.

Then you ask me to show you what the first edge looks like (the one on the left). I show it to you; either it's present, denoted by an up-edge, or it's not, denoted by a down-edge.

If it's present, then you can update your expectation for the expectation — by averaging out the four possibilities where the edge is present.

Now suppose I show you another edge, the second (on the right). Suppose it's not present. Now you update your expectation, because you're down to two possibilities. Finally, I show you the remaining edge, and at this point you see that the graph has chromatic number 2.

We can complete this whole picture, drawing a branching tree to all of the 8 possible graphs; we'll have some expectation at each node. An instance of this martingale is a walk along this tree — we start at the left, choose one of the two, and keep walking along these branches.

But it's obtained via this edge-exposure procedure, where we're revealing the edges one by one.

We can reveal information in more than one way. Instead of revealing the graph edge-by-edge, we can reveal it vertex-by-vertex — adding a new vertex and seeing what it looks like. In this case, initially you know nothing, so have expectation 2. Now I show you one of the vertices. You still have no information about any of the edges, so you still have expectation 2. Now I show you one more vertex, and all the edges between the shown vertices. Now at this point you update your expectation to 2.25. Finally, I show you the third and final vertex, and you can figure out what the answer is..

This is called the *vertex-exposure* martingale. Whereas the edge-exposure martingale has $\binom{n}{2}$ steps, the vertex-exposure martingale has n steps (or $n-1$ if you don't worry about the first step, which is trivial). Both are useful for different applications; we'll see some examples of each.

Eventually we'll come back to the proof of the bounded differences inequality, using concentration inequalities about martingales.

§19.3 Azuma's Inequality

Theorem 19.9

Suppose Z_0, …, Z_n is a martingale with the hypothesis that

$$|Z_{i+1} - Z_i| \le 1$$

for all i (so our step sizes never exceed 1). Then

$$\mathbb{P}(|Z_n - Z_0| \ge \lambda\sqrt{n}) \le e^{-\lambda^2/2}.$$

In a martingale, we're taking a bunch of steps. If you take n steps, in principle you could wander off by a distance of n. But Azuma's inequality tells you that's very unlikely — you do not expect to wander away by more than $O(\sqrt{n})$.

We also have some extensions and variations. For proving the version of bounded differences allowing different bounds on different coordinates, we'll change this slightly:

Theorem 19.10

Suppose Z_0, …, Z_n is a martingale with the hypothesis that

$$|Z_i - Z_{i-1}| \le c_i$$

for all i. Then

$$\mathbb{P}(|Z_n - Z_0| \ge \lambda) \le e^{-\lambda^2/2(c_1^2+\cdots+c_n^2)}.$$

We don't really care about the constant 2, but in order to get it, we'll prove a slight strengthening (which we can use to deduce the theorem just stated):

Theorem 19.11

Suppose Z_0, …, Z_n is a martingale such that for all i, conditioned on $(Z_0, \dots, Z_{i-1})$, Z_i lies inside an interval of length c_i (the location of this interval may depend on the interval revealed so far). Then

$$\mathbb{P}(|Z_n - Z| \ge \lambda) \le e^{-\lambda^2/2(c_1^2+\cdots+c_n^2)}.$$

Remark 19.12. If we don't care about the factor of 2 (which we are encouraged to ignore), then we shouldn't worry about the differences between these two versions. (To match the condition on the right, we should think of c_i as $c_i/2$; the point on the right is that the intervals don't have to be centered around the mean.)

The proof is in a way analogous to the proof of the Chernoff bound that we saw earlier — we will look at the exponential generating function and apply Markov. When we do this calculation, at some point we may need to estimate a certain exponential generating function — what happens when we take a single step. We'll isolate this piece of the proof:

Lemma 19.13 (Hoeffding's Lemma)

If X is a real random variable contained in an interval of length ℓ (where ℓ is some constant), and $\mathbb{E}X = 0$, then

$$\mathbb{E}e^X \le e^{\ell^2/8}.$$

We saw something like this when proving the Chernoff bound — we looked at e^X where X was the ±1 random variable, and this is exactly what happened. This is a slightly more general claim, and the proof is to consider the function e^X, and we have some interval where our random variable could lie; and we want to upper bound the mean. By convexity, it's upper-bounded by replacing e^X with a line (it's maximized if all the mass is concentrated at both ends). Then we just have to check that one case, and that's an inequality which we will not do in class.

With Hoeffding's Lemma, we are now ready to prove Azuma's inequality.

Proof of Azuma's inequality. Assume $Z_0 = 0$ for convenience (otherwise we could shift by a constant), and let $X_i = Z_i - Z_{i-1}$ be the size of the ith step. Then we know that

$$\mathbb{E}[X_i \mid Z_0, \dots, Z_{i-1}] = 0,$$

and we also know that conditioned on $Z_0, \dots, Z_{i-1}$, X_i lies in an interval of length c_i. Now let's look at the moment generating function

$$\mathbb{E}e^{tZ_n} = \mathbb{E}e^{t(X_n + Z_{n-1})}.$$

We can then split our expectation into two parts, as

$$\mathbb{E}[\mathbb{E}[e^{t(X_n)} \mid Z_0, \dots, Z_{n-1}] \cdot e^{tZ_{n-1}}].$$

(The term $e^{tZ_{n-1}}$ only depends on the first $n-1$ terms, so we hold them fixed and see how the last varies.)

For the inside expectation, we can use Hoeffding's lemma to get that

$$\mathbb{E}[e^{tX_n} \mid Z_0, \dots, Z_{n-1}] \le e^{t^2 c_n^2/8}.$$

This is true no matter what the first $n-1$ terms are — the location of this interval might vary depending on what we saw in the first few steps, but Hoeffding's lemma applies no matter where the interval is.

With that, we can collect everything and get the bound

$$EEe^{tZ_n} \le e^{t^2 c_n^2/8}\mathbb{E}e^{tZ_{n-1}}.$$

We can keep going (or apply induction), and we get

$$\mathbb{E}e^{tZ_n} \le \exp\left(-\frac{t^2}{8}(c_1^2 + \cdots + c_n^2)\right).$$

We've now managed to find an upper bound to our exponential generating function, and now we can apply Markov's inequality to estimate the tail. Applying Markov, we find that

$$\mathbb{P}(Z_n \ge \lambda) \le e^{-t\lambda}\mathbb{E}e^{tZ_n} \le e^{-t\lambda + t^2/8 \cdot (c_1^2 + \cdots + c_n^2)}$$

(for $t > 0$). Then setting

$$t = \frac{4\lambda}{c_1^2 + \cdots + c_n^2}$$

gets the desired bound of

$$\mathbb{P}(Z_n \ge \lambda) \le \exp\left(-\frac{2\lambda}{c_1^2 + \cdots + c_n^2}\right). \qquad \square$$

§19.4 Proof of the Bounded Differences Inequality

We'll deduce the bounded difference inequality from Azuma's inequality.

Proof of bounded differences inequality. The idea is to set up a martingale so that the initial term is $\mathbb{E}Z$, and the final term is Z. We'll do this through the Doob martingale: consider the Doob martingale where

$$Z_i = \mathbb{E}[Z \mid X_1, \ldots, X_i].$$

First, let's see the role played by the Lipschitz condition. For any fixed x_1, ..., x_{i-1}, we want to consider

$$\mathbb{E}f(x_1, \ldots, x_i, X_{i+1}, \ldots, X_n)$$

(where we feed in the first i coordinates, but let the remaining ones be random). We can see that if we changed the ith input to a different value x_i', then

$$\mathbb{E}f(x_1, \ldots, x_i', X_{i+1}, \ldots, X_n)$$

differs from the previous expression by at most c_i (since this is true for any values of X_1, ..., X_n).

This means that conditioned on knowing the first $i-1$ inputs X_1, ..., X_{i-1}, Z_i lies in an interval of length c_i. This is exactly the setup that we need for applying Azuma's inequality, so then Azuma gives us that

$$\mathbb{P}(Z - \mathbb{E}Z \geq \lambda) \leq \exp\left(\cdots\right),$$

which is precisely the expression we were trying to prove. □

Student Question. *WHy does the expected value inequality hold?*

Answer. For all inputs the inequality is true. Now if we hold the first i variables fixed and let the remainders be random, then since it's true pointwise, it's also true in expectation.

Remark 19.14. The right way to talk about martingales is filtrations — as you reveal more and more of the probability space, that defines the martingale. It's not just about revealing the previous terms, but all the information.

When we talked about martingales, we talked about conditioning on Z_1, ..., Z_{n-1}. But $\mathbb{E}[Z \mid Z_1, \ldots, Z_{n-1}]$ is not the same as $\mathbb{E}[Z \mid X_1, \ldots, X_{i-1}]$, since it's possible that multiple X_i give you the same sequence and you can't recover X_i from the Z's alone.

This can be fixed by conditioning on the revealed information instsead of the Z's in the Azuma proof, which is how you're actually supposed to do things when you talk about martingales.

For now, we just need the bounded difference inequality. For certain applications, we will have to go back to the martingale version — sometimes that's genuinely more powerful. But for many applications, it's simpler and sufficient to just apply the bounded difference inequality, and that's what we'll do next.

§19.5 Chromatic number of $G(n, p)$

Last time, we saw the landmark theorem of Bollobás that

$$\chi(G(n, 1/2)) \sim \frac{n}{2\log_2 n}$$

with high probability. But historically, before this result, it was proved that the chromatic number concentrates; and that's what we'll see now.

Theorem 19.15 (Shamir–Spencer 1978)

For every λ, $Z = \chi(G(n,p))$ satisfies that

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \lambda\sqrt{n-1}) \leq 2e^{-2\lambda^2}.$$

In other words, we have a concentration window of length $O(\sqrt{n})$ for the chromatic number of a random graph.

Proof. We'll apply the bounded differences inequality.

First let's do something that doesn't quite work, to show us the choices to be made. Suppose we think of $G(n,p)$ as a collection of $\binom{n}{2}$ edge-indicator variables, so then Z is a function of these $\binom{n}{2}$ edge-indicator variables. Changing one edge changes the chromatic number by at most 1 — if we add in one edge, our previous coloring might not work, but we can always introduce a new color. Then applying the bounded difference inequality (in the form without the c_i) tells us that

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-2t^2/\binom{n}{2}}.$$

This is not quite good enough — it gives us a window of width n, which is a pretty terrible result because the chromatic number is at most n.

SO instead, what we want to do is group the edges in a more clever way. What we want to do is use the idea of the *vertex exposure* martingale. Define a function $f\colon \Omega_2 \times \cdots \times \Omega_n$ where $E_i = \{(j,i) \mid j < i\}$ and $\Omega_i = \{0,1\}^{E_i}$. So instead of having a function that looks edge-by-edge, we group these edges together — where X_1 corresponds to whether the purple thing is an edge, X_2 corresponds to the configuraiton in the orange edges, X_3 corresponds to the configuration in the three blue edges, and so on. Of course, f is supposed to be the chromatic number of the resulting graph.

Now it's still true that changing one coordinate corresponds to changing edges only incident to one fixed vertex. So the chromatic number changes by at most 1. (You could have kept the same coloring, or used a new color for just that vertex.)

So now we are back to the same situation, except that instead of having a situation with $\binom{n}{2}$ input variables, we now have $n-1$ input variables. Everything else then works in exactly the same way, and now we do get the result that is claimed. This finishes the proof. □

We see that we are applying bounded differences, but in the first attempt we had too many variables, so we used a different organization of how to input variables into the function so that there's fewer variables, but bounded difference still applies.

There will be other problems where we *don't* want to do vertex exposure, and we want edge exposure instead. This is because for some other problems, this change might be too much. There will also be problems where the choice of the function is actually non-obvious.

We've now shown that the chromatic number concentrates in a window of width $\sqrt{n}$. We can ask whether this is actually true — the clique number had $O(1)$ concentration, which is much smaller. So potentially the window could be muvh smaller than $O(\sqrt{n})$. On a started problem, you will show that it's at most $\sqrt{n}/\log n$ — a small improvement of this upper bound. For a long time it was a major upper bound whether you have muvh better concentration. Many people believed that like the clique number, the concentration window should be over a constant-sized window. (There's a story where a mathematicial presented a talk claiming to have proved that; it turns out that was wrong.)

A recent breakthrough of Heckel 2022 says that the concentration window for $\chi(G(n,1/2))$ is wider than $n^{1/4-\varepsilon}$ — so it is *not* true that the concentration is less than $n^{1/4}$. So we do have an actual amount of variance in χ, in sharp contrast to the problem of the clique number (where there's 2-point concentration).

We don't actually understand the truth; what seems to be believed now is that the window fluctuates between $n^{1/4}$ and $\sqrt{n}$ depending on various properties of n.

One thing we'll see next time is that in $G(n, p)$ when p gets to decrease sufficicently quickly as a function of n, then we have much sharper concentration — we will show four-point concentration. So different regimes behave very differently, and some of these are difficult results.

§20 November 14, 2022

Last class, we started talking about concentration of measure, and we stated and proved an inequality about bounded differences. We'll state the most basic version, because that's the version we'll use today; but later on we'll use the more general versions from last time.

Theorem 20.1 (Bounded Differences Inequality)
Suppose $X_1 \in \Omega_1, \ldots, X_n \in \Omega_n$ are independent random variables, and suppose $f\colon \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ is a function such that changing one coordinate of the input to f changes its value by at most 1 (in absolute value). Then the random variable $Z = f(X_1, \ldots, X_n)$ satisfies the tail bounds

$$\mathbb{P}(Z - \mathbb{E}Z \geq \lambda) \leq e^{-2\lambda^2/n}$$

for all $\lambda \geq 0$.

There's also a similar lower tail form, which can be deduced from the upper tail form by replacing f with $-f$.

Last time, we stated this theorem, and we proved it using Azuma's inequality, which says that if we have a *martingale* with bounded steps, then the nth term of the martingale satisfies the same inequality.

> **Remark 20.2.** Last class, we weren't so careful with something: we stated the Azuma inequality allowing for different setp bounds c_i, and we stated the hypothesis that each step lies in an interval of size at most c_i depending on the previous values. The issue is that when we use this to prove the bounded differences inequality, it could be that several values of X may lead to the same Z. The easy way to fix this isi to do the proof directly for the Doob martingale. We should not worry about this; the theorems are correct, and we don't need to worry about the constants.

§20.1 Chromatic Number

Last time, we saw one application of the bounded differences inequality to chromatic numbers:

Theorem 20.3 (Shamir–Spencer 1987)
If $Z = \chi(G(n, p))$, then

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \lambda\sqrt{n-1}) \leq 2e^{-2\lambda^2}.$$

The proof was via applying the bounded differences inequality to vertex exposure — we consider the function $f\colon \Omega_1 \times \cdots \times \Omega_{n-1}$ where Ω_i describes which of the edges with right endpoint i are used, which outputs the chromatic number. Then f has the property that changing one coordinate of f (which only changes the edges around one vertex) cannot affect the chromatic number by more than 1, and therefore the chromatic number of the *random* graph satisfies this tail bound.

This result came before Bollobás's theorem from 1988 that

$$\chi(G(n,1/2)) \sim \frac{n}{2\log_2 n}$$

with high probability. We proved this theorem using Janson's inequality. In particular, a key step in the solution was the following claim:

Lemma 20.4

If $k_0 = k_0(n) \sim 2\log_2 n$ is the largest integer such that $\binom{n}{k_0} \cdot 2^{-\binom{k_0}{2}} \geq 1$, then

$$\mathbb{P}(\omega(G(n,1/2)) < k_0 - 3) = e^{-n^{2-o(1)}}.$$

This theorem says that with very high probability, we can find a clique of size around $2\log_2 n$. Once we have that, then we can look at $G(n,1/2)$ and keep plucking out large independent sets (cliques and independent sets are essentially the same) until we have very few vertices remaining, and we can then color those vertices each with their own color.

We proved this result using Janson's inequality. We'll now see a different proof of this lemma; this was Bollobás's original proof, and it goes through the bounded differences inequality.

When we used the second moment method, the approach was to count the number of cliques of the size that we want, look at its expectation and variance, and show that the number of cliques is 0 with low probability. But it *doesn't* work to try this again here — to set our random variable Z to be the number of k-cliques, where $k = k_0 - 3$, and attempt to show that this random variable is 0 with low probability. So instead, we'll need a more clever choice of a random variable.

Proof. Let $Y = Y(G)$ be the maximum size of an *edge-disjoint* collection of k-cliques in G. The reason we care about this is that at the end of the day, we don't care about how many k-cliques there are; we only care about whether there *is* a k-clique or not. So instead of counting k-cliques, we'll see whether we can fit many k-cliques in G. This behaves more 'smoothly' with G, but at the same time captures what we want to capture — $Y = 0$ is the same as $\omega(G) < k$.

The reason we might not want to consider Y is that unlike the number of k-cliques, there isn't such an easy way to compute $\mathbb{E}Y$ via linearity of expectations. It's actually not too bad; we did a homework problem about finding the number of edge-disjoint triangles, and a similar idea works here.

By the bounded differences inequality, where $G \sim G(n,1/2)$, we have

$$\mathbb{P}(\omega(G) < k) = \mathbb{P}(Y = 0) = \mathbb{P}(Y - \mathbb{E}Y \leq -\mathbb{E}Y).$$

We can then use the bounded differences inequality to bound this event. We need to check the hypothesis of the bounded differences inequality. If we change one edge, then Y can change by at most 1 — if we destroy one edge, then we lose at most one k-clique (we may be able to rearrange and not lose any, but we definitely lose at most 1). So bounded differences gives us the tail bound

$$\mathbb{P}(Y - \mathbb{E}Y \leq -\mathbb{E}Y) \leq \exp\left(-\frac{2(\mathbb{E}Y)^2}{\binom{n}{2}}\right).$$

(This is why we want to consider edge-disjoint k-cliques rather than trying to count all k-cliques; if we did count, then Y would change much more.)

Our goal is to show that this is nearly quadratic in the exponent, so we want to show that $\mathbb{E}Y \geq n^{2-o(1)}$. This now looks very much like the homework problem we did, of showing that $G(n,p)$ has lots of edge-disjoint triangles.

We'll now see two different approaches for proving that typically G has lots of edge-disjoint k-cliques.

First let's create an auxiliary graph $\mathcal{H}$ whose vertices are the set of k-cliques in G, and whose edges connect pairs of edge-overlapping k-cliques. Then we have $Y = \alpha(H)$. At this point, there are a couple of approaches.

One approach is to use the Caro–Wei inequality, which we proved in the chapter on linearity of expectations, and which gives us the lower bound

$$\alpha(H) \geq \sum_{v \in V(\mathcal{H})} \frac{1}{1 + d_{\mathcal{H}}(v)}.$$

By convexity, we can then write

$$\alpha(H) \geq \sum_{v \in V(\mathcal{H})} \frac{1}{1 + d_{\mathcal{H}}(v)} \geq \frac{|V(\mathcal{H})|}{1 + \overline{d}} = \frac{|v(\mathcal{H})|^2}{|v(\mathcal{H})| + 2\,|E(\mathcal{H})|}.$$

(Here $\overline{d}$ is the average degree.)

Another approach is the following: let H' be the graph obtained from H by keeping every vertex with probability q. Then we can use the cheap bound $\alpha(\mathcal{H}') \geq |V(\mathcal{H}')| - |E(\mathcal{H}')|$. On the other hand, any independent set we find in $\mathcal{H}'$ is also an independent set in $\mathcal{H}$, which means

$$\alpha(\mathcal{H}) \geq |V(\mathcal{H}')| - |E(\mathcal{H}')|.$$

Taking the expectation of both sides, we have

$$\mathbb{E}\,|V(\mathcal{H}')| - |E(\mathcal{H}')| = q\,|v(\mathcal{H})| - q^2\,|E(\mathcal{H})|.$$

This is true for all $q \in [0, 1]$, so provided that $|E(H)| \geq |v(H)|/2$, we can take

$$q = \frac{|v(\mathcal{H})|}{2E(\mathcal{H})}$$

to get that

$$\alpha(\mathcal{H}) \geq \frac{|v(H)|^2}{4\,|E(H)|^2}.$$

We want to show that the independence number is large, and we have some tools for getting lower bounds on the independence number in terms of some graph statistics. These statistics have come up a couple of times: we know that

$$|v(\mathcal{H})| \sim \mathbb{E}\,|v(\mathcal{H})| = \binom{n}{k} 2^{-\binom{k}{2}} = n^{3-o(1)}.$$

We also know that

$$\mathbb{E}\,|E(\mathcal{H})| = n^{4-o(1)}$$

by a somewhat hairy second moment calculation (that we've omitted a few times in class). Now plugging everything into either of our lower bounds on the independence number, we have

$$\mathbb{E}Y \gtrsim \mathbb{E}\frac{v(\mathcal{H})}{E(\mathcal{H})} \gtrsim \mathbb{E}\frac{n^{6-o(1)}}{|E(H)|}.$$

(For example, by considering when the number of vertices is within a factor of 2 from its expectation.) We can use a convexity argument to bring the expectation to the denominator, which then gives us

$$\geq \frac{n^{6-o(1)}}{\mathbb{E}\,|E(H)|} = \frac{n^{6-o(1)}}{n^{4-o(1)}} = n^{2-o(1)}.$$

□

This gives us a different proof of the key step in our original proof. The key point is that we are not looking at the number of k-cliques — that's too expensive from teh point of view of the bounded differences inequality, since it can change a lot. Instead, we look at the maximum number of *edge-disjoint* k-cliques. Even though this quantity is harder to work with, it doesn't change very much when we change one edge — it has a Lipschitz property that's more amenable to bounded differences. Then we plugged into the bounded differences inequality and showed that $\mathbb{E}Y$ is large, so $Y = 0$ with large probability.

§20.2 Chromatic Number for Sparse Graphs

It turns out that when p gets smaller, we actually have much better concentration inequalities.

Theorem 20.5 (Shamir–Spencer 1987)
Fix a constant $\alpha > 5/6$. Then for $p < n^{-\alpha}$, $\chi(G(n,p))$ is concentrated at 4 values with high probability — in other words, there exists some u (depending on n and p) such that

$$\mathbb{P}(u \le \chi(G(n,p)) \le u+3) \to 1$$

as $n \to \infty$.

So when p decreases rapidly, the chromatic number of $G(n,p)$ has much better concentration than shown by the previous theorem — it's concentrated on a finite set of values.

We will still use the vertex-exposure martingale, but if we apply it naively, we only get $\sqrt{n}$ concentration. So we will do something more clever.

Proof. By standard ε–δ nonsense, it suffices to prove that for all ε, there exists u depending on n, p, and ε, so that provided $p < n^{-\alpha}$ and n is sufficiently large,

$$\mathbb{P}(u \le \chi(G(n,p)) \le u+3) \le 1 - 3\varepsilon.$$

If we prove that this probability can be made arbitrarily close to 1, then letting ε decay very slowly as a function of n (where 'very slowly' depends on what sufficiently large means), we will get the 'with high probability' statement.

Define u to be the *least* integer such that

$$\mathbb{P}(\chi(G(n,p)) \le u) > \varepsilon.$$

If we let u be really large, then this probability will get very large (up to 1); so we can start with $u = 0$ and increase by 1 until the probability exceeds ε for the first time, and that's the u we choose. (It exists just by monotonicity, but it gives no information on what u is as a function of n, p, and ε.)

Now we make a very clever choice of a random variable to apply the bounded differences inequality. Let $G \sim G(n,p)$, and let $Y = Y(G)$ be the minmum size of a subset $S \subseteq V(G)$ such that $G \setminus S$ is u-colorable. (u is some number, say 100; the graph may not be 100-colorable, but if we remove some vertices we can make it 100-colorable, and we define Y to be the minimum number of vertices we'd need to remove.)

First note that Y changes by at most 1 if we change the edges around a single vertex of G — similarly to what happened earlier in the vertex exposure proof. (If we only change one vertex, then we might want to include or exclude the vertex, but this won't affect the size by more than 1, even if we get to make alternate choices.)

This is useful because we can now use the bounded differences inequality with respect to vertex exposure, and obtain the tail bounds

$$\mathbb{P}(Y \le \mathbb{E}Y - \lambda\sqrt{n}) \le e^{-2\lambda^2}$$

and

$$\mathbb{P}(Y \geq \mathbb{E}Y + \lambda\sqrt{n}) \leq e^{-2\lambda^2}.$$

We will use both bounds in succession, and for different purposes.

Choose $\lambda = \lambda(\varepsilon)$ such that $e^{-2\lambda^2} = \varepsilon$. First we consider the event $\mathbb{P}(Y = 0)$. If $Y = 0$, then G itself is already u-colorable, so

$$\mathbb{P}(Y = 0) = \mathbb{P}(\chi(G) \leq u) > \varepsilon = e^{-2\lambda^2}.$$

On the other hand, rewriting $Y = 0$ as $Y \leq \mathbb{E}Y - \mathbb{E}Y$ and applying the lower tail form of the bounded differences inequality, we get

$$e^{-2\lambda} = \varepsilon < \mathbb{P}(Y = 0) \leq e^{-2(\mathbb{E}Y)^2/n}.$$

Comparing the left and right hand sides, we find an upper bound on $\mathbb{E}Y$ — we get

$$\mathbb{E}Y \leq \lambda\sqrt{n}.$$

So we used the lower tail bounded differences inequality to get an upper bound on $\mathbb{E}Y$.

Now we apply the upper tail bounded differences inequality to show that Y is rarely large. We have

$$\mathbb{P}(Y \geq 2\lambda\sqrt{n}) \leq \mathbb{P}(Y \geq \mathbb{E}Y + \lambda\sqrt{n}) \leq e^{-2\lambda^2} = \varepsilon,$$

by applying the bounded differences inequality. This gives us some useful information about Y — it tells us that typically, Y is at most a constant times $\sqrt{n}$.

To recap what's happened so far, each of the following events occurs with probability at least $1 - \varepsilon$:

(1) There exists some $S \subseteq V$ with $|S| \leq 2\lambda\sqrt{n}$ such that $G \setminus S$ is u-colorable. (This is by the complement of the above event.)

(2) $\chi(G) \geq u$. (This is by the choice of u that we started with.)

(3) The induced subgraph $G[S]$ is 3-colorable.

For (3), we'll prove the following lemma:

Lemma 20.6

Fix $\alpha > 5/6$ and a constant C, and let $p < n^{-\alpha}$. Then with high probability, every subset of at most $C\sqrt{n}$ vertices in $G(n, p)$ is 3-colorable.

These three events together imply that $\chi(G) \geq u$, and $\chi(G) \leq u + 3$ — we can use u colors to take care of $G \setminus S$, and 3 new colors to take care of S.

So it remains to show this lemma — that if we draw $G(n, p)$, then with high probability every subset of at most $C\sqrt{n}$ vertices is 3-colorable.

Proof. If G is *not* 3-colorable, then we can choose a minimum-size subset $T \subseteq V(G)$ such that the induced subgraph $G[T]$ is not 3-colorable (which must always exist).

Claim — $G[T]$ has minimum degree at least 3.

Proof. If it had a vertex of degree 1 or 2, that wouldn't affect 3-colorability, so we could get rid of it. ■

So now the question is whether we can find a subgraph of $G(n, p)$ with minimum degree at least 3 — and it turns out that if p is sparse enough, then that isn't possible.

More precisely, we can bound the probability that $G(n, p)$ has a subgraph on $t \leq C\sqrt{n}$ vertices with at least $3t/2$ edges (meaning with average degree at least 3). We'll do this using a union bound — this is at most

$$\sum_{t=4}^{C\sqrt{n}} \binom{n}{t} \binom{\binom{t}{2}}{3t/2} p^{3t/2}$$

(looking at all possible induced subgraphs, all the places we could possibly put $3t/2$ edges, and finding the probability those occur). This expression is $o(1)$ if $\alpha > 5/6$ (by routine computation). Now we see that $G(n, p)$ typically does not have a subgraph on at most $C\sqrt{n}$ vertices with average degree at least 3, which proves the lemma. ■

Remark 20.7. 3-colorability is hard (it's a classic NP-hard problem). So when you see 3-colorability, you should either run away or do something naive. And that's what we do — we just look at a 3-core (which is very simple graph theoretically), because really understanding 3-colorability is a tall challenge.

This finishes the proof that $G(n, p)$ is 4-point concentrated. □

Remark 20.8. This is not the best thing we know. In fact, the original paper proved something slightly better. Thanks to further work, we now know that $G(n, n^{-\alpha})$ has 2-point concentration for all $\alpha > 1/2$.

Remark 20.9. Last time we mentioned that when p is constant (or slowly decreasing), we do not have 2-point concentration. A recent result of Heckel shows that $G(n, 1/2)$ is not more sharply concentrated than $n^{1/4}$. It is not well-understood what happens when $n^{-\alpha}$ is between 1 and $1/\sqrt{n}$.

Remark 20.10. We do not know the value of these two points — but a paper looks at what happens when $p = c/n$, and there it's been solved what the location of the chromatic number should be. But that's a difficult result, and in general we do not know the answer.

In the discussion on the second moment chapter, we had a lecture on thresholds. One of the questions was what we can say about the threshold for 3-colorability, or k-colorability in general. It's a sharp threshold — we know this by advanced tools — but we don't know the location. We know it's on the order of $1/n$, but we don't know the location. That's related to not knowing the value of the chromatic number here.

§20.3 Isoperimetric Inequalities

So far, what we've seen using the bounded difference inequalitiy is in the style of Chernoff bounds. But there is also a geometric story which is quite important.

To explain what's going on there, we'll talk about something very classical: isoperimetric inequalities.

Question 20.11. Given a subset $A \subset \mathbb{R}^n$ of fixed volume, how do we minimize its surface area?

'Isoperimetric' usually refers to the question the other way around (for fixed perimeter, how do you maximize the area)? The answer is to take a disk or ball. This can be made precise:

Definition 20.12. In a metric space X with metric d_X, if we have a set $A \subseteq X$, let

$$A_t = \{x \in X \mid d_X(x, A) \leq t\}$$

be the t-neighborhood of A.

Here $d_X(x, A)$ denotes the minimum distance from x to some point in A — so we take our shape, and we grow it out by a distance t.

Theorem 20.13 (Isoperimetric Inequality in Euclidean Space)
Let $A \subseteq \mathbb{R}^n$, and let $B \subseteq \mathbb{R}^n$ be a ball such that $\operatorname{Vol} A = \operatorname{Vol} B$. Then for all $t \geq 0$,

$$\operatorname{Vol} A_t \geq \operatorname{Vol} B_t.$$

(All sets in this discussion will be measurable.)

Remark 20.14. This may not be of the form we recognize — which has to do with surface area or perimeter. That follows from this formuation — the surface area (or surface volume), or the volume of the boundary of A, is at least the volume of the boundary of B. The reason this is true is that we can write

$$\operatorname{Vol}_{n-1}(\partial A) = \frac{d}{dt}|_{t=0} \operatorname{Vol}_n(A_t).$$

(As t starts from 0, if we grow it a little bit, then how quickly we grow initially is the surface area. You can use this to derive the surface area of a sphere from its volume, or vice versa.) By calculus, this is

$$\lim_{t \to 0} \frac{\operatorname{Vol}(A_t) - \operatorname{Vol}(A)}{t}.$$

We can then imagine the same expression with A replaced by B. Then we have an inequality term-wise (for each t), and as $t \to 0$ we hvae the inequality we're looking for.

Remark 20.15. If we only had the inequality for the boundary, then we could integrate to get the original (the t-neighborhood of a ball is still a ball). So the two formulations are actually equivalent.

The reason we discuss this is that there's a similar isoperimetric inequality in many other settings. In particular, there is *Harper's theorem*, which gives an isoperimetric inequality on the hypercube (and which we saw earlier). This is not the most general case, but it will be enough for our use.

Theorem 20.16 (Harper's Theorem)
Consider the Boolean cube $\{0, 1\}^n$ with the Hamming distance. If $A \subseteq \{0, 1\}^n$ is a subset of the unit cube, and B is a Hamming ball (everything within a certain Hamming distance of a single point), and if $|A| \geq |B|$, then for all $t \geq 0$, we have $|A_t| \geq |B_t|$.

We're not going to prove Harper's theorem. There is an even more precise statement if we're given the size of A — the answer is that we start building a Hamming ball, and then put the rest of the points in in lexicographical order.

But the reason to see this theorem is to see the connection between this result and what we just did — namely the bounded differences inequality. And for any consequence of Harper's theorem (where we don't care about constants), we can deduce it from the bounded differences inequality — so the bounded differences inequalitiy cna be viewed as analogous to isoperimetry.

The slogan of this section is that isoperimetry and concentration inequalities are really two sides of the same coin. Let's first see some easy facts.

Theorem 20.17
For every $t > 0$ and $A \subseteq \{0,1\}^n$, suppose that $|A| \geq 2^{n-1}$. Then $|A_t|/2^n \geq (1 - e^{-2t^2/n})$.

Proof. Let $B = \{x \in \{0,1\}^n \mid \mathrm{wt}(x) < n/2\}$ (i.e. elements with more 0's than 1's). Then we see $|B| \leq 2^{n-1} \leq |A|$. So Harper's theorem tells us that

$$|A_t| \geq |B_t|,$$

because B is a Hamming ball around the 0 vector. But we have a precise description of B_t — it's the set of all binary vectors with $\mathrm{wt}(x) < n/2 + t$. And we can use the Chernoff bound to bound vertices beyond this threshold; we then find a bound of

$$|A_t| > (1 - e^{-2t^2/n})2^n.$$

□

What this says is that if we start with half of the cube, the cube has diameter n. So if we expand by 1% of n, then we get almost everything.

So if we start by $1/2$ of the cube and want to get 99%, we only need to expand by $\sqrt{n}$. This is a high-dimensional phenomenon that w'ell explore more in upcoming lectures.

§21 November 16, 2022

Today we will continue to explore the concentration of measure phenomenon. In particular, we'll see the connection between probability and geometry.

This phenomenon, which is really about something happening in high dimensions, was stressed by Millman in the 1970s — he stressed the importance of the techniques and perspectives we'll get a glimpse of today.

In particular, we've been looking at concentration of Lipschitz functions of random variables. It turns out that this topic is related to isoperimetric inequalities in geometry — and in a way, these are two sides of the same coin.

This is a very important topic, in that it comes up all over the place — it comes up in combinatorics, probability, and theoretical computer science, but it also comes up in a lot of other areas. On one hand, there's *high-dimensional probability*, and on the other hand (in the circle around Millman), you may also hear the terms *asymptotic geometric analysis*. In a way they refer to different problems, or different angles; but at their core they're really about the same things.

§21.1 A Recap

Definition 21.1. In a metric space (X, d_X), for a subset $A \subseteq X$, we define the **radius-t neighborhood** of A as

$$A_t = \{x \in X \mid d_X(x, A) \leq t\}.$$

Today we'll mostly be working over probability spaces with a metric.

The classic isoperimetric inequality in $\mathbb{R}^n$ says the following:

Theorem 21.2

If $A \subseteq \mathbb{R}^n$ is measurable, and $B \subseteq \mathbb{R}^n$ is a ball with $\operatorname{Vol} A = \operatorname{Vol} B$, then

$$\operatorname{Vol} A_t \geq \operatorname{Vol} B_t.$$

Taking the derivative at 0, we essentially get the boundary volume. It's a classic fact that among all bodies of the same volume, the ball has the smallest surface area; that is equivalent to this version.

(This result was known to the Greeks, though the actual proof didn't come until 1900.)

Now let's move to the cube. On the cube, we're looking at the set $\{0,1\}^n$ together with the Hamming distance metric. (The cube is also equipped with a probability measure, where we take a uniform random point on the cube.) Then Harper's theorem gives an isoperimetric inequality on the boolean cube:

Theorem 21.3 (Harper's Isoperimetric Inequality)

If $A, B \subseteq \{0,1\}^n$ such that B is a Hamming ball and $|A| \geq |B|$, then for all $t \geq 0$, we have $|A_t| \geq |B_t|$.

This is analogous to what happens in the Euclidean case, but now instead of a Euclidean ball, we're looking at a Hamming ball.

One consequence of this result, which we showed at the end of last lecture, si a claim about what happens if we start with half the ball:

Theorem 21.4 (Rapid expansion from $1/2$ to $1-\varepsilon$)

If $A \subseteq \{0,1\}^n$ with $|A| \geq 2^{n-1}$, then $|A_t| \geq (1 - e^{-2t^2/n}) \cdot 2^n$.

The diameter of the space is n, so if we want to expand from $1/2$ of the cube to 99% of the cube, we only need to expand by a radius on the order of $\sqrt{n}$ — even if we expand just a little bit, as long as we start with half of the cube, we'll end up covering everything.

This is a rapid expansion phenomenon that we derived quickly from Harper's isoperimetric inequality. Much of today will be about an elaboration of this idea, and connecting it to what we did earlier about concentrations of Lipschitz functions.

§21.2 Rapid Expansion

If we start with half the space and expand a bit, we get almost everything. It turns out the same is true even if we start out with 1% of the cube:

Theorem 21.5 (Rapid expansion from ε to $1-\varepsilon$)

Let $\varepsilon > 0$, and let $C = \sqrt{2\log 1/\varepsilon}$. If $A \subseteq \{0,1\}^n$ with $|A| \geq \varepsilon \cdot 2^n$, then $\left|A_{C\sqrt{n}}\right| \geq (1-\varepsilon)2^n$.

So even if we start with 1% of the cube, if we expand on the order of $\sqrt{n}$, then we get almost everything. This is the kind of high-dimensional phenomenon that may be kind of counterintuitive, especially if we're used to thinking about a cube in low dimensions.

But in a way, this is not new to us. We'll give two proofs — one using Harper's isoperimetric inequality, and a second using the bounded differences inequality. Through these two proofs, hopefully we will see that these two subjects are connected.

Proof 1 via Harper's isoperimetric inequality. Let $t = \sqrt{\log(1/\varepsilon)/2 \cdot n}$ so that $e^{-2t^2/n} = \varepsilon$. Now we'll apply (*). We start with a small set A, and we want to show it'll expand rapidly to the entire set. We'll do this in two steps — we'll first show it expands rapidly to half the set, and then expand that rapidly to almost everything.

First let's suppose A doesn't expand very much. Then we can look at the complement $A' = \{0,1\} \setminus A_t$. Then when we expand the complement A' by t, we have a width-t gap, so when we expand A' by t, we are not going to hit A — in particular, we are not going to hit almost everything, because A has volume at least ε. This would contradict the theorem unless A' is less than half the space — so we see that $|A'| \leq 2^{n-1}$ (or else A'_t would intersect A, which is not possible). So t-expansion first gets us at least half the space.

Then once we have half the space, we can apply the theorem again to get that

$$|A_{2t}| \geq (1-\varepsilon)2^n.$$ □

Proof 2 via Bounded Differences Inequality. Pick a point $x \in \{0,1\}^n$ uniformly at random, and consider the random variable $Z = \operatorname{dist}(x, A)$. (Here the distance means the Hamming distance.)

Then Z changes by at most 1 if one coordinate of x changes. So this gives us the setting to apply the bounded differences inequality — this tells us

$$\mathbb{P}(Z - \mathbb{E}Z \leq -\lambda) \leq e^{-2\lambda^2/n} \text{ and } \mathbb{P}(Z + \mathbb{E}Z \geq \lambda) \leq e^{-2\lambda^2/n}.$$

Now consider the event that $Z = 0$ — meaning that the distance from x to A is 0, or in other words $x \in A$. We know from the hypothesis that $\mathbb{P}(x \in A) \geq \varepsilon$, but on the other hand setting $\lambda = \mathbb{E}Z$, we get an upper bound

$$\varepsilon \leq \mathbb{P}(x \in A) = \mathbb{P}(Z = 0) \leq \exp\left(-2(\mathbb{E}Z)^2/n\right).$$

This tells us $\mathbb{E}Z$ is not very large — more precisely

$$\mathbb{E}Z \leq \sqrt{\frac{\log(1/\varepsilon)}{2}n} = \frac{C}{2}\sqrt{n}.$$

We're looking at this random variable defined as the Hamming distance to A, and we've now shown that this random variable can't have high mean, since it's 0 a reasonable fraction of the time. But now applying the upper tail inequality, we have

$$\mathbb{P}(x \notin A_{C\sqrt{n}}) = \mathbb{P}(Z > C\sqrt{n}) \leq \mathbb{P}\left(Z > \mathbb{E}Z + \frac{C\sqrt{n}}{2}\right).$$

Applying the bounded differences inequality again, we get

$$\mathbb{P}(x \notin A_{C\sqrt{n}}) \leq e^{-2(C\sqrt{n}/2)^2/n} = \varepsilon.$$ □

We now have two proofs of this fact — that if we start with an ε fraction of the space and expand by something on the order of $\sqrt{n}$, then we get at least a $1-\varepsilon$ fraction of the space.

> **Question 21.6.** If we have a result of this form, can we deduce a corresponding concentration result?

Hopefully we see some hints of the connection from these two proofs — they're proving the same thing, and they look somewhat similar.

§21.3 Isoperimetry vs. Concentration

Recall the following definition:

Definition 21.7. Given two metric spaces (X, d_X) and (Y, d_Y), we say that $f\colon X \to Y$ is C-Lipschitz if for all $x, x' \in X$, we have

$$d_Y(f(x), f(x')) \le C d_X(x, x').$$

So in other words, f never stretches two points by a factor of more than C.

We saw this condition before in the bounded differences inequality, whose hypothesis required that f is 1-Lipschitz with respect to the Hamming distance. If that's the case, then we saw that

$$\mathbb{P}(|f - \mathbb{E}f| \ge n\lambda) \le 2e^{-2n\lambda^2}.$$

Intuitively, this says that f is *almost constant almost everywhere.* As long as we start with a Lipschitz function on the Boolean cube, for almost all of the cube, f doesn't vary very much.

There is a general, very simple connection between isoperimetry and concentration:

Theorem 21.8

Let $t, \varepsilon > 0$, and suppose $(\Omega, \mathbb{P})$ is a probability space with a metric. Then the following are equivalent:

(a) (Expansion/approximate isoperimetry). If $A \subseteq \Omega$ and $\mathbb{P}(A) \ge 1/2$, then $\mathbb{P}(A_t) \ge 1 - \varepsilon$.

(b) (Concentration of Lipschitz functions). If $f\colon \Omega \to \mathbb{R}$ is 1-Lipschitz, and some real number m satisfies $\mathbb{P}(f \le m) \ge 1/2$, then $\mathbb{P}(f > m + t) \le \varepsilon$.

These are two statements; one has to do with concentration and the other with isoperimetry, and they are equivalent. The proof is very simple, but the more important thing is the interpretation and application of these ideas.

In the second part, we can think of m as $\mathbb{E}f$; but this is not always correct. It is better to think of m as a *median*:

Definition 21.9. A **median** m of a random variable X is a number with the property that $\mathbb{P}(X \ge m) \ge 1/2$ and $\mathbb{P}(X \le m) \ge 1/2$.

This is a good definition, in that every real-valued random variable has a median; but it is not necessarily unique. (For the Bernoulli random variable with probability $1/2$, any real number in $[0, 1]$ is a median.)

If we have high concentration — if there exists m such that $\mathbb{P}(|X - m| \ge t) \le 2e^{-t^2/2}$ (for example) — then one can check that the median and mean are within $O(1)$ of n. So the mean and median can be different, but they're not that different if we have good concentration.

Now let's prove why or two versions are equivalent to each other.

Proof. To prove (b) from (a), suppose we have a function f which is 1-Lipschitz. We cna then construct a set

$$A = \{x \in \Omega \mid f(x) \le m\}.$$

Then we know that $\mathbb{P}(A) \ge 1/2$, by the hypothesis. Also, since f is 1-Lipschitz, we have that

$$f(x) \le m + t \text{ for all } x \in A_t.$$

(If we start in A we're at most m, so if we move at most a distance t away, we cannot exceed distance $m+t$.) By (a), we then have

$$\mathbb{P}(f > m + t) \leq \mathbb{P}(\overline{A_t}) \leq \varepsilon.$$

On the other hand, to prove that (b) implies (a), let $f(x) = \operatorname{dist}(x, A)$. Then we have

$$\mathbb{P}(f \leq 0) = \mathbb{P}(f = 0) = \mathbb{P}(A) \geq 1/2,$$

and f is 1-Lipschitz (by the definition of f and teh triangle inequality). So

$$\mathbb{P}(\overline{A_t}) = \mathbb{P}(f > 0 + t) \leq \varepsilon$$

(we're taking $m = 0$ here). This proves (a). □

So this gives us a simple connection between the two formal statements. In oen direction we take A to be a level set, and in the other direction we take f to be the distance to A; this shows us a connection between approximate isoperimetry on one hand, and concentration on the other hand.

The next thing we'll do is look at what this connection means in some other spaces. We've shown the bounded differences inequality using martingale techniques, but we can also ask what happens in other geometries, in particular the sphere. And starting with some known isoperimetric inequalities on the sphere, we'll be able to deduce some concentration inequalities.

§21.4 The Sphere

Previously, we stated the classic isoperimetric inequality in Euclidean space — among all bodies of the same volume, the ball has the smallest expansion. We can ask what happens in other spaces — what if we start with a sphere (meaning the *surface* of a sphere) $S^{n-1} \subseteq \mathbb{R}^n$. Then if we take a subset of the sphere with given surface area, which set has the smallest boundary?

We'd expect the answer to be a cap. This is indeed true:

Theorem 21.10 (Levy's Isoperimetric Inequality on S^{n-1})
On the unit sphere $S^{n-1} \subseteq \mathbb{R}^n$, if we have measurable sets $A, B \subseteq S^{n-1}$ such that $\operatorname{Vol}_{n-1} A = \operatorname{Vol}_{n-1} B$ and B is a spherical cap, then for all $t \geq 0$, $\operatorname{Vol}_{n-1} A_t \geq \operatorname{Vol}_{n-1} B_t$.

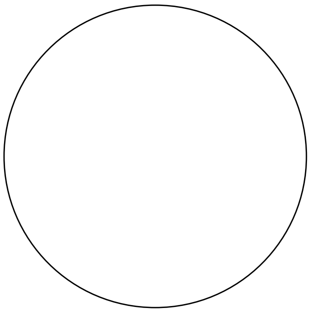
Note that a spherical cap is also a ball with respect to the metric. So this is completely analogous to what we saw earlier, although harder to prove because of the geometry of the sphere.

The theme of today is that if we have an isoperimetric inequality, we can use it to deduce concentration inequalities. So let's see what this gives us.

We'll be working on the unit sphere, and we'll use the natural probability measure — the uniform measure on S^{n-1} — and we'll use the *Euclidean* distance on the sphere. (This is not such a natural choice, because we might prefer the intrinsic geodesic distance; but it doesn't really matter.)

When we had a discussion about starting with half fo the cube, there's a calculation we did using the Chernoff bound — we started with the bottom half and expanded. We need to do a similar calculation for the sphere.

Suppose we have a sphere, and let C be a *hemisphere* (i.e., half of the sphere). Now let's suppose that we expand C by distance t, and look at the complement of C_t.



To do this, we will do some elementary geometry. We can consider the subtended angle θ that cuts off our length-t line segments; then we have

$$\theta = 2\sin^{-1}\frac{t}{2}.$$

Then we have a vertical segment of length

$$\cos\theta = 1 - 2\sin^2\frac{\theta}{2} = 1 - \frac{t^2}{2}$$

from the right endpoint of the t-segment to the halfway vertical line.

We wnat to know some bound on the surface measure of this cap, given this 2-dimensional cross-sectional picture. Now let's look at a ball centered at the point at the bottom of that vertical line, with radius going up to our point. Then the green ball contains the entire blue spherical cap — it contains not just the cap, but the entire sector.

So we have a green ball of radius $1 - t^2/2$, and it contains the sector spanned by $\overline{C_t}$. Now comparing Euclidean volumes, we find that

$$\frac{\operatorname{Vol}_{n-1}\overline{C_t}}{\operatorname{Vol}_{n-1}S^{n-1}} = \frac{\text{sector}}{\text{unit ball}} \le \frac{\text{green ball}}{\text{unit ball}} \le \left(1 - \frac{t^2}{2}\right)^n \le e^{-nt^2/2},$$

by comparing the solid ball volumes (and tkaing the ratio of the radii).

Student Question. *What happens if t is very close to $\sqrt{2}$?*

Answer. Something in the proof may have gone slightly wrong, but the bound is still true.

This is a very similar bound to the one we saw for the Hamming cube (this is not a coincidence).

If we start with $A \subseteq S^{n-1}$ such that $\operatorname{Vol}_{n-1} A \ge \frac{1}{2}\cdot\operatorname{Vol}_{n-1}S^{n-1}$, then we have

$$\operatorname{Vol} A_t \ge \left(1 - e^{-nt^2/2}\right)\operatorname{Vol}_{n-1}S^{n-1}.$$

So starting with $1/2$ and expanding, we get almost everything. Then by the result from earlier, we get the following concentration inequality:

Theorem 21.11

If f is a 1-Lipschitz function on S^{n-1}, then there exists m such that for the uniform measure on S^{n-1}, $\mathbb{P}(|f - m| > t) \le 2e^{-nt^2/2}$.

In other words, a Lipschitz function on a high-dimensional sphere is almost constant almost everywhere. This may be very counterintuitive; it is really about high dmensions. A sphere in three dimensions certainly doesn't have this property. But this statement says that in high dimensions, if we're willing to ignore some extremes of the sphere, most of the sphere should have similar values of f.

§21.5 Gauss Space

Now we'll look at a related space, known as Gauss space.

Definition 21.12. The space **Gauss space** is $\mathbb{R}^n$ with the Gaussian measure γ — the probability measure for $(Z_1, \dots, Z_n)$ where each $Z_i \sim N(0,1)$ is independent.

The Gaussian measure in 1 dimensions is the curve; this is the product of n of these. We can write down explicit formulas for the probability density function — that's

$$\frac{1}{(2\pi)^n} e^{-|x|^2/2}.$$

Question 21.13. What does an isoperimetric inequality on this space look like?

Suppose someone gives you a number; then you want to find a subset of the space which has exactly that measure, and for which the boundary is as small as possible according to the same measure.

Previously, in Euclidean space and Hamming space and on the sphere, we took balls. Here we still have the usual metric, so it's natural to guess that here we should still take a ball. But htat's not the right answer — the answer is the half-space. And today we will see why that's the natural answer here.

Theorem 21.14 (Gaussian Isoperimetric Inequality)

Suppose $A, H \subseteq \mathbb{R}^n$ are measurable subsets, and H is a half-space — the set $\{x \mid a \cdot x \le b\}$ for some constants a and b (i.e., the points on one side of a hyperplane), and $\gamma(A) = \gamma(H)$. Then we have

$$\gamma(A_t) \ge \gamma(H_t)$$

for all $t \ge 0$.

Corollary 21.15

If $f\colon \mathbb{R}^n \to \mathbb{R}$ is a 1-Lipschitz function (in Euclidean space) and Z is a n-dimensional vector with i.i.d. normal coordinates, then there exists m such that

$$\mathbb{P}(|f(Z) - m| \ge t) \le 2e^{-t^2/2}.$$

In order to deduce this result from the Gaussian isoperimetric inequality, we need to fiugre out the measure of a set obtained by expanding a half-space of volume $1/2$. If $H = \{x_1 \le 0\}$, then H has volume $1/2$, and

$$h_t = \{x_1 \le t\}.$$

The coordinates are all independent, and we can check that we then have $\gamma(H_t) \ge 1 - e^{-t^2/2}$. In particular, there is no dependence on the dimension.

We will not prove the Gaussian isoperimetric inequality, but we'll see some intuition now for why it is reasonable.

The Gaussian measure is important because it's an approximation for the binomial distribution. So instead of working with the Gaussian measure on $\mathbb{R}^n$, let's look at something that we can think of as generating it — and that's replacing each coordinate by m independent Boolean bits, and taking n copies (one for each coordinate) — so we instead think of $\{0,1\}^{mn}$ for large m. For a random point in $\{0,1\}^{mn}$, we then have m blocks; if we sum them up and normalize, then we get an approximation for the normal vector.

If we have a function on the Boolean cube, which is really only a function of the means of these blocks, then it shouldn't look so different from a function on the Gauss space — because by the central limit theorem, one is an approximation of the other.

But now let's look at this boolean cube.

> **Question 21.16.** What's an isoperimetric minimizer for the Boolean cube?

We saw that the isoperimetric minimizer is a Hamming ball. And a Hamming ball is the set of vectors whose sum of entries is at most some value b (by definition) — so these are isoperimetric minimizers for the boolean cube.

Running through this connection in the CLT connection, in the Gauss space this becomes a set of vectors whose sum is at most some value. That is a half-space. We can then rotate the half-space to get a usual half-space that only depends on one coordinate.

> **Student Question.** *If you start with this and run the connection backwards (on a rotated half-plane), what do you get?*
>
> **Answer.** You may end up with somewhat different scaling factors when you go backwards.

The Gauss space is cool because it has a couple of nice properties. In particular:

- Rotational invariance.
- Independence of coordinates.

In fact, requiring these two properties forces you to be the Gauss space.

The squared length of a random Gaussian vecotr is given by

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2.$$

We can check (by running the proof of the Chernoff bound on this sum) that this has mean n and sub-exponential concentration with window $O(\sqrt{n})$.

So a typical Gaussian vector has length $\sqrt{n + O(\sqrt{n})}$.

> **Exercise 21.17.** $\sqrt{n + O(\sqrt{n})} = \sqrt{n} + O(1)$.

WHat this says is that a typical Gaussian vector has length concentrated around $\sqrt{n}$. So if we take a typical Gaussian vector, it's basically a point very close to some sphere. In a way, this means Gauss space is very close to the sphere of radius $\sqrt{n}$ (and in fact, almost all points are within constant distnace from this sphere). For lots of problems where we want to understand distributios on a sphere, the analysis can be quite tricky (because the sphere is quite constrained). Meanwhile theh Gaussian vector is often easier to deal with, because coordinates are independent. So a common strategy is instsead of analyzing a random unit vector, we analyze a random Gaussina vector; this is usually much easier and gives you a good enough approximation result.

We can also relate the isoperimetric inequality on the sphere to the one in Gauss space. In the Gauss space, we have a half-space. When we restrict to the half-space on the sphere, then we get a spherical cap — which is also consistent with everything we've said so far. So everything we've said today is at least intuitively consistent with each other.

Student Question. *Can you use this to prove the Gaussian isoperimetric inequality?*

Answer. Prof. Zhao is under the impression that one method of proving it is to go through this connection (starting with a discrete isoperimetric inequality and then taking the central limit theorem). But there are also other and fancier techniques; in the world of high-dimensional probability and asymptotic geometric analysis, there are some continuous methods (such as semigroup methods and semigroup martingales and using stochastic calculus).

§22 November 21, 2022

§22.1 Concentration of Measure on the Sphere

Suppose $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is chosen uniformly on the unit sphere S^{n-1}.

Theorem 22.1

For all $\varepsilon > 0$, we have

$$\mathbb{P}(x_1 \geq \varepsilon) \leq e^{-n\varepsilon^2/2}.$$

This is a bit differemt from the theorem we saw last time, but the spirit is the same. We can draw a cross-section, and consider a spherical cap obtained by pushing the equatorial plane over by ε; then we want to show that the region past it has exponentially small relative surface area.

Proof. We divide into two cases:

Case 1 ($\varepsilon \in [0, 1/\sqrt{2}]$). In this case, we use the same method as last time — notice that the vertical distance is $1/\sqrt{1-\varepsilon^2}$, and if ε is small enough then we can encapsulate our blue sector inside a yellow ball. So the volume of the blue sector over the volume of the entire sphere is at most the volume of the yellow ball over the white ball, which is

$$(\sqrt{1-\varepsilon^2})^n \leq e^{-\varepsilon^2 n/2}.$$

However, we have some issues when ε is large (although this is most interesting when ε is small).

Case 2 ($\varepsilon \in [1/\sqrt{2}, 1]$). Now we can draw a similar picture, where the blue cap is now small, and we still have a blue cone. Now we are going to embed this cone inside a ball that exactly contains it. Let r be the radius of this yellow ball, so that we have a picture with a right triangle of base ε and hypotenuse 1, and we have two radii of length r. Then using similar triangles, we get an equation for r. Then the answer we get is

$$r^n \leq \frac{1}{(2\varepsilon)^n} \leq e^{-\varepsilon^2 n/2}$$

(in the final step we are claiming that $e^{x^2} \leq 2x$ for all $x \in [1/\sqrt{2}, 1]$ — we see that the left-hand side is convex, so it suffices to check the endpoints, which we can do). □

This is sharp when ε is close to 0 — the point is that in high dimensions, the volume of the spherical sector is basically the radius to the power of n, and this radius is what comes up in the expression.

The key message from last lecture is that once you have an isoperimetric inequality (which we do have for the sphere) and an estimate on how half the space grows, this leads to various concentration inequalities. So this gives the following:

Corollary 22.2

If $A \subseteq S^{n-1}$ and $\operatorname{Vol}(A) \geq \frac{1}{2}\operatorname{Vol}(S^{n-1})$, then

$$\frac{\operatorname{Vol} A_t}{\operatorname{Vol} S} \geq 1 - e^{-nt^2/4}$$

for the Euclidean distance.

This is because we're comparing the horizontal distance to the distance between the actual points; this gives a factor of at most 2, which we put in the exponent. (We generally don't worry about constants in the exponent.)

Once you have this isoperimetric inequality, we also have a concentration of measure inequality:

Corollary 22.3

If $f\colon S^{n-1} \to \mathbb{R}$ is 1-Lipschitz, then there exists m such that

$$\mathbb{P}(|f - m| > t) \leq 2e^{-nt^2/4}.$$

So this gives us a concentration of measure result on the sphere — on almost the entire sphere, most of f takes roughly the same value (f is nearly constant on nearly all the sphere).

What we'll do next is use this concentration inequality to prove a result known as the Johnson–Lindenstrauss lemma, which is a pretty useful result in many areas, in particular computer science. You may have a high-dimensional data set and want to preserve some information; but recording it in high dimensions can be expensive, and you want to compress it to lower dimensions without too much loss. So we want to embed a high-dimensional set of points into a lower dimension with small distortion, and the lemma tells us how to do this.

Theorem 22.4 (Johnson–Lindenstrauss Lemma)

There exists a constant c such that the following is true: let $\varepsilon > 0$, and let $X \subseteq \mathbb{R}^m$ have exactly N points. Then as long as $d \geq C\varepsilon^{-2}\log N$, there exists a way to embed X in d dimensions such that for every pair of points in X, if we look at their original distance $|x - y|$ in m dimensions, and consider their new distance $|f(x) - f(y)|$ in the embedding, they are basically the same up to a small factor —

$$(1 - \varepsilon)\,|x - y| \leq |f(x) - f(y)| \leq (1 + \varepsilon)\,|x - y|\,.$$

Note that m is irrelevant in the conclusion. So this tells us as long as we take N points, with roughly $\log N$ dimensions we can embed our points such that the pairwise distances are roughly preserved.

The proof follows from things we've just discussed, and this turns out to be an important tool in many areas. THey first came up with the result in the context of functional analyis, and it is important in computer science for dimension reduction. Up to a constant c, this bound $\varepsilon^{-2}\log N$ is essentially optimal.

Proof. The idea is to take a projection onto a random d-dimensional subspace. If we do this, the lengths of the vectors will get smaller; but we can determine how much they get smaller by, and to correct for thsi we multiply by $\sqrt{m/d}$ (we will see where this comes up in the end).

THe claim that this works hinges on the following lemma, telling us how distance is preserved under a random projection:

Lemma 22.5

Let $m \geq d$, and let $P\colon \mathbb{R}^m \to \mathbb{R}^d$ be an orthogonal projection onto the first d coordinates. Let z be a uniform point on the unit sphere in $\mathbb{R}^m$, and let y be the projection of z onto the first d coordinates (so $y = Pz$), and let $Y = |y|$. Then for all $t \geq 0$,

$$\mathbb{P}(\left|Y - \sqrt{\frac{d}{m}}\right|) \leq e^{-cmt^2},$$

where c is some constant.

Once we have this lemma, we can then apply the lemma to all the vectors of the form $z = (x-x')/|x-x'|$ for all pairs of distinct points in x. What we wrote is the distribution of a random unit vector onto a fixed plane; this is the same as if we start with a fixed vector and project onto a random d-dimensional subspace. So Y in the lemma is also the distribution of the length of the projection of z onto a random d-dimensional subspace, where by 'random' we mean uniformly randomly chosen from the space of d-dimensional subspaces.

We can then set t to be $\varepsilon\sqrt{d/m}$, and find that

$$\mathbb{P}\left(\left|\sqrt{\frac{m}{d}} \cdot Y - 1\right| \geq \varepsilon\right) \leq 2e^{-c\varepsilon d}.$$

So as long as d is at least $c\varepsilon^{-2}\log N$, we see that this quantity is less than $2N^{-cC}$. So as long as C is large enough, this is less than $2/N^2$. Then we can union bound over all $\binom{n}{2}$ pairs of points in X to get a union bound probability less than 1.

Now the key remaining ingredient is to prove the lemma, and for this we will use the concentration result on the sphere.

Here Y is really a function of a point on the unit sphere — it is the length of the projection of this point on the sphere to d dimensions. In particular, the map P sending $z \mapsto Y$ is 1-Lipschitz (projection is 1-Lipschitz, and taking lengths is 1-Lipschitz).

So by Levy concentration, we see that

$$\mathbb{P}(|Y - \mathbb{M}Y| \geq t) \leq 2e^{-mt^2/4}$$

(where $\mathbb{M}Y$ is the median of Y).

We now need to estimate the median of $\mathbb{Y}$. This is somewhat annoying, but $\mathbb{E}[Y^2]$ is easy to compute — we have

$$\mathbb{E}[Y^2] = \mathbb{E}[z_1^2 + \cdots + z_d^2].$$

We know that $z_1^2 + \cdots + z_m^2 = 1$, and all $\mathbb{E}z_i^2$ are the same by symmetry, so each must be $1/m$, and therefore $\mathbb{E}[Y^2] = d/m$.

So at this point, you would expect that Y is very concentrated around $\sqrt{d/m}$, because on one hand it has good concentration properties, and on the other hand its squared expectation is $\sqrt{d/m}$. So it is natural to say that this is true — and it is, but it's surprisingly annoying.

We have control on $\mathbb{E}[Y^2]$, and we also know that there is concentration around the median, so how do we relate these? We can do the following: we'll show that

$$\mathbb{M}Y = \sqrt{\frac{d}{m}} + O\left(\frac{1}{\sqrt{m}}\right),$$

which would imply our result in the lemma — as long as $t \geq 1/\sqrt{2cm}$ then $\mathbb{P}(\left|Y - \sqrt{d/m}\right| \geq t) \leq \mathbb{P}(|Y - \mathbb{M}Y| \geq t/2)$ (we can lose half of t and bring it inside — shifting over loses at most half of t)

and apply the Levy bound to get $2e^{-mt^2/4}$, while if t is small then by choosing c we can make the inequality vacuously hold (because the right-hand side is greater than 1).

To control the median, we first have by the triangle inequality that

$$|\mathbb{M}Y - \mathbb{E}Y| \le \mathbb{E}\,|Y - \mathbb{M}Y| = \int \mathbb{P}(|Y - \mathbb{M}Y| \ge t)\,dt \le \int 2e^{-mt^2/4}\,dt = O(1/\sqrt{m}).$$

But $\mathbb{E}Y$ is not the information that we have; instead we have $\mathbb{E}Y^2$. To relate these, we can do one more similar calculation, where we look at $\operatorname{Var} Y = (\mathbb{E}Y)^2 - \mathbb{E}Y^2$. We know

$$\operatorname{Var} Y = \mathbb{E}[(Y - \mathbb{E}Y)^2].$$

But we know that whatever we put inside, the expectation is the thing that minimizes the sum of squares; so we can only increase the expression by putting in the median instead, meaning

$$\operatorname{Var} X \le \mathbb{E}[(Y - \mathbb{M}Y)^2].$$

Then applying the same tail bound as earlier, we get

$$\int_0^\infty 2e^{-mt/2}\,dt = O(1/m).$$

Putting these together, we have learned that $\mathbb{E}Y = \sqrt{d/m + O(1/m)} = \sqrt{d/m} + O(1/\sqrt{dm})$. We also know that $\mathbb{M}Y = \mathbb{E}Y + O(1/\sqrt{m})$. So therefore $\mathbb{M}Y = \sqrt{d/m} + O(1/\sqrt{m})$, which is what we want. □

This concludes the proof of the Johnson–Lindenstrauss lemma. The last part of this calculation — the reason Prof. Zhao did it was to convince us of the message that when we have concentration around the mean or median, then everythign else becomes quite insensitive to what you take — the mean, the median, the root mean squared, and so on — because we have such good concentration that all these things are fairly close to each other.

This has a quick corollary, somewhat related to a homework problem we did earlier:

Corollary 22.6

There exists a constant c such that for every d, there exists a set of $e^{c\varepsilon^2 d}$ points in $\mathbb{R}^d$ whose pairwise distances are all in $[1-\varepsilon, 1+\varepsilon]$.

So we have a nearly equilateral set of exponentially many points, if we think of ε as being fixed.

Proof. Start with a unit simplex of N points in $\mathbb{R}^{N-1}$, where all points have distances exactly 1. Then apply Johnson–Lindenstrauss to project down to logarithmically many dimensions; the distances are distorted by no more than ε. □

§22.2 Sub-Gaussian Distributions

Before moving on to the next topic, we will introduce some new terminology — partly to summarize what we've done so far, but it's also convenient for what we'll discuss next.

We have been encountering these things we've been calling *sub-Gaussian tails*, so now we'll define this.

Definition 22.7. We say that a random variable X is K-**sub-Gaussian** around its mean if

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-t^2/K^2}.$$

If K is a constant, then this is comparable to a usual Gaussian distribution; otherwise we can think of K as the window of concentration. Usually we don't care about constant factors that much, so we talk about distributions being $O(K)$-subGaussian. If we don't care about constant factors, we also don't usually care about whether we have a mean or median, and the 2 can be any other constant — we're allowed to lose constants left and right.

We've seen a bunch of these so far, which we will summarize now: if we have a space with some distance, then a 1-Lipschitz function is sub-Gaussian with some window by the bounded differences in equality.

Space	Distance	subGaussian
$\{0,1\}^n$	Hamming	$O(\sqrt{n})$
S^{n-1}	Euclidean distance	$O(1/\sqrt{n})$
Gauss $\mathbb{R}^n$	Euclidean distsance	$O(1)$

§22.3 Talagrand's Inequality

Misha Talagrand is one of the great probabilists of modern time, and has done a lot of important work in probability and beyond. We'll talk about one of his results, which has wide-ranging applications.

This result is not easy — the section in the Alon–Spencer book on it is quite dense and difficult to parse. So Prof. Zhao will present it to us more slowly, with more concrete and geometric examples, before jumping to the combinatorial setting — which is important for applications, but harder to grasp.

We will actually not prove Talagrand's inequality. We can find the proof in Alon–Spencer, or in various resources online; but because the inequality is so intricate to state and apply, we'll focus on the statement and application rather than the proof.

Question 22.8. Suppose we fix a d-dimensional subspace $V \subseteq \mathbb{R}^n$, and let $x \sim \text{Unif}\{-1,1\}^n$. How well is $\text{dist}(x, V)$ concentrated?

It's not hard to compute some moments. But what we really want to know is if we can get a sub-Gaussian tail.

We can start by doing some easier calculations. Let $P = (p_{ij}) \in \mathbb{R}^{n\times n}$ be the matrix for projection onto $V^\perp$. Then we have

$$\text{dist}(x, V)^2 = |x \cdot Px|.$$

We can write this down explicitly, as

$$\sum_{i,j} x_i x_j p_{ij}.$$

Then we can check that if we take $x_i \in \{\pm 1\}$, then

$$\mathbb{E}\sum_{i,j} x_i x_j p_{ij} = \sum_i p_{ii}$$

(since the off-diagonal terms have mean 0). This is the trace of the projection matrix, and since it's a projection onto a $(n-d)$-dimensional subspace, the trace is $n-d$. So this tells us that the concentration we're expecting is around $\sqrt{n-d}$. (This is not the mean or median, but it is the root-squared mean, which if we have good concentration should be similar.)

Now let's think about the tightness of the concentration. We can look at this expression and try to analyze higher moments, but let's first look at some special cases.

In particular, this is much easier than V has codimension 1 (i.e., it is a hyperplane). In particular, when V is the coordinate hyperplane $\{x_1 = 0\}$, then we have $\operatorname{dist}(x, V) = 1$, which is not very interesting (there's absolute concentration). If we take $V = (1, 1, \ldots, 1)^{\perp}$,then

$$\operatorname{dist}(x, V) = \frac{|x_1 + \cdots + x_n|}{\sqrt{n}}.$$

We know that this is pretty nicely concentrated; in particular it is $O(1)$-subGaussian (by the central limit theorem, or the Chernoff bound). More generally, if we take V to be the orthogonal complement of some fixed vector α, then

$$\operatorname{dist}(x, V) = |x \cdot \alpha|\,.$$

We see that changing x_i changes the value by $2\,|\alpha_i|$. A version of Azuma's inequality tells us that if we know the individual coordinate changes, we can find that the window is teh sum of the squares of the original changes; so in this case, we alos have $O(|\alpha_1|^2 + \cdots + |\alpha_n|^2) = O(1)$.

In codimension 2, we can do something similar. But this doesn't work in codimension k.

We can consider the function $\{-1, 1\}^n \to \mathbb{R}$ sending $x \mapsto \operatorname{dist}(x, V)$. If we change one coordinate of x, then the result changes by at most 1, so this is 1-Lipschitz. By the bounded differences inequality, this tells us that it's $O(\sqrt{n})$-subGaussian. This is not very good — the diameter of the whole cube is actually only $\sqrt{n}$, so the range of values is $\sqrt{n}$ anyways. So this is almost a useless result.

Why was the bounded differences inequality not efficient here? The reason is that the distance that's relevant for the bounded differences inequality is the Hamming distance. But if you didn't knwo the bounded differences inequality, you would naturally care about the Euclidean distance here. The function si 1-Lipschitz not just for the Hamming distance, but for the Euclidean distance — and we have not used that fact in this proof.

SO thisi s a bit strange — on one hand we are looking at a random point on the unit cube, but we now need to think about the Euclidean metric on the uniti cube.

Talagrand developed a powerful inequality which allows us to answer this question. The general form is hard to state, so we will first state a consequence that implies this result:

Theorem 22.9

$\mathbb{P}(\left|\operatorname{dist}(x, V) - \sqrt{n-d}\right| \geq t) \leq Ce^{-ct^2}$.

So in other words, this is $O(1)$-subGaussian around the mean.

> **Remark 22.10.** Prof. Zhao likes the convention that C is a big constant and c is a little constant. This is not universal, but it is useful to keep in mind.

What Talagrand's inequality is about is the concentration of *convex Lipschitz functions* of independent random variables. There are a few key words here, and we will see how they appear.

This is not the most general statement, but it's a bit more palatable:

Theorem 22.11 (Talagrand)

If $A \subseteq \mathbb{R}^n$ is convex, and $x \in \operatorname{Unif}\{0, 1\}^n$, then

$$\mathbb{P}(x \in A) \cdot \mathbb{P}(\operatorname{dist}(x, A) \geq t) \leq e^{-t^2/4}.$$

There are some subtleties here.

Remark 22.12. A is really a convex set. It's *not* just a subset of the cube vertices — it's really a blob in Euclidean space. And this is important — we will see an example where if we don't assume this, the theorem fails.

Remark 22.13. The bounded differences inequality is much worse in this setting — it gives a tail bound of t^2/n in the exponent, which is $\sqrt{n}$ concentration — much worse than what is claimed here.

Talagrand's inequality fails for non-convex sets.

Example 22.14

Take A to be the set $\{x \in \{0,1\}^n \mid \mathrm{wt}(x) \leq n/2 - \sqrt{n}\}$, which is a subset of points on the Boolean cube (not a convex set). Then for every $y \in \{0,1\}^n$ with $\mathrm{wt}(y) \geq n/2$, we can check that $\mathrm{dist}(y, A)$ is the distance from y to the closest point in A in Euclidean distance. To start with y and get the closest point in A, we would change some 1's to 0's — in partyicular we'd change $\sqrt{n}$ 1's to 0's — so the distance is $n^{1/4}$. But we also see that A has $\mathbb{P}(x \in A) \geq c$ due to the central limit theorem, and $\mathbb{P}(\mathrm{wt}(x) \geq n/2) \geq 1/2$. So the product si fairly large; meanwhile $t = n^{1/4}$. So the inequality completely fails.

What happened here? We cna think about this geometrically. We have a set of points forming A, which is not a convex set. Maybe we care about this set for combinatorial reasons, but it's not a convex set. What we should actually do is look at their convex hull. Once we do that, the distance from another point to this convex hull is the Euclidean distance to the convex hull, rather than the Euclidean distance to one of the ppoints — and in high dimensions that is a big difference.

So geometrically, we should start with a convex set — or a convex hull of a discrete set — and then Talagrand's inequality holds.

Let's draw some concentration of measure consequences.

Corollary 22.15

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function which is *convex* and 1-Lipschitz with respect to the Euclidean metric. Let $x \sim \mathrm{Unif}\{0,1\}^n$. Then for all $r \in \mathbb{R}$ and $t \geq 0$, $\mathbb{P}(f(x) \leq r)\mathbb{P}(f(x) \geq r + t) \leq e^{-t^2/4}$.

This is actually an equivalent statement to the theorem here — the equivalence is basically teh same as the argument we saw earlier with the two notions of concentration of measure (one with expansion adn the other with concentration of functions).

Proof. First let's suppose we start with teh theorem and want to get to the corollary. Let A be the set of points where $f(x) \geq r$; then $A \subseteq \mathbb{R}^n$ and because f is convex, A is convex as well. (If we take a convex function and look at all points below a certain value, that set will be convex.)

Then we see that for every x with $\mathrm{dist}(x, A) \leq t$, then we have $f(x) \leq r + t$ (if we start with some point in A, wander by less than t, then since f is 1-Lipschitz the target value cannot change by more than the distance we've moved). So then our first set is A, and the second set is contained in the complement of the t-neighborhood of A; so then the corollary follows from the theorem.

We can do a very similar argument in reverse — if we start with the corollary, then we can take f to be the distance from x to A, and arrive at a very similar conclusion (setting $r = 0$). □

Student Question. *How much does the fact that we're taking a point on the hypercube matter?*

Answer. Not much. As long as each coordinate of x is in $[0, 1]$ independently, the same result holds.

We'll now derive a form that's a bit more useful than knowing the product of two probabilities. Namely, by setting r in this inequality to be a median of f, we can deduce the following:

Corollary 22.16

We have $\mathbb{P}(|f(x) - \mathbb{M}f(x)| \geq t) \leq 4e^{-t^2/4}$.

Remark 22.17. The constants 4 and 4 are not important; the dimensionless of this expression *is* important.

Proof. Set r to be $\mathbb{M}f$. Then we see that the first factor is at least $1/2$, so $\mathbb{P}(f \geq \mathbb{M}f + t) \leq 2e^{-t^2/4}$. Likewise, by setting $r = \mathbb{M}f - t$, we get a corresponding lower tail of $\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2e^{-t^2/4}$. Putting them together gives the desired bound. □

Once we assume this corollary, then our original theorem about $\operatorname{dist}(n, V) - \sqrt{n-d}$ follows — $\sqrt{n-d}$ is not the median but is intsead the root-mean-squared, but once we have concentration about one of them, we have concentration about any other.

We'll finish with another quick application. Let A be a random matrix — if we like, we can also consider what happens for symmetric matrices, but it doesn't really matter — with independent entries in $[-1, 1]$ (which don't have to be uniform or even identical).

Assuming only the form we've already stated, we'll assume i.i.d uniform entries in $\{-1, 1\}$; but it turns out you can be more flexible.

Question 22.18. How well is the operator norm of A concentrated?

By definition,

$$\|A\|_{\text{op}} = \sup_{|x|=1} |Ax|.$$

So this is some function fo amatrix, and we want to see how wsell it's concentrated.

The function $f\colon A \mapsto \|A\|_{\text{op}}$ is a function $\mathbb{R}^{n^2} \to \mathbb{R}$ (or $\binom{n}{2}$ if we take symmetric matrices) with some nice properties:

- It is 1-Lipschitz with respect to the Hilbert–Schmidt norm — the Euclidean distance where we view the matrix as a vector of its entries — because the Hilbert–Schmidt norm upper-bounds the operator norm.
- It's also convex — if we look at it on the set of all matrices, then it's convex because it's a norm (it satisfies teh triangle inequality).

Therefore, we can apply Talagrand's inequality to deduce that $\|A\|_{\text{op}}$ is $O(1)$-subGaussian around its mean.

This is pretty interesting, because theh operator norm of A is typically on teh order of $\sqrt{n}$. In fact, the truth is even better (although we can't show it using this method) — the actual window of concentration is much smaller. The typical fluctuation is actually $n^{-1/6}$, which is a deep result (there's a result about the distribution of the top eigenvalue of a random matrix).

This gives a geometric interpretation fo Talagrand. Next time we will see the more general form,w hich is a bit harder to grasp but allows us powerful combinatorial consequences.

§23 November 28, 2022

§23.1 Talagrand's Inequality

> **Remark 23.1.** Talagrand introduced this inequality in the 1990s as a new way to look at independence, and to prove concentration bounds for certain combinatorial optimization problems.

Last time, we looked at a few cases of Talagrand's inequality that were geometric, making them easier to think about:

- If $A \subseteq \mathbb{R}^n$ is convex, and $x \sim \operatorname{Unif}\{0,1\}^n$, then
$$\mathbb{P}(x \in A) \cdot \mathbb{P}(\operatorname{dist}(x, A) \geq t) \leq e^{-t^2/4}.$$
  Note that A being convex is important.
- As a corollary, by reinterpreting this isoperimetric statement in terms of Lipschitz functions, we deduce that if $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and 1-Lipschitz with respect to the Euclidean metric, then if $x \sim \operatorname{Unif}\{0,1\}^n$, we have
$$\mathbb{P}(|f - \mathbb{M}f| \geq t) \leq 4e^{-t^2/4}.$$

We aren't going to prove these inequalities; the proofs are in the textbook.

This isn't really the point of Talagrand's inequality, though it's a simpler special case it's easier to get a feel for. The point of Talagrand's inequality is a generalization of these that can be used for combinatorial optimization. For that, we need to introduce a notion of *convex distance*.

§23.2 Convex Distance

Convex distance is a bit of a subtle notion, so we'll define it in a few steps.

We'll consider a product probabilitiy space $\Omega = \Omega_1 \times \cdots \times \Omega_n$ (of independent random variables).

> **Definition 23.2.** Given a vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n_{\geq 0}$ and $x, y \in \Omega$, the **weighted Hamming distance** $d_\alpha(x, y)$ is defined as
> $$d_\alpha(x, y) = \sum_{i : x_i \neq y_i} \alpha_i.$$

When $\alpha = (1, 1, \ldots, 1)$, this is the Hamming distsance; in general it's like the Hamming distance, but with weights attached.

> **Definition 23.3.** For a point x and a set $A \subseteq \Omega$, we define $d_\alpha = \inf_{y \in A} d_\alpha(x, y)$.

Talagrand then introduced the notion of a *convex distance*:

> **Definition 23.4.** For a set $A \subseteq \Omega$, the **convex distance** $d_T(x, A)$ is defined as
> $$d_T(x, A) = \sup_{\alpha \in \mathbb{R}^n_{\geq 0}, |\alpha| = 1} d_\alpha(x, A).$$

Here $|\alpha|$ is the Euclidean norm (which we're requiring to be 1), so $\sum \alpha_i^2 = 1$.

Here's another way to view this definition.

Example 23.5

Suppose $\Omega = \{0,1\}^n$ is the Hamming cube (this is a combinatorial set, and doesn't come with Euclidean structure), and suppose $A \subseteq \Omega$ and $x \in \Omega$. Then $d_T(x, A)$ is the Euclidean distance between x and the *convex hull* of A.

Here we have a point x and some set A, which right now is a discrete set of points. We're asking to point a unit vector α in the optimal direction — and once we point α, we'll measure the distance to one of these points in that direction. The way to optimize α is to point it to the nearest point in the convex hull of A, in which case we get this Euclidean distance.

In general, we can actually reduce down to this special case. Given $x, y \in \Omega$, we can define $\phi_x(y)$ as the binary vector of disagreements — $\phi_x(y) = (\mathbf{1}_{x_1 \neq y_1}, \mathbf{1}_{x_2 \neq y_2}, \ldots) \in \{0,1\}^n$. For a set A, we then define $\phi_x(A) := \{\phi_x(y) \mid y \in A\}$, which is a subset of $\{0,1\}^n$.

For every $\alpha \in \mathbb{R}^n_{\geq 0}$, in $d_\alpha(x, A)$, whether we're summing a certain term only depends on agreements or disagreements — it doesn't depend on any other information. So it's really the weighted Hamming distance

$$d_\alpha(x, A) = d_\alpha(0, \phi_x(A)).$$

(The distance on the left is defined on Ω, and the distance on the right is defined on the Hamming cube.)

Now when we optimize over α, taking $\sup_\alpha$, we find that the left-hand side becomes the Talagrand convex distance, and the right-hand side becomes the Euclidean distance between the origin and the convex hull of $\phi_x(A)$. This is why it's called the convex distance. (We can view this as another definition — to measure the distance between x and A, we think of x as the origin, and for every point in A we measure its disagreement vector. Then this gives a subset of the Euclidean cube, and we measure the Euclidean distance from the origin to the convex hull of these disagreement vectors.)

§23.3 Talagrand's Inequality

Here's the general form of Talagrand's inequality:

Theorem 23.6 (Talagrand's Inequality)

If $A \subseteq \Omega = \Omega_1 \times \cdots \times \Omega_n$ (a probabilitiy space equipped with a product probability measure) and x is chosen according to this probability measure (i.e. with independent coordinates), then

$$\mathbb{P}(x \in A) \cdot \mathbb{P}(d_T(x, A) \geq t) \leq e^{-t^2/4}.$$

We see that the space we're working in is very general — it captures lots of combinatorial setups, not just things on the Boolean cube. And the conclusion is some sub-Gaussianity, although we have to explain how to interpret and use this inequality.

We'll first discuss a few things.

Question 23.7. How does this recover the form with Euclidean convex distance?

We'll call this (T), and the Euclidean statement (A).

To prove that (T) implies (A), we'll see the following relationship between convex and Euclidean distance:

Lemma 23.8

For a subset $A \subseteq [0,1]^n$ with $x \in A$, we have

$$\operatorname{dist}(x, \text{convex hull of } A) \leq d_T(x, A).$$

When we looked at just the vertices of the cube, this is actually an equality; but we can also have a solid cube, and the same is true except with an inequality. Recall that the definition of the convex distance doesn't care about the specific values of the coordinates (there's no metric on the coordinates), just about whether they're equal or not equal.

Proof. For every $\alpha \in \mathbb{R}^n$ and every $y \in [0,1]^n$, note that

$$|(x-y)\cdot\alpha| \le \sum |\alpha_i| \cdot |x-y| \le \sum_{x_i \neq y_i} |\alpha_i|$$

(since we can bound $|x_i - y_i|$ by 0 or 1). Now taking inf over all $y \in A$, and then taking sup over all unit vectors α, when we compare definitions we see that the left-hand side becomes the Euclidean distance, adn the right-hand side becomes the convex distance. □

Corollary 23.9

Suppose $x \in [0,1]^n$ is a random point with independent random coordinates (they don't have to be uniform or identically distributed). Then if $A \subseteq [0,1]^n$ is a convex set, then

$$\mathbb{P}(x \in A)\cdot\mathbb{P}(\operatorname{dist}(x,A) \ge t) \le e^{-t^2/4}$$

(where this denotes Euclidean distance) for all $t \ge 0$.

So this is a bit more general than our original statement, which only allowed points on the vertices of the Boolean cube — we can have points anywhere inside the cube, as long as the coordinates are independent.

Corollary 23.10

If $f\colon [0,1]^n \to \mathbb{R}$ is convex and 1-Lipschitz, then

$$\mathbb{P}(|f(x) - \mathbb{M}f(x)| \ge t) \le 4e^{-t^2/4}.$$

The second statement can be deduced from the first by the same trick as last time — we can deduce the second from the first by taking A to be various sub-level sets of f.

This is still more related to geometric applications; for the rest of the lecture, we'll focus more on combinatorial applications.

§23.4 A Useful Form of Talagrand's Inequality

Theorem 23.11

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$, and suppose $f\colon \Omega \to \mathbb{R}$. Suppose that for every $x \in \Omega$, there is some $\alpha = (\alpha_1(x), \ldots, \alpha_n(x)) \in \mathbb{R}^n_{\geq 0}$ such that for every $y \in \Omega$,

$$f(y) \geq f(x) - \sum_{i | x_i \neq y_i} \alpha_i(x).$$

Then for every $t \geq 0$,

$$\mathbb{P}(|f(x) - \mathbb{M}f(x)| \geq t) \leq 4e^{-t^2/k^2},$$

where

$$k = 2 \sup_{x \in \Omega} |\alpha(x)|.$$

(Here $|\alpha|$ denotes the Euclidean norm.)

Often f is the result of some optimization problem — next time we'll look at travelling salesman, where we have n random points in the unit square and f is the length of the shortest tour — f may be complicated and not easy to compute, but we'll see a way to get a concentration bound provided some control on the set.

Previously we saw the expression $\sum_i \alpha_i(x)$ as $d_\alpha(x, y)$.

We'll see an example in a second, but first let's think about what this says. We have a function f, which is the objective function to some optimization problem. This condition says that for every x, we can assign to it some weight vector, and this weight vector has the following effect: if we start with $f(x)$ and we change some of the coordinates of x to get y, we want to keep track of how much $f(x)$ can change (i.e. how much an adversary could change f by changing some coordinates). The answer is not very much — α tells us how resilient f is to changing some of its coordinates. In the travelling salesman problem f will be the length of the shortest tour; if I change a few of the points, the shortest tour may change, but not by that much — it's somewhat resilient to small changes in the input. This is often the case with many optimization problems.

If this is the case, then we have a concentration bound — where f is k-subGaussian around its median, where k is the Euclidean norm of the largest α vector. So $\alpha(x)$ records the resilience of f to changing some coordinates of x.

Remark 23.12. By replacing f by $-f$, the hypothesis changes to a similar one with signs reversed — if $f(y) \leq f(x) + \sum_{i|x_i \neq y_i} \alpha_i(x)$, then the same is true.

Remark 23.13. The power of Talagrand's inequality comes from the fact that we can choose a *different* vector for each x. In comparison to the bounded differences inequality, the bounded difference inequality corresponds (up to constants) to taking a *fixed* α-vector for all x. It says that if we start with $f(x)$ and flip one coordinate, and f changes by no more than α_i, then we have sub-Gaussianity with window $|\alpha|$. So this theorem is a significant generalization of bounded differences — in bounded differences it's often an issue that typical changes are small but the worst-case effect of switching a single coordinate is gigantic. And this is one way of sometimes dealing with that situation, where we can assign a different α to each x.

First let's prove this form, which we call (C), from (T).

Proof. Let $r \in \mathbb{R}$, and let $A = \{y \in \Omega \mid f(y) \leq r - t\}$ (where f is the given function in the theorem statement). Now consider some point $x \in \Omega$ with $f(x) \geq r$. Then $f(x)$ and $f(y)$ have very different values, so our condition should tell us that there is some weight-vector that witnesses this distinction between the values of f.

By hypothesis, there exists some $\alpha(x) \in \mathbb{R}^n_{\geq 0}$ such that $d_{\alpha(x)}(x, y) \geq f(x) - f(y)$ — this is the same as the inequality given in the theorem statement of (C). And $f(x) - f(y) \geq t$ for all $y \in A$. Now by taking inf over all $y \in A$, we find that for this α,

$$d_{\alpha(x)}(x, A) \geq t.$$

But the left-hand side is at most $|\alpha| \cdot d_T(x, A)$ — $d_T(x, A)$ was defined by taking sup over all unit vectors α, and here we didn't normalize α so we need to remember its norm.

So

$$d_T(x, A) \geq \frac{t}{|\alpha(x)|} \geq \frac{2t}{k}.$$

Thus by (T),

$$\mathbb{P}(f(x) \leq r - t) \cdot \mathbb{P}(f(x) \geq r) \leq \mathbb{P}(A)\mathbb{P}(d_T(x, A) \geq \frac{2t}{k}).$$

Talagrand's inequality tells us an upper bound on the latter expression of at most e^{-t^2/k^2}.

So now we have some tail bounds, and the next part is the one we saw last time — we can set r to be the median and the median plus t.

First, taking $r = \mathbb{M}f + t$, we obtain

$$\mathbb{P}(f \geq \mathbb{M}f + t) \leq 2e^{-t^2/k^2}$$

(since f is at least the median with probability at least $1/2$). Likewise, taking $r = \mathbb{M}f$ we find

$$\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2e^{-t^2/k^2}.$$

Putting these together gives the desired result. □

Student Question. In (C) is f required to be convex?

No, the only requirement is this inequlaity, which we can view as some combinatorial encapsulation of both Lipschitzness and convexity. (The notion of convexity doesn't make sense in this setting.)

The rest of this lecture is about applications.

§23.5 Largest Eigenvalue of a Random Matrix

Theorem 23.14

Let A be a random symmetric matrix (a_{ij}) with independent entries in $[-1, 1]^{n\times n}$. Let $\lambda_1(A)$ be the largest eigenvalue of A. Then

$$\mathbb{P}(|\lambda_1(A) - \mathbb{M}\lambda_1(A)| \geq t) \leq 4e^{-t^2/c}$$

for some constant.

When we say *independent* entries, we mean independent subject to being symmetric — for example, you can choose all the diagonal entries and all the entries above the diagonal, and define the rest of the matrix based on the symmetry.

Remark 23.15. If we take a random matrix with binary entries, what's the typical largest eignevalue? THis is actually related to the homework problem about flipping lights; the answer is that if we start with $\{-1, 1\}^n$, then the largest eignevalue is on the order of $\sqrt{n}$ (and we even know what the constant is). So having a constant-sized window is not too bad.

We will apply Talagrand's inequality; we wan tto verify the hypothesis of f, where f is the λ_1 function.

This means we'd like, for every matrix A, to come up with a weight vector $\alpha(A)$ — where $\alpha(A)$ defines the resilience of the eigenvalue under changing some coordinates. Here we only look at entries weakly below the diagonal; so we look at $\alpha_{ij}(A)$ for $i \leq j$. The property we watn is that

$$\lambda_i(B) \geq \lambda_i(A) - \sum_{a_{ij} \neq b_{ij}} \alpha_{ij}.$$

To do this, we want to recast the first eigenvalue in terms of some optimization problem. There is a standard way to do this, coming from the min-max characterization of eigenvalues — let $v = v(A)$ be the unit eigenvector of A for $\lambda_i(A)$. Then we know that $v^\intercal Av = \lambda_i(A)$, and that

$$v^\intercal Bv \leq \lambda_1(B).$$

So now we observe that $\lambda_1(A) - \lambda_1(B) \leq v^\intercal(A - B)v$. We can expand out the matrix multiplication (as an upper bound) as

$$\sum_{a_{ij} \neq b_{ij}} |v_i|\,|v_j|\,|a_{ij} - b_{ij}|$$

(since we only need to sum over things which are nonzero). We have $|a_{ij} - b_{ij}| \leq 2$, so this is at most

$$2 \sum_{a_{ij} \neq b_{ij}} 2\,|v_i|\,|v_j|\,.$$

This is basically the form that we're looking for — this informs us how to choose the weight of α (since this is the kind of format that we're looking for). So we should choose $\alpha(A)$ to be $(\alpha_{ij})_{i \leq j}$ as

$$\alpha_{ij} = \begin{cases} 4\,|v_i|\,|v_j| & i < j \\ 2\,|v_i|^2 & i = j \end{cases}$$

(we deal with the on-diagonals and off-diagonasl separately since they appear a different number of times in the sum, but if we ignore constants it doesn't matter). Then we have

$$\sum_{i \leq j} |\alpha_i|^2 \leq 8 \sum_{i \leq j} |v_i|^2\,|v_j|^2 = 8(|v|^2)^2 = 8.$$

This gives the constant we were looking for.

So this tells us that if we have some f reslieitn to changing coordinates in this way, then we have a sub-Gaussianity determined by the worst-case Euclidean norm of the resilience vector. Here we chose the resilience vector coming from the Courant–Fischer characterization of eigenvalues. In particular we see that v depends on A; so the individual entries of α can vary dramatically, but in the L^2 average everything is okay.

Student Question. Why is the second inequality true?

It's a fact that $\lambda_1(A) = \sup_{|v|=1} v^\intercal Av$. This is attained because of compactness; so taking v to be the one attaining it for A we get the first thing.

Remark 23.16. This is much better than we could have gotten using other methods, but it's not the truth. It's known that for a uniformly ±1 random symmetric matrix, the scale of fluctuation of the top eigenvector is not constant, but actually $n^{-1/6}$ (and that is the true scale). This is a deep fact, and when you normalize appropriately to get

$$\frac{X - \mathbb{E}X}{cn^{-1/6}},$$

this converges to a universal distribution that we haven't seen yet, the Tracy–Widon distribution. It comes up here, as well as in many other contexts; especially problems that arise from optimization.

§23.6 Another Application

Question 23.17. Given a random permutation of $[n]$, let X be the length of the longest increasing subsequence. What is its distribution?

Example 23.18

For the permutation 462315, the longest increasing subsequence has length 3.

Exercise 23.19. The answer is $\Theta(\sqrt{n})$ with high probability.

But now the question is concentration — how well is this concentrated? If we use the bounded differences inequality, as a permutation the entries are not independent. But there is a standard way to make them independent — namely, we take n uniform random points and look at the permutation induced by the ordering of those points. Then we do have independent coordinates.

Under this formulation, if we change one coordinate, the length could change by 1. So the bounded differences inequality gives us an $O(\sqrt{n})$-subGaussian result, which is terrible because the actual length is $\sqrt{n}$.

We will see that Talagrand will tell us the concentration is $O(n^{1/4})$ — roughly the square root of the actual length.

This problem has a special structure — to *certify* that $X \geq k$, we only need to see k coordinates. If someone generates a random permutation by generating random real numbers, and says they have an increasing subsequence of length at least 10 and wants to prove this to us, they only have to reveal 10 coordinates. Not all problems have this structure, but htis problem does; we'll see that any problem with this structure has very good concentration results.

Theorem 23.20

Suppose f is a function $\Omega = \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ such that:

- f is 1-Lipschirz with respect to the Hamming distance.
- The event $\{f \geq r\}$ is r-certifiable for every r (we will definen this in a moment, but we can guess what this means — I can certify $f \geq r$ by revealing r coordinates, not necessarily fixed).

Then for every t,

$$\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2e^{-t^2/4\mathbb{M}f},$$

while

$$\mathbb{P}(f \geq \mathbb{M}f + t) \leq e^{-t^2/4(\mathbb{M}f+t)}.$$

This is similar to the asymmetry in Chernoff bounds, where the upper tail is sub-exponential but not sub-Gaussian. So we'll show that it's $n^{1/4}$-concentrated, but not sub-Gaussian.

First we need to define certifiability. To certify that a permutation has an incrasing sequence of lneght k, we only need to see k coordinates. In general:

Definition 23.21. We say $A \subseteq \Omega$ is s-**certifiable** if for all $x \in A$, there exists some subset $I(x) \subseteq [n]$ (which could depend on the element — the coordinates you reveal to show that something's an increasin gpermutation will depend on the permutation) with $|I(x)| \leq s$, such that for every y that agrees with x on the coordinates in I (i.e. $x_i = y_i$ for all $i \in I(x)$), one has $y \in A$.

For increasing subsequences, to say the set of permutations with increasin gsubsequence size at lst k is certifiable means that for any such permutation, I can reveal s coordinates so that anything that agrees with my revealed coordinates also has the same property (here $s = k$).

First here's a corollary of Talagrand for certifiable functions:

Corollary 23.22

If f is 1-Lipschitz with respect to the Hamming distance, and $\{f \geq r\}$ is s-certifiable, then for all $t \geq 0$,

$$\mathbb{P}(f \geq r - t) \cdot \mathbb{P}(f \geq r) \leq e^{-t^2/4s}.$$

We will skip the proof. You can deduce this from Talagrand by using this definition as a way to choose the α going into the definition. This is not dificult; most of it is unravelling the definitions.

Then from the corollary, we can deduce the theorem, which implies the result about the concentration.

In both cases, fundamentally what's happening is — previously in the class we looked at concentration for the number of triangles in the longest graph, which is statistical. But here we're looking at quantities which are results of optimization problems – we start with a random matrix and try to find hte largest eignevalue, which is an optimization problem applied to the random matrix ($v^\intercal Av$). Here longest increasin gsequence means we start with a random object and aply an optimziation process to aht.

Talagran'd sinequality is effective when applied to the objective of an optimization problem — that's because there's som uch flexibility in the statement.

Remark 23.23. For LIS, it turns out that the actual window of concentration — Vershik–Kerov 1977 tells us that $\mathbb{E}X \sim 2\sqrt{n}$. A much deeper result due to Baik–Deift–Johansson 1999 shows that the correct window is $\Theta(n^{1/6})$-concentration, and that once we normalize by the scale factor,

$$\frac{X - 2\sqrt{n}}{n^{1/6}} \to \text{Tracy–Widon}.$$

So in this sense, this si auniversal distribution — it comes up in many unrelated-appearing problems.

§24 November 30, 2022 — Euclidean Travelling Salesman Problem

We'll consider the following problem:

Problem 24.1. Suppose $x_1, \ldots, x_n$ are points in the unit square $[0, 1]^n$. We want to find $L(x_1, \ldots, x_n)$, the length of the shortest tour through the points $x_1, \ldots, x_n$ (where a *tour* must begin and end at the same point).

The travelling salesman problem in general is hard to even approximate. But in the *euclidean* version, there's a PTAS — in polynomial time we can approximate the problem up to a factor of $1 + \varepsilon$.

There are lots of practical uses, but also fun challenges.

Example 24.2 (Mona Lisa TSP Challenge)

Someone generated a dot picture of the Mona Lisa, and the challenge is to find a shortest tour. At some point, a prize of $1000 was offered to beat the current best tour.

We won't be concerned with the *worst* case, but what happens with *random* input.

Definition 24.3. In other words, $L(x_1, \ldots, x_n) = \min_\sigma \left|x_{\sigma(1)} - x_{\sigma(2)}\right| + \cdots + \left|x_{\sigma(n)} - x_{\sigma(1)}\right|$. We define L_n as $L(x_1, \ldots, x_n)$ for n i.i.d. points distributed uniformly in the square.

Exercise 24.4. With high probability, $L_n = \Theta(\sqrt{n})$.

(It's useful to think about why we'd expect this to be the right order of magnitude.)

In fact, much more is known about the leading constant factor, although there is still an open problem about this:

Theorem 24.5 (Beardwood, Halton, Hammersly 1959)

The ratio $L_n/\sqrt{n}$ converges to a constant.

However, this constant is not known.

We will focus on the *concentration* of L_n.

Demonstration 24.6

We have a live demo of the problem. If we put 100 random points in the unit square and ask Mathematica to draw a shortest tour, it'll do it (for small n it'll find the largest tour; for large n it'll approximate). The value we get here is about 7.7.

When we increase n to 1000, we get an interestring pattern, which is related to some physical phenomenon.

For 10000, the length is 72.4. If we run again, it becomes 72.39. Running it again, we get 72.41, and then 72.71, and then 72.88, then 71.88, then 72.68. So it doesn't seem to vary very much — the length of this tour seems to exhibit quite good concentration.

The goal of today is to prove some bounds on this concentration.

So far, we've looked at a few concentration methods, and a few theorems stood out — the bounded differences inequality, Azuma's inequality on martingales, and Talagrand's inequality. Let's see what happens when we try to apply these techniques.

With bounded differences, we can ask — if we change a single point, how much can the length of the tour change?

The worst case is of constant order — if all the points are at some corner, and then you bring one point somewhere else, we see that the cost of that point is of a constant size. So the bounded differences inequality will tell us a concentration window of $\sqrt{n}$ — the bounded differences inequality tells us that if changing x_i changes f by at most c_i, then f is k-subGaussian about its mean, with k on the order of $\sqrt{\sum c_i^2}$. Here all the c_i are constant, so $k \asymp \sqrt{n}$. This gets us $O(\sqrt{n})$-subGaussian, which is a trivial result as the actual size is about $\sqrt{n}$. So we would like to do better.

Nevertheless, the idea of changing one point will still come up, so let's summarize what we know. For every set of points S in the square, and one point x in the box, we know

$$L(S) \leq L(S \cup \{x\}) \leq L(S) + 2\text{dist}(x, S).$$

Of course adding a point can't shorten our tour (we could just cut out that point); on the other hand, for the upper bound if we have a tour of S, then we could take a detour from the shortest point to go to x, and then come back.

Let's now try to apply Azuma's inequality for this problem. SO far in this chapter, we have not yet seen the difference between applications of Azuma and applications of bounded differences (we've used them the same way). WHat we're about to see next is instructive in seeing that you can get more out of martingales than just the bounded differences inequality.

Azuma's says that if we have a martingale (here we'll use the Doob exposure martingale) which satisfies some bounded differences properties — that $|Z_i - Z_{i-1}| \le c_i$ for all i — then we have the same conclusion as before. So we need to look at the worst-case change in each step of the martingale.

The martingale we'll look at is the Doob martingale, where the ith term is $\mathbb{E}[L(x_1, \dots, x_n) \mid x_1, \dots, x_i]$. Towards the very end, once you've revealed all but a few points, the possible change to this number in the worst case is just like before. But in the very beginning, if someone tells you where x_1 is, whether it's in one corner or another or the center, this doesn't change your remaining expectation by all that much. So the *initial* points don't affect the expectation by much.

So we can gain over the bounded differences inequality at the very beginning, and that'll be enough to get us something useful.

Theorem 24.7 (Rhee–Talagrand 1987)

L_n is $O(\sqrt{\log n})$-subGaussian about its mean.

This is a much better result — it's $\sqrt{\log n}$ rather than something polynomial in n.

TO do this, let's first consider the following lemma:

Lemma 24.8

Suppose that S is a set of k i.i.d. uniform random points in $[0,1]^2$. Then for all points $y \in [0,1]$ (here y is non-random),

$$\mathbb{E}\mathrm{dist}(y, S) \lesssim \frac{1}{\sqrt{k}}.$$

So if we fix a point y and lay down k uniform random points, and ask how far is y from the closest such point, it turns out the answer is around $1/\sqrt{k}$.

Proof. It's possible to calculate this almost exactly — we have

$$\mathbb{E}\mathrm{dist}(y, S) = \int_0^{\sqrt{2}} \mathbb{P}(\mathrm{dist}(y, S) \ge t)\, dt.$$

For the expression inside, this is the same as saying we *don't* have any points within a t-radius ball around y. So this is the same as

$$\int_0^{\sqrt{2}} (1 - \mathrm{area}(B(y,t) \cap [0,1]^2))^k \, dt.$$

We can use the inequality transforming $1 - x \le e^{-x}$, so this is at most

$$\int_0^{\infty} e^{-k \cdot \mathrm{area}}\, dt.$$

Now no matter where we begin with y, this area grows quadratically in t — so we have

$$\int_0^{\infty} e^{-\Omega(kt^2)}\, dt.$$

By a change of variables, this final quantity is on the order of $1/\sqrt{k}$. □

> **Student Question.** Since the areas are at least 1, how can we say they grow quadratically?
>
> We can change back to the $\sqrt{2}$ and do the quadratic approximation to a finite amount, and cap the integral there.

Proof of theorem. We will consider the Doob martingale; let

$$L_{n,i}(x_1, \ldots, x_i) := \mathbb{E}[L_n(x_1, \ldots, x_n) \mid x_1, \ldots, x_i].$$

This is a function of the i input points.

> **Claim** — $L_{n,i}$ is $1/\sqrt{n-i+1}$-Lipschitz with respect to the Hamming distance.

When i is close to 1, this is claiming $O(1)$-Lipschitz. But when i is very small, this is $1/\sqrt{n}$, which is very small. So each individual point in the very beginning affects the answer only by a tiny bit.

Proof. If we look at $L(x_1, \ldots, x_n)$, by the inequality above this is at most

$$L(x_1, \ldots, x_i', \ldots, x_n) + 2\text{dist}(x_i, \{x_1, \ldots, x_n\}).$$

We can write this again, dropping all the points before n (unless $i = n$, where we use the bound 2 instead).

Now let's take the expectation over all points we have not yet seen. Then on the left-hand side, we get

$$L_{n,i}(x_1, \ldots, x_i).$$

On the right-hand side, we get that this is at most

$$L_{n,i}(x_1, \ldots, x_i') + 2\mathbb{E}\text{dist}(x_i, \{x_{i+1}, \ldots, x_n\}).$$

So we're finding the distance from a single point to a bunch of random points. And this is bounded by the previous lemma, so we get a bound of $O(1/\sqrt{n-i+1})$. This proves the claim in the case where we flip just x_i, but the same proof works if we change any of the other coordinates (one at a time). □

So the initial points have very little effect on the final outcome. In particular, this tells us that the Doob martingale satisfies the bounded differences

$$|Z_i - Z_{i-1}| = O\left(\frac{1}{\sqrt{n-i+1}}\right).$$

Letting this be c_i and summing over squares, we get a harmonic series, so the outcome (when we square-root) is $\sqrt{\log n}$. So Azuma's inequality gives us sub-Gaussian concentration with width $\sqrt{\log n}$. □

Through this example, hopefully we saw the power of using the martingale rather than just the bounded difference function itself — with bounded differences the worst case is bad at every coordinate, but here we could set up a martingale so that the initial points have very little effect, which makes things much better.

§24.1 A Better Result

We'll now see a better result:

> **Theorem 24.9** (Rhee–Talagrand 1989)
>
> L_n is $O(1)$-subGaussian about its mean.

Remark 24.10. THrough the profile of Talagrand posted on Piazza, Prof. Zhao found out that Rhee is Talagrand's wife; they met while working together.

We'll rephrase this in the following way — what this shows is that

$$\mathbb{P}(|L_n - \mathbb{E}L_n| > t) \leq 2e^{-ct^2}.$$

Remark 24.11. When $t \asymp \sqrt{n}$, we know this is sharp. The paper gives a nice argument — if you consider the event where all the points are crammed into one quadrant of the square, that gives you some relationship between L_n adn the shrunk version, which has exponentially small probability. So this is sharp for some values of t. Prof. Zhao does not know if it is sharp for smaller values of t.

Rhee and Talagrand showed this using martingales and Axuma technieuqes. But we will see anicer proof found later that uses Talagrand's inequality ,which is really nice and exemplifies what the point of Talagrand's inequality is. Prof Zhao was amazed he saw this proof because it's an object he had only seen as a mathematical curiosity, but now we're going to use it.

Definition 24.12. A **space-filling curve** is a continuous map $[0,1] \to [0,1]^2$ that is surjective.

This is a 19th century type object that many of us have heard of. We know they exist, and the first person who exhibited such a curve was Peano (1890), who gave a construction of such a curve. A more popular such curve is given by Hilbert in 1891 (similar to the Peano curve but it has some nice properties).

3b1b has a video on this —- it's a limit of discrete approximations, where we have a sequence of discrete curves that converge pointwise to some curve filling the whole space.

3b1b has a longer video, and one property was important for this to have a well-defined limit. We have this sequence of objects — we have a square with the first curve, and then the second curve where we boxify it, and there's fractal-like behavior. But if we start with a point 0.567 way along the curve, it doesn't move very much as we go to further iterations, so that the limit really does exist pointwise, and it is surjective.

This is something you may have seen as a mathemtaicl curiosity, but we will use it. We need a further property of this space-filling curve:

Definition 24.13. Given two metric spaces X and Y, a map $f\colon X \to Y$ is **Holder continuous** with exponent α if there exists some constant C which may depend on f and α such that

$$d_Y(f(x), f(x')) \leq C d_X(x, x')^\alpha.$$

When $\alpha = 1$, this is the same as Lipschitz continuity. Lipschitz continuity i sstronger than continuity (it's a form of uniform continuity), but it's too much to ask for; the curve is not Lipschitz continuous. Asking for smaller values of α is still asking for something, but it's a weaker condition.

Theorem 24.14

THe Hilbert curve is Holder continuous with exponent $1/2$.

Remark 24.15. If a space-filling curve is Holder continuous with exponent α, then $\alpha \leq 1/2$ — so this is the best kind of Holder continuity we can hope for. The reason for this is that if it's a space-filling curve, then if we chop up the unit interval into k equally spaced intervals

$$[(i-1)/k, i/k],$$

these k intervals must cover the square. So we have k ets which cover the unit square, which means one of the must have large diameter (or else you are not going to cover the entire square). So one of the images has diameter at least $\gtrsim 1/\sqrt{k}$ (or else by area considerations you can't cover the space). So there are some two points whose distance is at most $1/k$, such that their image has distance at least $\gtrsim 1/\sqrt{k}$. So you cannot be Holder continuous with any exponent better than $1/2$. In this sense, the Hilbert curve is optimal.

Proof Sketch. A property of the Hilbert curve, which you can see from the discrete approximation, is that — we define it first as line approximations on dyadic intervals $[(i-1)/4^n, i/4^n]$. In the limit, this interval is sent to some square $[(j-1)/2^n, j/2^n] \times [(k-1)/2^n, k/2^n]$. (This is true for the discrete approximation, so it's true for th elimit.)

So for any $x \neq y$, we can find the least n such that x and y belong to some interval of this form. This means x and y have distance $\Theta(4^n)$, but then their images have distance $|f(x) - f(y)| \lesssim 1/2^n$. (They could be much closer, but they're certainly contained in the box.) This is what we're looking for for Holder continuity. □

This basically proves that the Hilbert curve is Holder continuous with exponent $1/2$.

Student Question. Where does the diameter bound come from?

We have k sets that cover the square. If each of the k ets has diameter $c/\sqrt{k}$, then each is contained in a ball with area $1/k$ (times a small constant). So you cannot cover the whole square.

Student Question. What if we selected points $1/4 - \varepsilon$ and $1/4 + \varepsilon$?

We want to take the largest n instead of the least n.

Isn't this still a problem, if they're very close but separated early on?

You probably need to slide the window a little bit. The argument is not exactly right, but it gives you some flavor of why you should expect $1/2$.

Lst time we presented several formulations of Talagrand's inequality. The one that will be relevant to us is the following:

Theorem 24.16 (Talagrand's Inequality)

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ and $f\colon \Omega \to \mathbb{R}$. Suppose that for every $x \in \Omega$, there exists $\alpha(x) = (\alpha_1(x), \dots, \alpha_n(x)) \in \mathbb{R}^n_{\geq 0}$, such that $f(y) \geq f(x) - \sum_{i | x_i \neq y_i} \alpha_i(x)$ for all $y \in \Omega$. Then $f(x)$ is $O(K)$-subGaussian about its mean, with $K = \sup_x |\alpha(x)|$. (Here we're looking at the Euclidean norm of α.)

We saw this last time, and the interpretation was that the α's measure the resilience of f to changing some of its coordinates.

We would like to apply Talagrand's inequality ,so the remaining goal is to define α — given a set of n points, we want to design some α which works for us.

We *want* to show that there exists $\alpha\colon \Omega \to \mathbb{R}^n_{\geq 0}$ (here Ω is the set of configuration so fn points in the unit square) such that

$$L(x) \leq L(y) + \sum_{x_i \neq y_i} \alpha_i(x)$$

for all x and y in Ω, and furthermore $\sum \alpha_i(x)^2 = O(1)$ for all x. Once we can design an α with this property, we will be done.

Proof. We're going to use the space-filling curve — let $H\colon [0,1] \to [0,1]^2$ be the Hilbert space-filling curve. The intuition is that $\alpha_i(x)$ is something like the difficulty of serving x_i — so we want to give some measure of how difficult it is to serve the ith point. We coudl choose something more intuitive, like the distance from x_i to its nearest point (that would agree with what we want), but simply assigning that doesnt' really capture the structure of TSP. So instead what we'll use is the Hilbert space-filling curve.

We have a set of points, and we have some curve (which we are not going to be able to draw) which goes through the entire square. In particular, it will go through each of these points. The curve is not injective and may go through the points ultiple times, but that is fine.

Now we order the points along their appearance on H; let this order be σ, so that $x_{\sigma(1)}, \ldots, x_{\sigma(n)}$ appear in that order on the Hilbert curve.

Now let $\alpha_i(x)$ be twice the Euclidean distance from x_i to its previous point plus the distance to the next point — in other words, if we have our x_i, then α_i is the distance of the segment to the previous point plus the distance of the next segment, multiplied by 2.

This is some measure of the difficulty of serving the ith point. So we need to prove a couple of things.

First we need to prove the first inequality (1), that

$$L(x) \leq L(y) + \sum \alpha_i(x).$$

What this says is — first, the trivial case is when $x_i \neq y_i$ for all i. In that case, this is a trivial inequality, since the sum $\sum_i \alpha_i(x)$ is twice the length of the tour induced by the Hilbert curve, which upper bounds the TSP length (even ignoring $L(y)$ on the right).

Now let's suppose we pick an optimal tour through y. Now we need to figure out how to get a tour through x that doesn't cost too much more. We know x has a bunch of shared points, and some non-shared points. For the non-shared points, these are somewhere else, and we need to figure out how to reach them.

So what we can do is start by going along the tour for y, and now we're going to make some detours — we're going to try to reach the other points. TO do that, let's think about all these new points, and draw the two segments coming out of their Hilbert curve induced tour. We don't know which order the Hilbert curve goes through them, but we have all these extra points.

Now once we draw in all these extra edges, the whole thing becomes a connected graph. So we can first go around the blue thing, and then replay — whenever we need to go off-path to reach something, we go reach it, come back, and continue on the blue tour. So we simply add some excursions ot reach all the new points in $x \setminus y$. The cost in doing these excursions is basically twice the sum of these lengths that we have to consider, since the excursion involved maybe goin goff-path to somewhere and ocming back to continue our tour. So that gets us the first inequality.

Now we need to prove the second, that for every choice of x, $\sum \alpha_i(x)^2$ is at most a constant.

This quantity is, once we decide the points in x, there is some order determined by the Hilbert curve through them. So this is at most $\lesssim \sum_{i=1}^n (\lvert H(t_i) - H(t_{i+1})\rvert + \lvert H(t_i) - H(t_{i+1})\rvert^2)$, where we write $t_1 \leq \cdots \leq t_n$ to be the points on the interval mapping to our x_i. We can now bring the square inside and delete one of hte terms (losing some constants). Now using the fact that our thing is Holder continuous, this is at most $\sum_{i=1}^n \lvert t_i - t_{i+1}\rvert$. So here these are increasing points in the unit interval, which means the sum is at most 2. That finishes the proof. □

Intuitively, α records the difficulty of serving our points, and it will depend on the points. And the way we define it is by obtaining a tour through them based on the space-filling curve. THis is caleld the space-filling curve heuristic; it is not necessarily the shortest tour, but it is not necessarily bad. Then α will assign the length of the two edges coming out of the induced tour. We first show that if we start with some existing tour y and someone changes the points in y to some new points x, we have to prove an upper bound through stuff in x, not costing too much more than the tour through y. And those excursions involve only the edges that came up. So teh cost of these excursions is basically the sum of the α_i. That fihisnes the fisr part.

The second part is a bound on the L_2 norm of the α's and that pfollows from Holder continuity with exponent $1/2$, noting that we're travelling in this map along the segment in the same order, so that the sum is at most constant (we're not going back and forth on the interval).

Student Question. In the problem, the ordering of hte points doesn't matter, but in our application of Talagrand it does come up, right?

We can think of the points as unlabelled points if we like; we are then giving them an order, which is the order they appear on the Hilbert curve. You might be curious what people do if I actually give you n points and need to find a short tour through them. One method is called the **space-filling curve heuristic**, which is not exeactly this —- but you pretend there is some underlying space-filling curve, and try to go through the points as if travelling along that. If your points are densely populated in the unit square, that is often not a bad strategy.

Student Question. Does this work with any space-filling curve which is Holder continuous with exponent $1/2$?

Yes. In fact, the original Peano curve has the same property.

Student Question. Is there any chance of classifying space-filling curves that have exponent $1/2$?

Prof. Zhao doesn't knwo.

Next time we'll start a nnew chapter on the entropy method.

§25 December 5, 2022 — The Entropy Method

We'll spend the next three lectures on the entropy method, a powerful and beautiful technique.

§25.1 What is Entropy?

Definition 25.1. For a discrete random variable X, we define its **entropy** as

$$H(X) = \sum_{s \in S} -p_s \log_2 p_s,$$

where S is the set of possible values of X, and $p_s = \mathbb{P}(X = s)$.

This has an important meaning for information theory — Claude Shannon (the father of information theory) found that this captures something about the amount of information in a random variable. Informally speaking, $H(X)$ is the amount of information in the random variable X, or the amount of 'surprise.'

Example 25.2

If X is uniformly distributed among n elements, then each p_s is $1/n$, and

$$H(X) = \log_2 n.$$

A way to interpret this is that if I need to encode X, I need to send you $\log_2 n$ bits to tell you the outcome of this random variable.

Example 25.3

If X is deterministic, then there is no surprise (you learn nothing by revealing the value of X), and the entropy is 0.

This can be made formal, and this is known as Shannon's sourcing theorem. Roughly, to communicate n i.i.d. copies of X, we need to send $n(H(X) + o(1))$ bits — and it is possible to do so. So $H(X)$ is asymptotically the information content in X.

We will mostly be concerned with combinatorial applications (where we'll look at a problem that is not random, introduce randomness and analyze its entropy, and be able to get cool results this way).

§25.2 Basic Properties of Entropy

(These properties are fairly straightforward and can be proved by convexity, so we won't go over them; the proof is in the lecture notes.)

Proposition 25.4 (Uniform Bound)

$H(X) \leq \log_2 |\text{support}(X)|$.

The way to interpret this is that if you tell me your random variable takes n values, then the worst case (in the sense of having the most entropy) is the uniform random variable. So to maximize the surprise stored in the random variable (given the number of possible values it can take), you want to spread it out evenly.

Definition 25.5. We use $H(X, Y)$ to mean the **joint entropy** — the entropy of the random variable (X, Y).

Proposition 25.6

If X and Y are independent, then $H(X, Y) = H(X) + H(Y)$.

The intuition is that if X and Y are independent, then when you reveal X to me, I learn nothing about Y; the information I get from learning X and Y together is the same as from learning them separately.

Definition 25.7. The **conditional entropy** is defined as

$$H(X \mid Y) = \mathbb{E}_y H(X \mid Y = y) = \sum_y \mathbb{P}(Y = y) H(X \mid Y = y),$$

where $H(X \mid Y = y) = \sum_x \mathbb{P}(X = x \mid Y = y) \log_2 \mathbb{P}(X = x \mid Y = y)$.

Proposition 25.8

We have $H(X \mid Y) = H(X,Y) - H(Y)$.

The way to interpret conditional entropy is that it's the expected amount of new information learned from X after already knowing Y — suppose you reveal Y, and ask how much more information I'll learn by revealing X. In *expectation*, this gives us the conditional entropy.

We can also see this from the second formula — it's the total information learned from seeing X and Y jointly, minus the information we got from revealing Y (which has already happened and so shoudn't be counted).

Example 25.9

If $X = Y$, then $H(X \mid Y) = 0$ — there is no surprise anymore in X once I have already revealed Y.

Example 25.10

If X and Y are independent, then once someone tells us Y, we haven't learned anything about X — so anything we see when we reveal X is new, and $H(X \mid Y) = H(X)$.

Something we'll use quite a lot is the chain rule:

Proposition 25.11 (Chain Rule)

We have $H(X,Y) = H(X) + H(Y \mid X)$. Similarly,

$$H(X,Y,Z) = H(X) + H(Y \mid X) + H(Z \mid X,Y),$$

and so on.

Proposition 25.12 (Subadditivity)

We have $H(X_1, \ldots, X_n) \leq H(X_1) + \cdots + H(X_n)$.

One way to interpret this is through the chain rule — $H(Y \mid X) \leq H(Y)$, since the amount of information we learn from Y after seeing X is upper-bounded by the amount of information in Y on its own (seeing X can only reduce the surprise). We can rephrase this as *dropping conditioning*:

Proposition 25.13 (Dropping Conditioning)

We have $H(X \mid Y) \leq H(X)$. More generally $H(X\ midY, Z) \leq H(X \mid Z)$.

The amount of surprise in X after revealing Y certainly can't be greater than the amount of surprise in X on its own.

Remark 25.14. The data processing inequality is very much related to this property.

These are all the basic properties we'll use; the main point of today's lecture is to see some combinatorial applications.

§25.3 Tail Bounds of Binomial Distributions

Notation 25.15. We'll use $H(p)$ to be the entropy of the Bernoulli variable with probability p — so $H(p) := H(\mathrm{Ber}(p)) = -p\log_2 p - (1-p)\log_2(1-p)$.

(This is a bit of abuse of notation since we're using H in two different ways — you might complain that p is a constant, and H of a constant should be 0. But these are intended to be two different meanings; this abuse of notation is standard.)

Theorem 25.16

If $0 \le k \le \frac{n}{2}$, then

$$\sum_{i\le k}\binom{n}{i} \le 2^{H(k/n)n} = \left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k}.$$

One way to prove this inequality is to run the proof of the Chernoff bound for this setting. But we'll now see how to do it using entropy, to get a feel for where this number comes from without taking derivatives and optimizing some steps.

Proof. Let $(X_1,\dots,X_n) \in \{0,1\}^n$ be chosen uniformly, conditioned on $X_1+\cdots+X_n \le k$. So the number of psosibilities for this vector is precisely the desired sum, and it's uniformly distributed among these possible vectors. This means

$$\log_2\sum_{i\le k}\binom{n}{i} = H(X).$$

Now we can use some of these entropy inequalities — we have

$$H(X) = H(X_1,\dots,X_n) \le H(X_1)+\cdots+H(X_n)$$

by subadditivity. Now we wish to find $H(X_i)$. X_i only takes values 0 and 1, so it is a Bernoulli random variable with some probability. To find this probability, we have

$$\mathbb{E}[X_i] = \frac{\mathbb{E}[X_1+\cdots+X_n]}{n} \le \frac{k}{n}.$$

So then X_i is a Bernoulli random variable with $H(X_i) = H(p)$, where $p \le k/n$. Since $H(p)$ is an increasing function up to $1/2$, this means $H(X_i) \le H(k/n)$. Plugging this back in, we get

$$\log_2\sum\binom{n}{i} \le nH\left(\frac{k}{n}\right),$$

and comparing the two sides finishes the proof. □

Remark 25.17. There are also other forms of entropy — Rényi entropy and differential entropy.

Student Question. How big is this bound?

It's pretty good — if k/n is a constant, then this is the right constant. More generally, if we consider a binomial with p instead of $1/2$, then the right answer has to do with *relative entropy.*

In the rest of the lecture, we'll see some more clever and substantial applications, which are really gems of the field.

§25.4 Permanents

Definition 25.18. The **permanent** of a $n \times n$ matrix A is defined as

$$\operatorname{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)}.$$

So we look at all ways of choosing exactly one entry from each row and column, take their product, and sum over combinations. The determinant has a similar formula, but it's multiplied by the sign of the permutation; the permanent is essentially obtained from erasing the sign.

This is interesting for many reasons; it's difficult to compute, so it has a role in complexity theory.

As an interpretation, the permanent of a bipartite adjacency matrix is equal to the number of perfect matchings. (If the 1's and 0's correspond to edges and non-edges, then choosing all 1's corresponds to a perfect matching between the rows and columns.)

The main result we'll see is an inequality that upper-bounds the permanent of a $\{0,1\}$-matrix.

Theorem 25.19 (Bregman–Minc Inequality)
If $A \in \{0,1\}^{n \times n}$ has ith row sums d_i, then

$$\operatorname{per}(A) \le \prod_{i=1}^{n} d_i!^{1/d_i}.$$

In the bipartite case, someone gives us information about the degrees on the left, and we want to maximize the number of perfect matchings. This upper bound is in a sense tight — if our graph is a disjoint union of complete bipartite graphs, then the number of perfect matchings in each part is $d_i!$, and we have d_i such terms (which cancels out with the exponent of $1/d_i$). So equality occurs when our matrix consists of a bunch of square blocks which are all 1's, and 0's everywhere else. (The contribution from each block is precisely the size of the block factorial.)

The proof we'll see is not the original proof, but an entropy proof found later due to Radhakrishnan 1997.

Proof. We'll use entropy by first defining a random variable uniform on its support, whose support has size $\operatorname{per}(A)$ — let σ be a random permutation of $[n]$ which is uniform conditioned on $a_{i\sigma(i)} = 1$, or in other words a random transversal of 1's or a uniform random perfect matching (among all perfect matchings in the graph). Then σ is uniform on a set of size $\operatorname{per}(A)$, since $\operatorname{per}(A)$ counts the number of such permutations by definition. This means

$$H(\sigma) = \log_2 \operatorname{per}(A).$$

Just like before, we can decompose σ into its coordinates, as

$$H(\sigma) = H(\sigma_1, \dots, \sigma_n).$$

We're now going to apply some entropy inequalities.

As a first attempt, we can first apply the chain rule to write

$$H(\sigma) = H(\sigma_1) + H(\sigma_2 \mid \sigma_1) + H(\sigma_3 \mid \sigma_1, \sigma_2) + \cdots$$

and try to bound the individual terms. First, σ_1 tells me which 1 we should pick among all the 1's on the first row. This is a random choice, though it's quite likely not uniformly distributed; we don't have information on *how* it's distributed, but we can use the uniform bound to write $H(\sigma_1) \le \log_2 d_1$.

The second quantity says, after seeing which choice we made on the first row, which choice do we want to make on the second row? This is a conditional entropy, so it's complicated — it depends on how we chose the first row. Some choices of the first row might reduce the number of possibilities for the second row. So that is kind of complicated — it's more difficult than what we had to do on the first row. And if we ignore that and just use the uniform bound, then we have $\log_2 d_2$. And we can keep doing this — so then we get

$$\log_2(d_1 d_2 \cdots d_n).$$

This is pretty bad — it's what would happen if we replaced $d!$ by d^d, which is quite lossy.

But we didn't do anything smart — this would be the same as counting the number of possibilities in the first row, then the second, and so on (which is high school combinatorics). So nothing smart happened here, even though it was written in the language of entropy.

The trick is — here we revealed the rows one by one in a prescribed order. But instead we're going to reveal the rows in a uniform *random* order.

Let $\tau = (\tau_1, \ldots, \tau_n)$ be a uniform random permutation of $[n]$. We're going to use the chain rule, but we now reveal the choices in the order prescribed by the random permutation τ — so

$$H(\sigma) = H(\sigma_{\tau_1}) + h(\sigma_{\tau_2} \mid \sigma_{\tau_1}) + H(\sigma_{\tau_3} \mid \sigma_{\tau_1}, \sigma_{\tau_2}) + \cdots.$$

(This chain rule is true for *every* (deterministic) choice of τ.)

Let's now look at one of the terms — let's look at the term $H(\sigma_i \mid \cdots)$ corresponding to σ_i, which has some random number of conditioned information (depending on the choice of τ). So we fix i and define k such that $\tau_k = i$ (which is random). Then our relevant expression is

$$H(\sigma_i \mid \sigma_{\tau_1}, \ldots, \sigma_{\tau_{k-1}}).$$

After having seen σ_{τ_1}, ..., $\sigma_{\tau_{k-1}}$, the expected number of remaining choices for σ_i is uniformly distributed in $[d_i]$ — initially the ith row had d_i 1's. In some random order, the other rows get revealed. Sometimes one of the other revealed rows will knock out one of these possibilities. So we want to see, among the 5 1's, how many of them don't get knocked out before we reveal the row. To figure this out, if we fix σ which is supposed to assign some choice to each element, so we want to see of the five relevant rows, where in the relative order the ith row comes up in τ.

So the number of remaining choices for σ_i is uniformly distributed in $[d_i]$, with respect to the randomness of τ. (In fact this is true no matter what σ we have placed.)

> **Student Question.** What if you don't have 1's in any of those columns?

Answer. Then the permanent is 0. We're assuming the permanent is not zero.

Think about a fixed σ — we don't know what it is, but it's there (and hidden). Someone reveals to us the rows one by one. By the time the ith row comes up, we want to see how many choices we have left. If i came up first, then we haven't seen the selection of i, so we have 5 choices left. But if someone has shown us a certain other row, then we only have 4 left.

And given σ, that probability distribution depends on when i comes up relative to the other five rows, which is uniform.

Imagine you have a fixed σ below the table, and you're trying to reveal what it is (it's there but you don't know what it is). Then this is true. But then if we let σ vary, it's still true.

Now by the uniform bound,

$$H(\sigma_1 \mid \sigma_{\tau_1}, \ldots, \sigma_{\tau_{k-1}}) \le \mathbb{E}_{\sigma,\tau} \log_2 \operatorname{support}(\sigma_1 \mid \sigma_{\tau_1}, \ldots, \sigma_{\tau_{k-1}})$$

But even for a fixed σ, this is uniform. So this is *equal* to

$$\frac{1}{d_i}(\log_2 1 + \log_2 2 + \cdots + \log_2 d_i) = \frac{1}{d_i} \log d_i! \, .$$

This upper-bounds the term corresponding to σ_i, and summing over all i's gives the desired formula. □

This is a bit confusing, but the main point is that instead of doing the naive thing, we reveal the rows in a random order. That will allow us to much better analyze what's happening, and get the better bound.

Remark 25.20. Some clarifications: when we talk about the number of possibilities, think about a *fixed* σ. (The answer will be the same for every σ.) Then τ is the row reveal order, and we're fixing i and looking at the ith row.

Then when we say the number of possibilities for the ith row when it comes up, we mean the following: the ith row has, let's say 5 1's. Then what we mean is — σ is fixed, so σ is going to make some choice such as placing circles in some certain places on these rows. Then the number of possibilities is 5 if i comes up first among these five rows, 4 if it comes up second, 3 if it comes up third, and so on — because if the row above it appears first, then this eliminates the possibility. (You can naively no longer choose this entry, because you've already chosen an entry in the same column. So you can replace 'number of possibilities' by whatever this is.)

§25.5 Steiner Triple Systems (STS)

Definition 25.21. A **Steiner triple system** is a 3-uniform hypergraph on n (labelled) vertices, where every pair of vertices is contained in exactly one triple.

Another way to say this is that it's a decomposition of K_n into edge-disjoint triangles — every pair is used exactly once.

The first nontrivial STS is the Fano plane — we have 7 vertices, and if our edges are as in the Fano plane, then every pair of vertices is contained in exactly one triple.

There are some basic questions about when these exist. The number of edges of K_n is $\binom{n}{2}$, so then $\binom{n}{2}/3$ must be an integer. Likewise, $(n-1)/2$ must be an integer (fixing a vertex and looking at the edges coming out of it). Together, these are the same as requiring $n \equiv 1, 3 \pmod 6$. In fact this is an iff condition — Steiner triple systems exist if and only if $n \equiv 1, 3 \pmod 6$.

Question 25.22. How many STSs are there?

This turns out to be quite a difficult quwestion, and it was not fully resolved until a paper from 2018 (based on design theory).

We will look at another, somewhat recent, result:

Theorem 25.23 (Liniel, Luria 2013)

The number of STS on n vertices is

$$\mathrm{STS}(n) \le \left(\frac{n}{e^2 + o(1)}\right)^{n^2}.$$

(The matching lower bound was shown by Keevash.)

Exercise 25.24. Show that this is a reasonable bound (by trying to count and making unjustified assumptions).

Proof. As before, we will pick a uniform STS X of order n, and we will try to upper bound

$$H(X) = \log_2 \mathrm{STS}(n).$$

X is a decomposition of K_n into triangles, and as before we will try to reveal this decomposition edge-by-edge — and we will try to reveal it in some random order.

First we need some way of encoding X. We will encode X as a tuple $(X_{ij})_{i<j}$ which we can think of as putting labels on the edges — we initially have a vertex-labelled graph, and we'll now put a label on each of the edges so that each label is the vertex label of the third vertex in the triple containing the edge. For example, if we had the STS of one triangle, then we'd put a 3 on the edge 12.

Every STS has a label of this form, and this specifies the STS. We would like to know how many possibilities there are.

As before the break, the method is to reveal the labels in some random order, and use the chain rule on these random variables revealed in a random order. In order to actually do this proof cleanly, we're not just going to reveal them in a random order; rather, we are going to assign a random *time* where we reveal the edges. So we assign i.i.d. $y_{ij} \sim \mathrm{Unif}[0,1]$ corresponding to the time when we reveal the label on the edge ij. Then the random vector y determines an order of the edges of K_n, in the sense that we say $k\ell \prec ij$ if $y_{kl} > y_{ij}$. (So we're revealing from time 1 to 0.)

(We've seen this idea earlier in the class — when we looked at two-coloring hypergraphs, there was a proof using a greedy random coloring, where instead of choosing a permutation we chose random uniform labels. This is a similar idea.)

For a given ordering, we can write down the chain rule

$$H(X) = \sum_{ij} H(X_{ij} \mid X_{k\ell} \text{ for all } k\ell \prec ij).$$

(We're conditioning on all the edges we've already revealed.) Let N_{ij} be the number of possibilities for X_{ij} after revealing all th earlier $X_{k\ell}$ (for $k\ell \prec ij$). So then N_{ij} is also a random variable, for each fixed ij — through some process I slowly start to see the other labels and at some point I need to reveal ij, and before I reveal it, there's some number of possibilities that we'd like to upper bound.

The quantity N_{ij} will depend on both the STS we chose and the random variables y that we injected. By the uniform bound, we see that

$$H(X) \le \sum_{i,j} \mathbb{E}_X \log N_{ij}$$

for every choice of y (once we fix y, that determines the order of the chain decomposition, and then we have this rule). Taking an expectation over the randomness of y, we get

$$H(X) \le \sum_{ij} \mathbb{E}_{X,y} \log N_{ij}.$$

Write y_{-ij} to be y with its ij-coordinate removed (so we erase the coordinate, to get a vector with one fewer coordinates). We are going to upper bound

$$\mathbb{E}_{y_{-ij}} \log N_{ij}$$

as a function of y_{ij} — if someone tells us the label on ij (suppose I knew a priori that the label is 1/3), I now want to know, in expectation, by the time that edge comes up in the ordering, how many possibilities remain.

Define an event that $\{(i,j)$ shows up first in its triple$\}$ — i.e. $\{ij \prec ik, jk$ where $k = X_{ij}\}$. The probability of this event is

$$\mathbb{P}_{y_{-ij}}(\{(i,j) \text{ shows up first in its triple}\}) = \mathbb{P}(y_{ik}, y_{jk} < y_{ij}).$$

Here y_{ij} is fixed and the other values are uniform, so this is y_{ij}^2 (k is random as well, but this is true conditioning on any value of k).

If ij does *not* show up first in its triple, then we don't have any flexibility for ij — someone else has already told us where ij is supposed to be (part of the other triple). So ij has one possibility. In particular, this means $N_{ij} = 1$, so its log is 0.

We want to evaluate $E_{y_{-ij}} \log N_{ij}$. We saw that if i does not appear first in its triple, then this is 0, so we can ignore it. Otherwise, there's a probability y_{ij}^2 that it appears first in its triple, and we want to find $\mathbb{E}_{y_{-ij}}[\log N_{ij} \mid ij$ first in triple$]$.

At this point, we will now apply convexity to bring the expectation inside the log. So this is at most

$$y_{ij}^2 \log \mathbb{E}_{y_{-ij}}[N_{ij} \mid ij \cdots].$$

Now we're considering the expectation of the number of available possibilities. This is a count, so we can use linearity of expectation — for each $s \in [n] \setminus \{i,j,k\}$, if s remains a possibility for X_{ij} by the time that X_{ij} is revealed, then it must have not been used in anything that could have been a conflict.

The probability that any of the yellow edges show up before ij is precisely $1 - y_{ij}^6$ (consider the triangles involving is and js — then the 6 edges are the three from both triangles of is and js).

So then

$$EE_{y_{-ij}}[N_{ij} \mid ij \text{ first in triple}] \le 1 + (n-3)y_{ij}^6$$

(k is always a possibility, all the other s's come up with probability at most our bound, and we use linearity of expectation). Plugging into our bound, we get

$$\mathbb{E} \log N_{ij} \le \int_0^1 z^2 \log(1 + (n-3)z^3)\, dz.$$

At this point, we can change variables letting $t = z^3$. This does have a closed-form antiderivative, but it doesn't really matter since we only care about asymptotics; asymptotically this is

$$\int \log\left(\frac{1}{n-3} + t^2\right) dt \to \int \log(t^2)\, dt = -2$$

by monotone convergence. SO then our entire thing goes to

$$\frac{\log n - 2 + o(1)}{3}.$$

So now we have $\binom{n}{2}$ terms, each of which we have upper-bounded in this way — so then we get

$$H(X) \le \binom{n}{2}\left(\frac{\log 2 - 2 + o(1)}{3}\right) = \frac{n^2}{6} \log\left(\frac{n}{e^2 + o(1)}\right).$$

□

§26 December 7, 2022

Today we will continue our discussion of the entropy method.

§26.1 Review of Properties

Definition 26.1. Given a discrete random variable X, the entropy of X is defined as

$$H(X) = \sum_s -\mathbb{P}(X = s) \log \mathbb{P}(X = s).$$

If you like, log means $\log_2$ throughout. But it usually doesn't matter, as long as you're internally consistent. (Last time Prof. Zhao made a small mistake where he integrated log and got -1 instead of $-\log e$; but it doesn't affect the answer.)

We have a number of equalities and inequalities:

Proposition 26.2 (Uniform Bound)

We have $H(X) \le \log |\text{support}(X)|$.

In other words, if you know nothing about X except the possibilities it can have, its entropy is maximized when it's split evenly between the possibilities.

Proposition 26.3

If X and Y are independent, then $H(X, Y) = H(X) + H(Y)$.

Definition 26.4. The **conditional entropy** $H(X \mid Y)$ is $\mathbb{E}_y H(X \mid Y = y)$ — the expected entropy of the conditioned random variable X on Y, letting Y vary according to its distribution.

The conditional entropy is also $H(X, Y) - H(Y)$ — this is the typical additional amount of information you learn from X after seeing Y.

Proposition 26.5 (Chain Rule)

We have $H(X, Y) = H(X) + H(Y \mid X)$.

Proposition 26.6 (Sub-additivity)

We have $H(X, Y) \le H(X) + H(Y)$.

Proposition 26.7 (Condition Dropping)

We have $H(X \mid Y, Z) \le H(X \mid Z)$.

In other words, if we condition on some random variables and we erase some of these variables, the entropy can never go down.

Last time, we tried to emphasize the information theoretic meaning of each of these — they all hopefully have an intuitive interpretation in terms of the amount of surprise in the random variable, which will be useful for remembering these inequalities.

Our goal is to see several neat combinatorial applications of entropy. Last time we first used entropy to bound tails of binomial distributions. We also saw an application to bounding the permanent of a $\{0, 1\}$ random matrix, and an application to counting Steiner triple systems.

Today we will see a different line of applications, which has to do with a major conjecture in graph theory called *Sidorenko's conjecture* (this is one of Prof. Zhao's favorite conjectures in graph theory).

§26.2 Sidorenko's Conjecture

Definition 26.8. Given graphs F and G, a **graph homomorphism** from F to G is a map $\varphi\colon V(F) \to V(G)$ such that $\varphi(u)\varphi(v)$ is in $E(G)$ whenever uv is in $E(F)$ — i.e., it sends edges of F to edges of G.

What this is supposed to capture is — suppose we have a big graph G (such as a diamond with one line) and a small graph F (say a carrot), and we want to see the number of ways to map the vertices of F into vertices of G such that edges are preserved. The top vertex of the carrot could go to the top vertex of G (of degree 2); the first leaf could go to one of its neighbors; the other leaf could still go to either neighbor (the map doesn't have to be injective).

Definition 26.9. $\hom(F, G)$ is the number of homomorphisms from F to G.

This is closely related to subgraph counts — if F were a triangle, then it'd be 6 times the number of triangles in G.

Definition 26.10. The **homomorphism density** (or F-density in G) is defined as

$$t(F, G) = \frac{\hom(F, G)}{v(G)^{v(F)}}.$$

In other words, it's the probability that a uniform random map from the vertices of F to G is a homomorphism.

This notion captures a lot of different concepts. For today we will think of G as large, and F as small and fixed. Then $\hom(F, G)$ is basically the number of subgraphs of F in G. We need to make a few corrections — first, we need to multiply by a factor of $\operatorname{aut}(F)$ (for a triangle, we get a factor of 6). But we also have to account for the fact that a homomorphism doesn't have to be injective, while the number of copies counts injective maps. It turns out that's a small discrepancy, so the error term is of a lower order —

$$\hom(F, G) = \operatorname{aut}(F) \cdot \#\{\text{copies of } F \text{ in } G\} + O_F(v(G)^{v(F)-1}).$$

Question 26.11. What inequalities between these quantities are true? More specifically, given a fixed small graph F and a constant $p \in [0, 1]$, what is the minimum possible F-density in a graph with edge density at least p?

Here edge density means K_2-density — the number of homomorphisms from K_2 to G, divided by the total number of homomorphisms — or

$$\frac{2e(G)}{v(G)^2}.$$

So someone gives us p (for example $1/2$) and asks us, of all the graphs with edge density $1/2$, which one minimizes the number of copies of F?

For some F, we do know the answer.

Example 26.12

If F is a triangle and $p = 1/2$, then the graph G which minimizes the F-density is a complete bipartite graph — it has edge density $1/2$ and no triangles.

When $p > 1/2$, we also know the answer; this is discussed in Prof. Zhao's book.

Question 26.13. What happens when F is bipartite — for example a 4-cycle?

Here something very different happens. The complete bipartite graph has a lot of F's; it turns out that intuitively what we *expect* the answer to be is the following:

Conjecture 26.14 (Sidorenko's conjecture, informally) — The minimizing G is a random graph.

Essentially, for any bipartite F, to minimize F-density for a given edge density, we should take a random graph.

More precisely:

Conjecture 26.15 (Sidorenko's conjecture) — For any bipartite F and any G, we have

$$t(F, G) \geq t(K_2, G)^{e(F)}.$$

One way to interpret this is that if we fix F and ask which G minimizes this quantity provided that we know what the right-hand side is, we see that the right-hand side is basically the edge-density of F in a random G — so it's what you'd expect the edge-density to be in a random graph.

It is still a conjecture; but it has been proved for some classes of F, and we will see some examples today. But it is open for *most* graphs F, and there are even people who suspect that maybe the conjecture is false.

How might we prove this conjecture for certain graphs F? We want to prove a lower bound on F-density in G, or homomorphism count. Last time, we used entropy to prove *upper* bounds on counts in the following way — we had some space we wanted to count, and we wanted a bound $|\Omega| \leq \bullet$. We then let $X \in \text{Unif}(\Omega)$, so that on one hand $H(X) = \log |\Omega|$, and on the other hand we applied various entropy inequalities to prove an upper bound on $H(X)$; that yielded an upper bound on $|\Omega|$. But now we want a *lower* bound.

So how can we use entropy to prove lower bounds?

What we are going to do is construct μ to be a probability distribution supported on the set $\text{Hom}(F, G)$ — so we are going to construct a random homomorphism from F to G, but not uniformly. Then we know that $H(\mu)$ (i.e. H of a random variable drawn from this distribution) is upper-bounded by $\log |\text{support}(\mu)| \leq \log \hom(F, G)$. This allows us to get alower bound on teh number of homomorphisms if we can lower-bound $H(\mu)$.

This is true for any μ, but the point is that if we can cleverly construct μ such that we can get a lower bound on $H(\mu)$, then we win — so the cleverness is in constructing a random (not necessarily uniform) homomorphism.

The inequality we wish to show can be rewritten in terms of counts — as

$$\frac{\hom(F, G)}{v(G)^{v(F)}} \geq \left(\frac{2e(G)}{v(G)^2}\right)^{e(F)}.$$

So it suffices to construct a probability distrubution μ on $\text{Hom}(F, G)$ so that

$$H(\mu) \geq e(F) \log(2e(G)) - (2e(F) - v(F)) \log v(G).$$

If we can find μ such that this inequality is true, then plugging this in gives the desired inequality.

§26.3 Sidorenko for Trees

Theorem 26.16 (Blakey–Roy 1965)

Sidorenko's conjecture holds for a graph which is a 3-edge path.

This looks easy — it's a very concrete inequality about counting walks in paths, or what happens to the sum of entries in a matrix when we raise it to the third power. But it's actually not easy.

The proof we'll see using the entropy method is due to Li–Szegedy 2011.

Proof. Label the vertices of our path by x, y, z, w. Our goal is to find a distribution of walks of length 3 in a graph.

Define a random walk $XYZW$ in G as follows:

- XY is a uniform random edge — we choose one of the edges uniformly at random, and then decide on the direction uniformly at random. Note that X is not distributed uniformly randomly among vertices — it's distributed with probability proportional to its degree. Meanwhile, conditioned on X, Y is uniform among its neighbors. (You can think about counting.)
- We then pick Z to be a uniform random neighbor of Y.
- We then pick W to be a uniform random neighbor of Z.

This gives a random homomorphism from a 3-edge path to G. This is very different from picking a walk uniformly at random.

The key observation is the following: Z depends on Y but not on X (once we see Y we can forget about X and figure out what Z is). Similarly once we see Z we can forget about X and Y when trying to figure out what W is. So how is YZ distributed? It's still a uniform random edge. This is because Y is distributed the same way that X is — it's also distributed proportionally to its degree. So the distribution of YZ is the same as XY; in particular, it is uniform. And the same is true for ZW.

Another way to see this is that conditioned on the choice of Y, X and Z are independent and uniform neighbors of Y — we can make use of symmetry. (The word is a 'reversible Markov chain' — instead of first picking X you could first pick Y, and then X and Z are basically identical copies of each other.)

This observation tells us that $H(Z \mid X, Y) = H(Z \mid Y)$ (it does not make a difference if we forget about X — this is a statement about conditional independence, or the Markovian property. Intuitively once we learn X and Y and we're trying to figure out how much information there is in Z, we might as well forget X, because once we know Y then X doesn't influence Z anymore). Likewise, $H(W \mid X, Y, Z) = H(W \mid Z)$. (Both of these statements are true by conditional independence.)

Now we have

$$H(X, Y, Z, W) = H(X) + H(Y \mid X) + H(Z \mid X, Y) + H(W \mid X, Y, Z)$$

by the chain rule. And then by this conditional independence observation, we can simplify some of the terms by dropping the unnecessary conditioning, to get

$$H(X) + H(Y \mid X) + H(Z \mid Y) + H(W \mid Z).$$

In this expression, the last three terms are all the same — Y conditioned on X, Z conditioned on Y, and W conditioned on Z all have the same distribution. So

$$H(X, Y, Z, W) = H(X) + 3H(Y \mid X).$$

We can run the chain rule once again in reverse, to get that this is

$$3H(X, Y) - 2H(X).$$

Finally, since XY is uniformly distributed, then $H(X,Y) = \log(2e(G))$ (because XY and YX are different). Meanwhile we could figure out $H(X)$ explicitly because we know the explicit distribution, but it's enough to use the uniform bound $H(X) \leq \log v(G)$. Then we get precisely the inequality we were trying to show (once we compare coefficients) — this is

$$3\log(2e(G)) - 2\log v(G),$$

which is what we needed to show and we're done. □

That's a proof of Sidorenko's inequality for the three-edge path. (You can try this problem without these methods to convince yourself that this problem is not trivial.)

First, this works for any number of edges. But it actually does even more — the same proof also works for trees. Here we would define the random homomorphism by selecting the vertices one by one — we first pick a uniform edge, and then when we need to grow another edge we pick a uniform neighbor, and so on. (We don't need to worry about collisions.)

So this works basically verbatim for trees. We'll now look at the next interesting case in this context.

Theorem 26.17

Sidorenko's conjecture holds for a 4-cycle, and in fact for all $K_{s,t}$.

(We'll prove it for the 4-cycle, but the same proof works for $K_{s,t}$.)

It turns out this is actually not too hard to show using Cauchy–Schwarz; the entropy proof is actually trickier, but it leads to more things.

Unlike before, this process no longer works because we have a cycle — so we need to figure out how to construct a random homomorphism from the 4-cycle.

We'll again try to construct ν so that this inequality is satisfied. Label the vertices x_1, x_2, y_1, and y_2. We want a way to embed these vertices into G.

As before, we pick x_1y_1 to be a uniform edge, and then we choose y_2 to be a uniform neighbor of x_1.

Now we need to embed x_2. If we embed by growing from y_2 as a uniform neighbor, we will run into problems because we also need x_2 and y_1 to be adjacent.

But we already know that x_1 is a common neighbor of y_1 and y_2. So we will just make a copy of x_1 — x_2 is a conditionally independent copy of x_1, conditioned on y_1 and y_2.

What this means is — we've initially defined a probability distribution on a 2-edge path. But then we can ask for the marginal of this distribution on y_1 and y_2 — we can certainly generate this distribution by picking y_1 and y_2 according to some distribution, and then picking x_1. When we do that, we can instead throw in another copy — picking x_2 according to that distribution as well. Then we get that our two desired edges are present for free; so this is a well-constructed homomorphism from a 4-cycle to G.

So we can think of (x, y_1, y_2) as a distribution. We could generate this distribution instead by first picking (y_1, y_2) according to its marginal distribution; and then instead of generating one x_1, we generate two independently.

Now we can do a similar calculation to before. First, we have

$$H(X_1, Y_1, Y_2) = 2H(X_1, Y_2) - H(X_1) \geq 2\log(2e(G)) - \log v(G)$$

using the same calculation as before, for a 2-edge path.

Next, we now have

$$H(X_1, Y_1, Y_2, X_2) = H(Y_1, Y_2) + H(X_1, X_2 \mid Y_1, Y_2)$$

by the chain rule. But we see that X_1 and X_2 are conditionally i.i.d. copies given Y_1 and Y_2, so we can split the term — they contribute equally, so we can rewrite

$$H(X_1, X_2 \mid Y_1, Y_2) = 2H(X_1 \mid Y_1, Y_2).$$

This is great, because now there are no more cycles left — we have gotten rid of the cycle structure in the graph.

Like before, we can now use the chain rule again to split this as

$$2H(X_1, Y_1, Y_2) - H(Y_1, Y_2).$$

Previously XY was a random edge. Here this is some distribution we don't exactly know, but we have the inequality from before which ges in the correct direction — so the first term is at least $4\log(2e(G)) - 2\log v(G)$. For the second term, the uniform bound gives us $2\log v(G)$ — it's 2 vertices, so the entropy is at most $\log v(G)^2$. And that's precisely the bound we wanted — $4\log(2e(G)) - 4\log v(G)$.

You can check the final coefficients agree, but more subtly, this is not a lossy calculation — we haven't thrown away any terms that should not have been thrown away.

Theorem 26.18 (Conlon–Fox–Sudakov 2010)

Sidorenko's conjecture holds for a bipartite graph that has a vertex adjacent to all vertices in the other part.

In other words, we have one vertex which is complete to everything else; we have no restrictions on the other adjacencies.

This was proven by a different probabilistic method known as dependent random choice.

In particular, we will demonstrate this for a graph consisting of two cycles joined together at an edge — with cycles $x_0y_1x_2y_2$ and $x_0y_2x_2y_3$.

Proof. We'll adopt the same strategy, but we again need a way to select a random homomorphism from this graph to G.

We start by choosing x_0y_1 to be a uniform random edge. We then choose y_2 to be a uniform random neighbor of x_0. We then choose y_3 to also be a uniform random neighbor of x_0.

Now we need to figure out how to choose x_1. We can think of x_1 as — we chose y_3 already, but imagine we forget about y_3. Then it's the same situation as before — so we can choose x_1 as a conditionally independent copy of x_0 conditioned on (y_1, y_2). Likewise, we can choose x_2 as a conditionally independent copy of x_0 conditioned on (y_2, y_3).

First, this is a well-defined random homomorphism from our graph to G.

Now let's compute the entropy of this random homomorphism. Simiarly to the calculation we just did, first by the chain rule we can write this as

$$H(X_0, X_1, X_2 \mid Y_1, Y_2, Y_3) + H(Y_1, Y_2, Y_3).$$

Now we can use conditional independence — conditioned on Y_1, Y_2, and Y_3, X_0, X_1, and X_2 are conditionally independent (although not identically distributed) by construction. So we can split them — as

$$H(X_0 \mid Y_1, Y_2, Y_3) + H(X_1 \mid Y_1, Y_2, Y_3) + H(X_2 \mid Y_1, Y_2, Y_3) + H(Y_1, Y_2, Y_3).$$

We can then drop some unnecessary conditionings. X_1 is generated based on first picking Y_1Y_2 and then finding a conditional copy of X_0. If we drop Y_3, it makes no difference. (This is a statement about conditional independence. For how to justify it, conditioned on X_0, Y_3 is independent of Y_1, Y_2, and X_1. Alternatively, the way we constructed X_1 only depends on Y_1 and Y_2 — but it is important that Y_1, Y_2, and Y_3 have some conditional independence. Because if knowing Y_3 could affect Y_1 and Y_2, we'd be in trouble. But you can imagine construction X_0, Y_1, Y_2, and X_1 first, and Y_3 later; then this becomes clearer.)

Then we can use the chain rule, which gives

$$H(X_0, Y_1, Y_2, Y_3) + H(X_1, Y_1, Y_2) + H(X_2, Y_2, Y_3) - H(Y_1, Y_2) - H(Y_2, Y_3).$$

(One term of the first $H(X_0 \mid Y_1, Y_2, Y_3)$ cancels out with the last term.)

The proof we did earlier with trees actually lower-bounds the first three terms — the first term is at least $3\log(2e(G)) - 2\log(v(G))$, by the first proof applied to a star. The second term has a similar lower bound of $2\log(2e(G)) - \log(v(G))$, and the same is true for the third (here we are implicitly using what we did in the first proof, which bounds the first three terms). For the remaining terms, we can use teh uniform bound, which gives us a bound of $2\log v(G)$ and $2\log v(G)$. Putting everything together, we get

$$7\log(2e(G)) - 8\log v(G).$$

This is indeed the same as what we were looking for, which finishes the proof for this graph. □

This same proof works as long as we have one vertex adjacent to everything else — we can start this process, and the same proof works (since nothing is lossy).

What goes wrong without this condition is that we can't copy, so we might not have a good way to construct this graph homomorphism. As a concrete graph we don't know how to prove Sidorenko for, it's *not* known whether Sidorenko holds for the Möbius graph — graph-theoretically the Mobius graph is $K_{5,5}$ minus a 10-cycle, where each of the five vertices on the left is adjacent to its parallel neighbor, its up-neighbor, and its down-neighbor. (This is called the Möbius graph because it's the incidence graph of the simplicial complex of a Mobius strip — we can construct a Mobius strip by gluing together triangles ABCDE and naming the five vertices we glue together; then A is incident to the vertices 1, 2, 3; B to 2, 3, 4; and so on. We do not know whether Sidorenko's conjecture holds for this graph.

If we try to run this strategy from earlier, you can start off fine — you can embed the first vertex and its three uniform neighbors. But then you get stuck when trying to embed the next — because we don't have a template to work from.

Student Question. Are there any simpler examples?

Basically this is the smallest graph where we do not know how to prove Sidorenko's inequality — for every 'smaller' graph, through some methods we do know how to prove it.

Student Question. Do we know how good Sidorenko's inequality is when G gets large?

It really is about G being large — this statement is an exact statement for all F and G, but the point is that if you think about F as small and G as being large and being $G(n,p)$, then the left-hand and right-hand side are basically the same.

Student Question. You run into the same template error if you try to do this for C_6. Is there a method to do it?

There's a question on the homework to prove it not just for C_6 but for three of them (in a honeycomb).

§27 December 12, 2022

§27.1 Shearer's Lemma

Shearer's lemma is a simple yet powerful tool that allows us to apply the entropy method to a variety of situations. First we'll see a special case of Shearer's lemma:

Lemma 27.1
For any random variables X, Y, and Z,

$$2H(X,Y,Z) \le H(X,Y) + H(X,Z) + H(Y,Z).$$

By sub-additivity we know $2H(X,Y,Z) \le 2H(X) + 2H(Y) + 2H(Z)$; but this is a stronger statement.

Proof. By the chain rule,

$$\begin{aligned} H(X,Y) &= H(X) + H(Y \mid X) \\ H(X,Z) &= H(X) \qquad\qquad\quad + H(Z \mid X) \\ H(Y,Z) &= \qquad\quad H(Y) \qquad\ + H(Z \mid Y). \end{aligned}$$

Now we can use condition dropping — we have $H(X) + H(Z \mid X) \ge H(X) + H(Z \mid X, Y)$, and likewise $H(Y,Z) \ge H(Y \mid X) + H(Z \mid X, Y)$. If we sum these terms up, then on the left-hand side we get

$$H(X,Y) + H(X,Z) + H(Y,Z),$$

and on the right-hand side we get

$$\ge 2H(X) + 2H(Y \mid X) + 2H(Z \mid X, Y) = 2H(X,Y,Z)$$

by the chain rule again. □

The proof in the general case (which we will see soon) is the same.

§27.2 A Geometric Application

Question 27.2. Suppose we have a compact set $K \subseteq \mathbb{R}^3$ in three dimensions such that its shadows K_{xy}, K_{xz}, and K_{yz} in the three directions (i.e., the projections of K onto the coordinate planes). Suppose we know that the areas of these projections are all at most 1. How big can K be?

One example is if K itself is a unit cube — a unit cube has projection area 1 onto each of the three coordinate planes. And the question is whether some other example could have volume bigger tahn 1.

This si not obvious — there are funny shapes with projections that don't reflect the 3D picture — but the answer is yes.

Theorem 27.3 (Loomis–Whitney Inequality)
$(\operatorname{Vol} K)^2 \le \operatorname{area} K_{xy} \operatorname{area} K_{xz} \operatorname{area} K_{yz}$.

(Note that the dimensionality is correct — both sides have six dimensions.)

Proof. First, we will give a *finite* version of this result:

Theorem 27.4

If $S \subseteq \mathbb{R}^3$ is finite, then $|S|^2 \leq |\pi_{xy}S|\,|\pi_{xz}S|\,|\pi_{yz}S|$.

Proof. Let (X, Y, Z) be a point in S chosen uniformly at random. Then $\log_2 |S| = H(X, Y, Z)$. Now applying Shearer's lemma, we have

$$2\log_2 S = 2H(X, Y, Z) \leq H(X, Y) + H(X, Z) + H(Y, Z).$$

Now $H(X, Y)$ is some random variable supported on $\pi_{xy}S$, so by the uniform bound we have an upper bound of $H(X, Y) \leq \log_2 |\pi_{xy}S|$, and likewise for the subsequent terms. This gives

$$2\log_2 |S| \leq \log |S_{xy}| + \log |S_{xz}| + \log |S_{yz}|.$$

□

Now to deduce the compact set, we can discretize K to be a union fo axis-aligned grid-aligned unit cubes, and take a limit (essentially approximating K by boxes). □

§27.3 General Form of Shearer's Lemma

Theorem 27.5 (Shearer's Entropy Inequality)

Suppose we have subsets $A_1, \ldots, A_s$ of $[n]$, such that every $i \in [n]$ appears in at least k sets A_j. Then

$$kH(X_1, \ldots, X_n) \leq \sum_{j\in[s]} H(X_{A_j}).$$

Notation 27.6. $H(X_{A_j})$ means the tuple $(X_i)_{i\in A_j}$.

In our special case $n = 3$ and we should think of X, Y, X, Z, and Y, Z as corresponding to the subsets. For example X appears twice, Y appears twice, and Z appears twice, so $k = 2$.

Example 27.7

$3H(W, X, Y, Z) \leq H(X, Y, Z) + H(W, Y, Z) + H(W, X, Z) + H(W, X, Y)$, since each appears three times on the right-hand side.

We also have $2H(W, X, Y, Z) \leq H(W, X) + H(X, Y) + H(Y, Z) + H(Z, W)$, as each appears twice.

Proof. The proof is the same as the proof we saw in the special case, but with more variables. □

In our geometric example, we can do the same with more variables.

Theorem 27.8 (Loomis–Whitney Inequality)

If $S \subset \mathbb{R}^n$ is finite, then

$$|S|^{n-1} \leq \prod_{i=1}^{n} |\text{projection of } S \text{ onto } i\text{th coordinate hyperplane } x_i = 0|.$$

There is a different way to write this (slightly more general) that will have combinatorial applications:

Theorem 27.9

Suppose $A_1, \ldots, A_s \subseteq \Omega$ such that each $i \in \Omega$ appears in at least k sets A_j. Then for every $\mathcal{F}$ which is a family of subsets of Ω, one has

$$|\mathcal{F}|^k \le \prod_{j\in[s]} \left|\mathcal{F}|_{A_j}\right|.$$

Notation 27.10. We use $\mathcal{F}|_A$ to be the 'shadow of $\mathcal{F}$ in A' — we take any $F \in \mathcal{F}$ and we only look at $F \cap A$. So $\mathcal{F}|_A = \{F \cap A \mid F \in \mathcal{F}\}$.

Essentially we're initially in a large universe Ω but we restrict ourselves to A, and for any set F that comes up we ignore the rest and only look at A. (And we delete duplicates.)

This is a corollary of the Shearer inequality above — it's the same inequality but allowing mroe flexible indices.

§27.4 Triangle-Intersecting Families

Definition 27.11. A set $\mathcal{G}$ of graphs on $V = [n]$ (a fixed n-vertex labelled set of vertices) is **triangle-intersecting** if whenever $H, H' \in G$, $H \cap H'$ contains a triangle (this is the edge intersection of the two graphs).

Question 27.12. How large can such a set be — what is the size of the largest triangle-intersecting family?

One example of a triangle-intersecting family is if we fix a triangle, say on 123, and then take all graphs containing it. This set will certainly be triangle-intersecting in a trivial way (every pair of graphs intersects on this triangle).

The size of this set is

$$2^{\binom{n}{2}-3}$$

— so essentially 1/8 of all graphs will be in this set.

Question 27.13. Can we do better?

One reason we present this problem and see a partial solution is that this is the problem where Shearer's inequality originally arose.

First, here is an easy bound: a triangle-intersecting family is necessarily edge-intersecting (in the sense that every pair of graphs intersects in at least an edge). So we can forget about the graph stucture and think about the edge set. We then want to find the largest intersecting familly. This is at most half — we saw this in the second-lecture (when discussing Erdos–Ko–Rado); you can alwyas pair up a set and its complement, and you can get half by including everything with a certain element.

So this gives us a bound of $1/2 \cdot 2^{\binom{n}{2}}$. We will see an improvement to $1/4$.

Theorem 27.14 (Chung–Graham–Frankel–Shearer)

Every triangle-intersecting family has size less than $2^{\binom{n}{2}-2}$.

This is a factor of 2 improvement, but also a factor of 2 away from our construction.

Proof. Let $\mathcal{G}$ be our triangle-intersecting family. The main idea is to observe that when restricted to $\overline{K_{\lfloor n/2\rfloor,\lceil n/2\rceil}}$, originally the graph was triangle-intersecting, but now that triangle must show up as at least an edge — one of the two parts must have at least one edge inside it in the triangle.

So then this thing is edge-intersecting. And knowing that it's edge-intersecting gives us a 1/2 density bound.

At this point we knwo restricting to some specific partition, we get density 1/2. Before today's elcture, the only thing we knew what to do was to do a sampling bound, and get back 1/2. But instead using Shearer over all possible partitions, we can go from 1/2 to 1/4. That comes from the fact this set occupies 1/2 of all possible edges.

For every $S \subseteq [n]$ with $|S| = \lfloor n/2\rfloor$, let $A_S = \binom{S}{2} \cup \binom{\overline{S}}{2}$ (so the edge set of the bipartite graph's complement), and let r be theh number of edges in A_S (it's $\binom{\lfloor n/2\rfloor}{2} + \cdots \le \frac{1}{2}\binom{n}{2}$).

Now $\mathcal{G}|_{A_S}$ is edge-intersecting — every triangle has to have an edge in A_S. And thus $|\mathcal{G}|_{A_S}| \le 2^{|A_S|-1} = 2^{r-1}$.

Now we apply Shearer — every edge of K_n appears in exactly

$$k = \frac{r}{\binom{n}{2}}\binom{n}{\lfloor n/2\rfloor}$$

different A_S with $|S| = \lfloor n/2\rfloor$, by averaging. Now applying Shearer's inequality in this form, we see that

$$|\mathcal{G}|^k \le (2^{r-1})^{\binom{n}{\lfloor n/2\rfloor}}.$$

If you set back the value of k into this whole expression, we find that

$$|\mathcal{G}| \le 2^{\binom{n}{2} - \frac{\binom{n}{2}}{r}},$$

and $\binom{n}{2}/r \ge 2$. (Where the 2 comes from is that we get 1 from the easy bound and another from the paplication of Shearer, because all the restricted edge sets have density 1/2.) □

Remark 27.15. This gets us halfway between the easy answer and conjectured answer. It turns out the conjecture is true (1/8); it was proven by Ellis–Filmus–Friedgut 2012 that K_3-intersecting families have density 1/8 (The best example is fixing a triangle).

Prof. Zhao and Berger extended this result to K_4-intersecting. But in general the question for K_r is open. We expect the answer is what we get when we fix a K_r and take all graphs containing it, but this is open for larger values of r.

These proofs use a different method — Fourier analysis and linear programming bounds.

§27.5 Independent Sets in Regular Graphs

Question 27.16. Fix d. Among all d-regular graphs, which graph maximizes the number of independent sets (denoted as $i(G)$)?

(Note this is different from maximizing the size of the largest.)

We normalize by $i(G)^{1/n(G)}$, because having more vertices of course gives us more independent sets.

Conjecture 27.17 — If you take a $K_{d,d}$ (complete bipartite graph with d vertices on each side), this has some number of independent sets. If we now take many copies of this thing — $K_{d,d} \cup K_{d,d} \cup \cdots$.

Basically we're keepng these guys small but having a fair number of independent sets still, and then we have a bunch of copies and the numbers multiply. In this case we have

$$i(nK_{d,d}) = i(K_{d,d})^n = (2 \cdot 2^d - 1)^n$$

(we pick each side and take an arbitrary subset, and then remove 1 because we overcounted the empty subset).

It turns out this is the best you can do.

Theorem 27.18 (Kahn–Zhao)
If G is a n-vertex d-regular graph, then

$$i(G) \le i(K_{d,d})^{n/2d}.$$

In other words, the answer to this question is $K_{d,d}$.

Remark 27.19. The history of this result is that Kahn proved this in the case of bipartite graphs, which we will see soon, using entropy. When Prof. Zhao was an undergrad he found a way to extend the bipartite case to the general case. This is the paper where Prof. Zhao first got exposed to the entropy method. The proof appeared in a slightly different form in the paper, but it's not l ong but took Prof. Zhao 7 years to understand what the idea was, until roughly whne he taught this course.

Proof for Bipartite G. We will illustrate the proof for a 6-cycle; the proof method works in general. Supose our vertices are x_1, x_2, x_3 on the left and y_1, y_2¡ y_3 on the right (with edges $y_1x_1y_2x_3y_3x_2$).

To prove an upper bound on the number of independent sets, we first let $(X_1, X_2, X_3, Y_1, Y_2, Y_3) \in \{0,1\}^6$ be the indicator vector of an independent set chosen uniformly at random (we look at the collection of all independent sets, and pick one uniformly). In apрrticular $H(X_1, \ldots) = \log_2 i(G)$.

Note that this is different from what we did in the last lecture, where we wanted to lower bound the number of homomorphisms (in Sidorenko) and we came up with a special distribution. Here we start witih a uniform distribution.

Let's consider this random vector and calculate its entropy. We know

$$\log i(G) = H(X_1, X_2, X_3, Y_1, Y_2, Y_3).$$

By the Chain rule, we can rewrite this as

$$H(X_1, X_2, X_3) + H(Y_1, Y_2, Y_3 \mid X_1, X_2, X_3).$$

Now we apply Shearer. To make the notation more convenient, add a factor of 2. Using Shearer on the first term, we get

$$\le H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3) + 2H(Y_1, Y_2, Y_3 \mid X_1, X_2, X_3).$$

For the terms on the right $H(Y_1, Y_2, Y_3 \mid X_1, X_2, X_3)$, we are treating every independent set with equal weight. So if I tell you the choices for X_1, X_2, X_3, what's left to decide is — if X_1 or X_2 is included that rules out y_1, otherwise it has two possibilitis. But the choidds for Y_1, Y_2, Y_3 rae independent after you've decided X_1, X_2, X_3. (Think about counting — you can decide each choice separately and multiply the number of choices together.)

What that amounts to here is a claim about conditional independence —

$$H(Y_1, Y_2, Y_3 \mid X_1, X_2, X_3) H(Y_1 \mid X_1, X_2, X_3) + H(Y_2 \mid X_1, X_2, X_3) + H(Y_3 \mid X_1, X_,2, X_3).$$

We can actually say more — once we have deciede X_1, X_2, X_3 and we are trying to figure out what's going on with Y_1, we do not care about X_3. So we can delete some of the conditioned random variables. Now we have

$$H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3) + 2H(Y_1 \mid X_1, X_2) + 2H(Y_2 \mid X_1, X_3) + 2H(Y_3 \mid X_2, X_1).$$

Now we have three pairs of sums. It remains to prove that $H(X_1, X_2) + 2H(Y_1 \mid X_1, X_2) \leq \log i(K_{2,2})$ (which is the thing we're trying to compare with on the right-hand side). Then the same inequality holds for the other pairs, and we can plug in everything and that gives us at most $3 \log i(K_{2,2})$, which finishes the inequalities. (We haven't lost everything, and you can check the numbers work out correctly.)

We can interpret the $Y_1 \mid X_1, X_2$ by coming up with Y_1' which is a conditionally independent copy of Y_1 conditioned on (X_1, X_2) (similarly to in the last lecture — we have some distribution and we can condition on X_1 and X_2 and make a conditionally independent copy). This allows us to write the sum as

$$H(X_1, X_2) + H(Y_1, Y_1' \mid X_1, X_2).$$

By the reverse chain rule, this becomes

$$H(X_1, X_2, Y_1, Y_1').$$

Now we have X_1 and X_2 corresponding to whether these two vertices are in the independent set. We also have Y_1, and a conditionally independent copy Y_1'.

This si now some distribution on independent sets of $K_{2,2}$. It's not necessarily unifrom — it comes out of some strange distribution we're not going to bother to understand (look at that distribution, figure out the distribution on (X_1, X_2, Y_1), and make a conditionally independenet copy). This is some crazy distribution, but you now they generate an i ndependent set of $K_{2,2}$. So we can apply the uniform bound to get that this is *at most* $\log i(K_{2,2})$. □

The whole proof works for any d-regular bipartite graph.

Question 27.20. What happens if we start with a non-bipartite graph?

If the graph is non-bipartite, we would like to do the chainning step but we wouldn't know what to do here — picking an independent set in a non-bipartite grpahi s no longer a twopstep process where you first decide the left and then the right.

Next we will see the second part, which is the reduction from a general graph to a bipartite graph.

Lemma 27.21

For any graph G (niot even necessarily regular), we have

$$i(G)^2 \leq i(G \times K_2).$$

Definition 27.22. $G \times K_2$ is the graph where we start with a graph G (which we use as a Czech flag with two triangles instead of one, one on each end).

Then we draw the two copies of G, with one above the other (imagine a 3D picture — we draw it slightly above in a different color). This gives us $2G$.

We then look at all the pairs of parallel edges, and we replace them with crossing edges — so instead of a_1b_1 and a_2b_2, we now have a_1b_2 and a_2b_1. (Edges now go yellow-white instead of yellow-yellow and white-white.)

Imagine two sheets of paper where initially they all lie on the same sheet of paper, and then we cross all edges so that they now all go up and down.

Proof. We will construct an injection from $I(2G)$ (the set of independent sets of $2G$ to $I(G \times K_2)$.

Suppose we start with some $S \in I(2G)$. Basically, we are going to try to use the same set. There's a one-to-one correspondence in the vertex set, so we can color in the same vertices.

The problem is that this may no longer be an independent set — now we have some bad edges. Let $E_{\text{bad}}(S)$ be the set of edges in G that end up creating conflicts in $G \times K_2$.

In our example it's the top segment of the left triangle, and the top and right segments of the rectangle.

The point now is to try to fix these conflicts. We will try to do this in a consistent way.

One way to fix the conflict is to swap some of the vertices to their twin, since they come in pairs. So we are going to do this. Say we swap the two things in the two corners —- that would fix the situation.

There's a couple of things we need to consider — can we always swap to fix, and can we swap consisetntly so that we get a well-defined map that's actually an injection?

First, $E_{\text{bad}}(S)$ is always bipartite — because it's a bipartite graph consisting of the white vertices on one side, and the yellow vertices on the other (of $2G$), since all edges go white to yellow. This means we can always find some bipartition to swap the edges and fix teh bad sets.

So now teh question is, can we do this consistently (maybe some other S will also have problems and we fix them to end up with teh same resulting set).

The way to deal with this is: for each bipartite subgraph $F \subset G$, pick a canonical (e.g. lexicographically first) bipartition A, B. When we have that, then we swap $A(E_{\text{bad}}(S))$. — SO we decide in advance, whenever I see a bipartite subgraph, which vertices to swap (and that's a function of the bipartite subgraph, but not S).

Why does this work? We need to show that this procedure creates an injective map. Roughly the point is that we can reverse it. Suppose I see some independent set in $G \times K_2$, and I know that it's an image of ϕ (our injection) and we want to be able to explicitly see where it comes from.

So we look at our picture after we've done the swapping.

First, you can still figure out where the bad edges are — because you can look at the old graph and see if there was a bad edge. And then you can figure out which vertices were swapped, And ten you can reverse those swaps.

So given an element in the image of ϕ, we can first determine the bad edges. This determines the swapped vertices, and this lalows you to recover the pre-image. And if you an always recover the pre-image, then it's an injective map. □

This then gives the result $i(G) \le i(K_{d,d})^{n/2d}$ for general graphs:

Proof. $i(G) \le i(G \times K_2)^{1/2}$. This graph is bipartite, so applying the bipartite version of the problem gives $i(K_{d,d})^{n/2d}$, which finishes the inequality. □

This entropy proof tells us more. It tells us that

Theorem 27.23 (Galvin–Tetali 04)

For all graphs G which are n-vertices and d-regular and bipartite, adn H is any graph allowing loops, we have

$$\hom(G, H) \le \hom(K_{d,d}, H)^{n/2d}.$$

You may wonder what's going on here. There are some special cases.

Graph homomorphisms are a geeral quantity that captures a lot of stuff. THe number of homomorphisms from G to the graph with two vertices, one edge and one loop, is precisely the number of independent sets of G — the reason is that the vertices of G mapped to the right vertex must form an independent set. (If they're adjacent, they cannot both map to the right vertex.)

Moreover, the number of homomorphisms from G to K_q is the number of proper q-colorings of G with q labelled colors. (For example when $q = 3$, we can think of the three vertices as red, blue, yellow. Two vertices which are blue cannot both be adjacent, because there is no loop.)

Remark 27.24. For independent sets we saw that the bipartite hypothesis can be relaxed. A natural question is, can it be relaxed in general (here)? The answer is no. For example, when H consists of two loops, then a homomorphism from G to H – every vertex of G is assigned the left loop or the right loop, and there shouldn't be any crossing edges. So this is $2^{\#CC \text{ of } G}$.

If we want to maximize this homomorphism count raised to the $1/v(G)$, which is basically this problem (among d-regular graphs), we are trying to minimize the component sizes. So the minimizing G is K_{d+1} and not $K_{d,d}$. So teh bipartite condition cannot be dropped ing eneral.

The next question is whether it can be dropped for some interesting cases. Cna it be dropped for colorings? It turns out it can:

Theorem 27.25 (Sah–Sawhney–Stoner–Zhao 2020)
The bipartite condition can be dropped for H being K_q (corresponding to counting proper $q$coloings).

(The proof is much more involved.)

§28 December 14, 2022 — The Container Method

The container method is a relatively new development. It was developed in the last decade or so, and it's one of the most important recent developments.

Question 28.1. How many triangle-free graphs are there on n (labelled) vertices?

For example, $K_{n/2,n/2}$ is triangle-free; so all its subgraphs are triangle-free as well. It has $2^{n^2/4}$ subgraphs, so the number of triangle-free graphs is *at least* $2^{n^2/4}$.

It turns out this is the right answer:

Theorem 28.2 (Erdős–Kleitman–Rothschild 1973)
The number of triangle-free n-vertex graphs is

$$2^{n^2/4+o(n^2)}.$$

So in the exponent, we have the right first-order term. The original proof used some method; another method is by Szemeredi's graph regularity lemma (if we want to learn more about this we can take 18.225). Today we will see a proof using the container method, which is very powerful and allows us to do much more.

Remark 28.3. Erdős, Kleitman, and Rothschild proved more — they showed that a $1-o(1)$ fraction of all triangle-free graphs on n vertices are bipartite. This gives a characterization of the typical structure of a typical triangle-free graph. We won't discuss it here, though it can be derived using the techniques we've seen along with lots of other things.

(It won't really matter from now whether we talk about labelled or unlabelled graphs — this affects the count by $n! \approx 2^{n \log n}$, which is tiny compared to $2^{n^2/4}$.)

§28.1 The Container Method

The container method has to do with independent sets in hypergraphs.

Lots of problems in combinatorics can be phrased in terms of independent sets in graphs and in hypergraphs.

For this problem, we can form a 3-uniform hypergraph H whose vertices are all unordered pairs of elements $\{1, \dots, n\}$ — so $V(H) = \binom{[n]}{2}$ — and whose edges are triples of the form $\{xy, xz, yz\}$, where x, y, and z are distinct elements of $[n]$.

In other words, we start with a complete graph and look at its edge set. The edges now become vertices in the 3-graph, and the edges of the hypergraph are triples of edges which form a triangle.

Definition 28.4. An independent set in a hypergraph is a subset of vertices that contain no edges.

(This is consistent with the notion of independent sets in a graph.)

So in other words, if we have a 3-edge, then we can't have all three vertices in our independent set.

Then a triangle-free graph on $[n]$ is precisely the same as an independent set in H — an independent set corresponds to a subset of edges of a clique containing no triangles, which is the same as a triangle-free graph.

There are lots of naturally occurring combinatorial problems that can be phrased in terms of independent sets in hypergraphs.

The container theorem is a general result about independent sets in hypergraphs. It was first developed in a graph form. Kleitman–Winston developed this method and used it to count the number of C_4-free graphs (which is $2^{O(n^{3/2})}$); it was later extended by Sapozhenko 2001, who used it to prove other results (such as the number of independent sets in a d-regular graph). The modern form was found in the last decade independently by Balogh–Morris–Samoty 2015 and Saxton–Thomason 2015; it is one of the most important developments in extremal and probabilistic combinatorics in the last decade.

The tool that we will use is a container theorem for triangle-free graphs (which is a corollary of the general theorem):

Theorem 28.5 (Containers for Triangle-Free Graphs)

For every $\varepsilon > 0$, there exists $C > 0$ such that for every n, there exists a collection $\mathcal{C}$ of graphs on n vertices with the following properties:

- $|\mathcal{C}| \le n^{Cn^{3/2}}$;
- Every $G \in \mathcal{C}$ has $(\frac{1}{4} + \varepsilon)n^2$ edges;
- Every triangle-free graph on n vertices is a subgraph of some $G \in \mathcal{C}$.

What this is saying is that there's a collection of *containers* for triangle-free graphs. The first statement says that there's not too many containers — we are relatively efficient in terms of the number of containers that are used. The second is that the containers are small — each container is small (it doesn't have too many edges). And they're called containers because every triangle-free graph can be contained. (A triangle-free graph can have as many as $n^2/4$ edges; so we should have containers that are at least as large, but probably a bit larger to have some extra room.)

The set of all triangle-free graphs would satisfy the second and third conditions, but we'd have too many. The point is that if we allow a bit of wiggle room to make the containers a bit bigger, then we can contain all the triangle-free graphs, but much more efficiently — in the sense of not requiring so many containers.

> **Remark 28.6.** In particular, the containers are not necessarily triangle-free. But it turns out that each container contains a small number of triangles.

We are going to take this statement for granted; it follows from a general result.

Now we'll see how to prove the Erdős–Kleitman–Rothschild result using this container theorem:

Proof. Let $\varepsilon > 0$ be an arbitrarily small constant, and let C be as above. Then every $G \in \mathcal{C}$ has at most $(\frac{1}{4} + \varepsilon)n^2$ edges, and every triangle-free graph is contained in some $G \in \mathcal{C}$. This tells us that the number of triangle-free graphs on n vertices is at most

$$\sum_{G \in \mathcal{C}} 2^{e(G)}.$$

(We may overcount, but it's certainly an inequality in the correct direction.) We know that there are not too many containers — the number of containers is $n^{Cn^{3/2}}$ — and on the other hand, we also know each container doesn't have too many edges. So then we have

$$\sum_{G \in \mathcal{C}} 2^{e(G)} \le n^{Cn^{3/2}} \cdot 2^{(1/4+\varepsilon)n^2} = 2^{(1/4+\varepsilon)n^2 + Cn^{3/2}\log n}.$$

By choosing ε small enough, this is then the same as our goal. □

The point of containers is that we were very efficient in the union bound. If we summed over all maximal triangle-free graphs, then there'd be too many terms. But here we only have a few containers, so we can use the sum to get the correct answer.

> **Student Question.** *How many maximal triangle-free graphs are there?*
>
> **Answer.** The number of maximal triangle-free graphs is $2^{(1/8+o(1))n^2}$. This is not obvious.

> **Student Question.** *How was EKR proved originally?*
>
> **Answer.** Using some combinatorial arguments that are hard to encapsulate with a name (they proved it not just for triangle-free graphs but also for cliques).

There are still open conjectures — what if we wanted to avoid a different graph?

> **Definition 28.7.** The **extremal number** $\operatorname{ex}(n, H)$ is the maximal number of edges in a n-vertex H-free graph.

> **Theorem 28.8** (Erdős–Stone–Simonovits)
>
> For every fixed graph H,
>
> $$\operatorname{ex}(n, H) \asymp \left(1 - \frac{1}{\chi(H) - 1} + o(1)\right)\binom{n}{2}.$$

(This comes from a complete multipartite graph.)

The hard situation is when H is bipartite, when the above only gives us $o(n^2)$; this is not a very precise asymptotic, because we'd want the actual growth rate. And for most graphs H, we don't know what the answer is.

Theorem 28.9

For every fixed H,

$$\#H\text{-free graphs on } n \text{ vertices} = 2^{\text{ex}(n,H)+o(n^2)}.$$

But for H bipartite, this is not satisfying because we don't actually get precise asymptotics in the exponent.

Conjecture 28.10 — For all bipartite H with a cycle,

$$\#H\text{-free graphs on } n \text{ vertices} = 2^{O(\text{ex}(n,H))}.$$

(We do not know this. The C_4-case is a special case — the extremal number for C_4 is $O(n^{3/2})$, and Kleiteman–Winston showed that the number of C_4-free graphs is $2^{O(n^{3/2})}$ as well.)

Remark 28.11. People generally believe this conjecture, since there is some evidence — we know it's true for many classes of H, and for H satsifying certain assumptions.

Student Question. *Do we believe that teh constant should be 1?*

Answer. There's a lower bound of 1; but there are examples of H whwere the constant is lower-bounded greater than 1. (In the C_4 case this isn't known.) So the answer is no in general, but in some specific cases we don't know.

§28.2 Mantel's Theorem in Random Graphs

Theorem 28.12 (Mantel's Theorem)

The number of edges in a triangle-free graph is at most $n^2/4$ — i.e.,

$$\text{ex}(n, K_3) = \left\lfloor \frac{n^2}{4} \right\rfloor.$$

(This comes from a bipartite graph with equally split parts.)

We will see the following statement:

Theorem 28.13

If $p \gg 1/\sqrt{n}$, then with high probability, every triangle-free subgraph of $G(n,p)$ has at most $(\frac{1}{4}+o(1))pn^2$ edges.

You can get this many edges by taking a bipartition of the vertices and looking at all edges in between. This theorem tells us that Mantel is still true if instead of looking at all possible edges, we restrict to the edges inside a $G(n,p)$; and the answer is that it's basically the same thing.

Remark 28.14. The hypothesis is necessary — the statement is false if $p \ll 1/\sqrt{n}$. This is because the number of triangles is typically $O(n^3p^3)$, whereas the number of edges is around $n^2p/2$. So if p is small, then there are way more edges than triangles. Then the statement is false because we can get a K_3-free subgraph of $G(n,p)$ by removing all triangles; we don't change the number of edges by much when we do this (since there's so few triangles, this is a negligible number of edges).

Remark 28.15. There is a much stronger form of this statement, due to DeMarco and Kahn — the maximum triangle-free subgraph of $G(n,p)$ is actually bipartite with high probability. So it's not just that the number is this; the best thing you can do is actually to take a bipartite subgraph. (This si a much stronger statement we won't get into.)

Let's see how the container theorem for triangle-free graphs can help us prove this result.

Proof. We'll prove this for $p \gg \log n/\sqrt{n}$ (this has to do with a technicality to how we stated the theorem; we can get all the way by using a more refined version of the result).

Let $\varepsilon > 0$ be arbitrarily small, and C as in the container theorem. For every $G \in \mathcal{C}$, we know that $e(G) \leq (\frac{1}{4} + \varepsilon)n^2$. By the Chernoff bound, if we look at a fixed G, then if we take a p-random subset of edges, the number of edges remaining concentrates very well — in particular

$$\mathbb{P}(G \cap G(n,p) \geq (\frac{1}{4} + 2\varepsilon)n^2p) \leq e^{-\Omega_\varepsilon(n^2p)}.$$

Then we can take a union bound — to find the probability that $G(n,p)$ contains a triangle-free subgraph with more than $(1/4+2\varepsilon)n^2p$ edges (which we want to show is unlikely), we can union bound over containers. If it has a triangle-free subgraph with lots of edges, it must be contained in some container, so we can zoom into that container, and inside that container we must have lots of edges — so this is at most

$$\sum_{G\in\mathcal{C}} \mathbb{P}(e(G \cap G(n,p)) \geq (\frac{1}{4} + 2\varepsilon)n^2p) \leq |\mathcal{C}| \cdot e^{-\Omega_\varepsilon(n^2p)} = e^{O_\varepsilon(n^{3/2}\log n) - \Omega_\varepsilon(n^2p)}.$$

If $p \gg \log n/\sqrt{n}$, then the second term is large and beats out the first term; in particular this goes to 0 as $n \to \infty$. This is true for every ε, so we derive the desired consequence. □

Again the point is that we're using the union bound. If we union bounded over all triangle-free grpahs we'd have too many terms; but the container theorem allows us to be very efficient and only use a small number of terms.

Question 28.16. Does this show us that 3/2 is optimal in teh container theorem?

Yes. (There's an extra log; there's a more technical theorem that allows you to get rid of the log, but otherwise the 3/2 is optimal.)

§28.3 Graph Containers

This result is basically due to Kleitman–Winston from the 1980s.

Theorem 28.17

For every $c > 0$, there exists $\delta > 0$ such that the following holds: let $G = (V, E)$ be a graph with average degree d and maximum degree at most cd. (We can think of this as an almost-regular graph.) Then there exists a collection $\mathcal{C}$ of subsets of V with the following properties:

- $|\mathcal{C}| \le \binom{|V|}{\le 2\delta|V|/d}$ (i.e., the number of subsets of $|V|$ with size at most $2\delta\,|V|\,/d$) — this should be small, because d will not be too small.
- Every independent set I of G is contained in some $C \in \mathcal{C}$.
- Every $C \in \mathcal{C}$ has $|C| \le (1-\delta)\,|V|$.

First, this does not prove our triangle theorem — because that is about independent sets in three-uniform hypergraphs. (There was an analoghy where triangle-free graphs correspond to independent sets in a specific 3-uniform hypergraph; triangles involve 3-edges.) Nevertheless, we have to start somewher; so we start with independent sets in graphs.

Secnod, the second bullet point in the triangles is supposed to correspond to the third here. It may not look the same because there we had really small containers, and here we only have containers a little bit smaller than the trivial upper bound of $|V|$. It turns out that in applications when you actually use containers, we iterate the theorem so that the containers becomes maller and smaller and smaller, until they get to the minimum size; then it becomes of that form. (So we start with one giant container, namely the universe — of course this contains everything, and doesn't do anything useful. Then in one round iof this application we make the containers a bit smaller; we get more containers but not too much more. Then inside each container we apply this again; we keep geting more containers but not too much more. We go on some constant number of times; then all containers become the saize we want, and we don't have too manycontainers; then we are happy.)

(This is one of the statements that may becoeme clearer when we see the proof; the message is we're trying to build containers for independent sets.)

Proof. We will use an algorithm (i.e., someone gives us a graph and we run an algorithm to output a container).

Fix the graph G.

Algorithm 28.18 (Graph Container Algorithm) — Input: an independent set $I \subset V$.

Output: a 'fingerprint' $S \subseteq I$ with $|S| \le \frac{2\delta|V|}{d}$. (This is a subset of I which is supposed to tell us some information; of course it may be lossy because we threw away other information). Together with the fingerprint, we also output a container $C \supseteq I$, such that C depends only on S.

So the output will be a way to generate a fingerprint from I, and a way to generate (even if we don't see all of I and only see the fingerprint — and the original graph) — a container that's guaranteed to contain the independent set I.

Throughout the algorithm, we'll maintain some data that will mutate throughout the algorithm, and a partition of $V = A \cup S \cup X$, where:

- A is supposed to be the 'available' vertices — initially $A = V$.
- S is supposed to be the 'current fingerprint' — we'll build the fingerprint one element at a time, and initially S is the empty set.
- X is the set of excluded vertices, also initially $\emptyset$.

When we say the 'max-degree order' of $G[A]$, we order the vertices of A in decreasing degree in $G[A]$. (We break ties according to some initially fixed arbitrary order.)

Our algorithm does the following: while $|X| < \delta |V|$, we:

(1) Let v be the first vertex of $I \cap A$ in the max-degree order. (A is the set of available vertices, so we just look at A; everythign else we have somehow gotten rid of. We sort in decreasing degree according to the induced subgraph; so we have some vertices, adn I might be some subset of these. We then pick the first one, and call that v.)

(2) Add v to S.

(3) Add the *neighbors* of v to X. (So we throw away all the neighbors of v. Note that v cannot neighbor another thing in the independent set, bedcause it's an independnet set.)

(4) Add the vertices *before* v in the max-degree order on $G[A]$ to X as well (if we had additional vertices before v, we also throw them out).

(5) Remove from A all the new vertices added to S or X. (This corresponds to A being the set of availabel vertices — we pick the independent set vertex with the maximum degree in the remaining thing, throw out its neighbors, and throw out all vertices that come before it. Once you throw them out, you're left with some A; that's going to be the remaining A.)

We now have a new set A, and we repeat until we have excluded a large number of vertices, when we terminate.

This is somewhat greedy, but it's a bit sophisticated. Essentially, we are trying to build a fingerprint — someone gives an indpeendnet set, and we're trying to identify the most important vertices in it; and the idea is to look or the highest degree vrtices, but in some very precise sense.

> **Claim** — When the loop terminates, we get a partition $V = A \cup S \cup X$ such that $|X| \geq \delta |V|$ and $|S| \leq 2\delta |V| / d$.

The first is true by definition (or else the loop would not hav eterminated).

Intuitively the second makes sense because we are picking a vertex with high degree, and then everythign else we throw away; the overall degree shouldn't have changed by all that much. So the graph should remain roughly regular, and each time we're throwing in a vertex of high degree. So we shouldn't have a fingerprint that's too large since each time we add a vertex to the fingerprint, we get rid of a lot of vertices.

Proof Sketch. By the degree hypothesis, in every iteration, at least $\frac{d}{2}$ bew vertices are added to X (provided that $d \leq 2\delta |V|$). So each time we're adding a lot of vertices, and we shouldn't be able to do this too many times. (There is some subtlety that we won't get into, but the intuition makes sense — we shouldn't have too many things in the fingerprint because each added vertex is quite costly.) □

Now here are some key facts about what we have left: in the goal, the container is supposed to depend just on the fingerprint. It could be that two independent sets generated the same fingerprint. But:

- If two different independent sets I and I' generate the same fingerprint, then they should necessarily produce the same partition $V = A \cup S \cup X$.

This is because if I and I' generated the same fingerprint, then some of the remaining vertices of I may not have been seen, but that doesn't matter — the algorithm never sees them, so the output is the same.

- The final set $S \cup A$ contains I.

This s true becus ethe only things we've thrown into the excluded vertices are necessarily non-I elements.

> **Student Question.** What happens if I is small and we run out of I after a few rounds in the algrithm (i.e. it becomes empty)?
>
> Because the degrees are bounded, we can first enlarge I into a maximal independent set, and then it will be quite large.

And so this is the point earlier — this $S \cup A$ will be our container C. It icontains I, and it is a function of the ifngerprint alone. Therefore, the total number of possibilities for containers $S \cup A$ is at most the number of possibilities for S —- which is at most $\binom{|V|}{\le 2\delta|V|/d}$ (since S is small).

And furthermore, $|A \cup S| \le |V| - |X| \le (1-\delta)|V|$, since we excluded a lot of vertices. This finishes the proof of the graph container theorem. □

So we can find a small number of containers that contain all the independent sets, and each one of them is not so large.

To finish, we'll state without proof the extension of the container theorem to 3-uniform hypergraphs, which can then be used to establish the result above about triangle-free graphs. The proof is more complicated, but it follows the same strategy of trying to run a container-producing algorithm. (The 3-uniform algorithm will call the 2-uniform algorithm —- i'ts much more involved.)

Theorem 28.19 (Container Theorem for 3-graphs)

For every $c > 0$, there exists $\delta > 0$ such that the following holds: let H be a 3-graph with average degree $d \ge 1/\delta$ and such that $\Delta_1(H) \le cd$ (the maximum number of edges containing one vertex) and $\Delta_2(H) \le c\sqrt{d}$ (the maximum number of edges containing a given pair of vertices). Then there exists a collection $\mathcal{C}$ of $V(H)$ such that:

- $|\mathcal{C}| \le \binom{V(H)}{\le V(h)/2\sqrt{d}}$
- Every independent set of H is contained in some container;
- All the containers have size at most $(1-\delta)v(H)$.

You should see this as a direct extension fo the graph container theorem — in a hypergraph with some control on maximum degrees, we can find a colleciton of containers containing all independent sets where there's not too many, and each container is at most 99% of the size of the universe. By iterating downwards, we can keep on doing this and eventually you can get all the way down to $(\frac{1}{4} + \varepsilon)n^2$ (in our example). So that heorem follows from iterating this result.

This was one of the big breakthroush in probaiblistic and extremal combinatorics in teh last decade, and it has lots of applications.

To wrap things up, n this course we went tohrough a lot of content — a few cute applications of the probabilistic method, ltos of basic methods (linearity of expectation, second moment method) and important tools (Lovasz Local method, Janson, concentration inequaliites, entropy). These techinques are broadly applicatble; the probaiilistic method is an important tool in combinatorics.

If you want to expore related topics, next term there is the class 18.218 on Ramsey theory, adn next fall there will be 18.225 o n graph theory and additive combinatorics (which will explore some topics mentioned todya such as the regularity lemma, as well as additive combinatorics).