

Lecture notes (MIT 18.226, Fall 2022)

Probabilistic Methods in Combinatorics

Yufei Zhao

Massachusetts Institute of Technology

yufeiz@mit.edu

<http://yufeizhao.com/pm/>

Preface

Last updated: December 14, 2022

These notes were created primarily for my own lecture preparation. The writing style is informal. These notes are not meant to be a replacement of the lectures or the textbook.

The main textbook reference for this class is

Alon and Spencer, *The Probabilistic Method*, Wiley, 4ed.

Please report errors via the Google Form <https://bit.ly/pmnoteserror>.

Asymptotic notation convention

Each line below has the same meaning for positive functions f and g (as some parameter, usually n , tends to infinity)

- $f \lesssim g$, $f = O(g)$, $g = \Omega(f)$, $f \leq Cg$ (for some constant $C > 0$)
- $f/g \rightarrow 0$, $f \ll g$, $f = o(g)$ (and sometimes $g = \omega(f)$)
- $f = \Theta(g)$, $f \asymp g$, $g \lesssim f \lesssim g$
- $f \sim g$, $f = (1 + o(1))g$
- *whp* (= *with high probability*) means with probability $1 - o(1)$

Warning: analytic number theorists use \ll differently to mean $O(\cdot)$ (Vinogradov notation)

Subscripts (e.g., $O_s(\cdot)$, \lesssim_s) are used to emphasize that the hidden constants may depend on the subscripted parameters. For example, $f(s, x) \lesssim_s g(s, x)$ means that for every s there is some constant C_s so that $f(s, x) \leq C_s g(s, x)$ for all x .

We write $[N] := \{1, \dots, N\}$.

Contents

1	Introduction	1
1.1	Lower bounds to Ramsey numbers	1
1.2	Set systems	7
1.3	2-colorable hypergraphs	10
1.4	List chromatic number of $K_{n,n}$	12
2	Linearity of expectations	15
2.1	Hamiltonian paths in tournaments	15
2.2	Sum-free subset	16
2.3	Turán's theorem and independent sets	17
2.4	Sampling	19
2.5	Unbalancing lights	21
2.6	Crossing number inequality	23
2.7	Dense packing of spheres in high dimensions	26
3	Alterations	29
3.1	Dominating set in graphs	29
3.2	Heilbronn triangle problem	30
3.3	Markov's inequality	31
3.4	High girth and high chromatic number	32
3.5	Random greedy coloring	33
4	Second moment method	37
4.1	Does a typical random graph contain a triangle?	37
4.2	Thresholds for fixed subgraphs	42
4.3	Thresholds	46
4.4	Clique number of a random graph	55
4.5	Hardy–Ramanujan theorem on the number of prime divisors	57
4.6	Distinct sums	61
4.7	Weierstrass approximation theorem	63
5	Chernoff bound	65
5.1	Discrepancy	67
5.2	Nearly equiangular vectors	69

5.3	Hajós conjecture counterexample	72
6	Lovász local lemma	75
6.1	Statement and proof	75
6.2	Coloring hypergraphs	79
6.3	Independent transversal	85
6.4	Directed cycles of length divisible by k	86
6.5	Lopsided local lemma	88
6.6	Algorithmic local lemma	93
7	Correlation inequalities	101
7.1	Harris–FKG inequality	101
7.2	Applications to random graphs	104
8	Janson inequalities	107
8.1	Probability of non-existence	107
8.2	Lower tails	113
8.3	Chromatic number of a random graph	116
9	Concentration of measure	119
9.1	Bounded differences inequality	119
9.2	Martingales concentration inequalities	120
9.3	Chromatic number of random graphs	125
9.4	Isoperimetric inequalities: a geometric perspective	129
9.5	Talagrand’s inequality	142
9.6	Euclidean traveling salesman problem	152
10	Entropy method	159
10.1	Basic properties	159
10.2	Permanent, perfect matchings, and Steiner triple systems	165
10.3	Sidorenko’s inequality	171
10.4	Shearer’s lemma	176
11	The container method	185
11.1	Containers for triangle-free graphs	187
11.2	Graph containers	190
11.3	Hypergraph container theorem	192

1 Introduction

The **probabilistic method** is an important technique in combinatorics. In a typical application, we wish to prove that a certain object/choice exists. We introduce randomness, and show that a random construction works with positive probability.

Let us begin with a simple example of this method.

Theorem 1.0.1 (Large bipartite subgraph)

Every graph $G = (V, E)$ contains a bipartite subgraph with at least $|E|/2$ edges.

Proof. Assign every vertex a color, either black or white, uniformly and independently at random.

Let E' be the set of edges with one black endpoint and one white endpoint. Then (V, E') is a bipartite subgraph of G .

Every edge belongs to E' with probability $1/2$. So by the linearity of expectation, the expected size of E' is

$$\mathbb{E}[|E'|] = \frac{1}{2} |E|.$$

Thus there is some coloring with $|E'| \geq \frac{1}{2} |E|$. Then (V, E') is the desired subgraph. \square

1.1 Lower bounds to Ramsey numbers

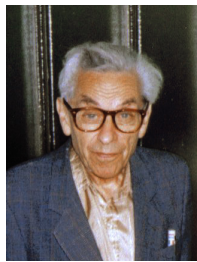
Ramsey number $R(k, \ell)$ = smallest n such that in every red-blue edge coloring of K_n , there exists a red K_k or a blue K_ℓ .

For example, $R(3, 3) = 6$ (every red/blue edge-coloring of K_6 has a monochromatic triangle, but one can color K_5 without any monochromatic triangle).

Ramsey (1929) proved that $R(k, \ell)$ exists (i.e., is finite). This is known as **Ramsey's theorem**.

Finding quantitative estimates of Ramsey numbers (and its generalizations) is generally a difficult and often fundamental problem in Ramsey theory.

1 Introduction



Paul Erdős (1913–1996) is considered the father of the probabilistic method. He published around 1,500 papers during his lifetime, and had more than 500 collaborators. To learn more about Erdős, see his biography *The man who loved only numbers* by Hoffman and the documentary *N is a number* (You may be able to watch this movie for free on [Kanopy](#) using your local public library account).



Frank Ramsey (1903–1930) wrote seminal papers in philosophy, economics, and mathematical logic, before his untimely death at the age of 26 from liver problems. See a recent profile of him in [the New Yorker](#).

Erdős' original proof

The probabilistic method started with the following theorem proved by Erdős in his seminal paper:

P. Erdős, Some remarks on the theory of graphs, *Bull. Amer. Math. Soc.*, 1947.

Remark 1.1.1 (Hungarian names). Many Hungarian mathematicians, starting with Erdős, made foundational contributions to this field. So we will encounter many Hungarian names. Here is some commonly encountered issues.

How to type “Erdős” in L^AT_EX: `Erd\H{o}s`
(*incorrect*: `Erd\"os`, which produces “Erdös”)

How to pronounce Hungarian names:

Hungarian spelling	Sounds like	Examples
s	<i>sh</i>	Erdős, Simonovits
sz	<i>s</i>	Szemerédi, Lovász

Theorem 1.1.2 (Lower bound to Ramsey numbers; Erdős 1947)

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

In other words, there exists a red-blue edge-coloring of K_n with no monochromatic K_k .

In the proof below, we will apply the **union bound**: for events E_1, \dots, E_m ,

$$\mathbb{P}(E_1 \cup \dots \cup E_m) \leq \mathbb{P}(E_1) + \dots + \mathbb{P}(E_m).$$

We usually think of each E_i as a “bad event” that we are trying to avoid.

Proof. Color edges of K_n with red or blue independently and uniformly at random.

For every fixed subset S of k vertices, let A_S denote the event that S induces a monochromatic K_k , so that $\mathbb{P}(A_S) = 2^{1-\binom{k}{2}}$. Then, by the union bound,

$$\mathbb{P}(\text{there is a monochromatic } K_k) = \mathbb{P}\left(\bigcup_{S \in \binom{[n]}{k}} A_S\right) \leq \sum_{S \in \binom{[n]}{k}} \mathbb{P}(A_S) = \binom{n}{k} 2^{1-\binom{k}{2}} < 1.$$

Thus, with positive probability, the random coloring gives no monochromatic K_k . So there exists some coloring with no monochromatic K_k . \square

Remark 1.1.3 (Quantitative bound). By optimizing n as a function of k in the theorem above (using Stirling’s formula), we obtain

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

The above argument can be also phrased as counting instead of randomness: of all $2^{\binom{n}{2}}$ possible colorings, not all are bad.

Indeed, Erdős’ 1947 paper actually was phrased in terms of counting (phrasing the argument at that time in terms of probability might have been heresy).

In this course, we mostly consider finite probability spaces. While in principle the finite probability arguments can be rephrased as counting, some of the later more involved arguments are impractical without a probabilistic perspective.

Remark 1.1.4 (Constructive lower bounds). The above proof only gives the existence of a red-blue edge-coloring of K_n without monochromatic cliques. Is there a way to find algorithmically find one? With an appropriate n , even though a random coloring achieves the goal with very high probability, there is no efficient method (in polynomial

1 Introduction

running time) to certificate that any specific edge-coloring avoids monochromatic cliques. So even though there are lots of Ramsey colorings, it is hard to find and certify an actual one. This difficulty has been described as *finding hay in a haystack*.

Finding constructive lower bounds is a major open problem. There was major progress on this problem stemming from connections to randomness extractors in computer science (e.g., [Barak et al. 2012](#), [Chattopadhyay & Zuckerman 2016](#), [Cohen 2017](#))

Remark 1.1.5 (Ramsey number upper bounds). Although Ramsey proved that Ramsey numbers are finite, his upper bounds are quite large. [Erdős–Szekeres \(1935\)](#) used a simple and nice inductive argument to show

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

The current best bound is due to [Sah \(2020+\)](#):

$$R(k+1, k+1) \leq e^{-c(\log k)^2} \binom{2k}{k}.$$

The above bounds all have the form $R(k, k) \leq (4 + o(1))^k$. It is a major open problem whether $R(k, k) \leq (4 - c)^k$ is true for some constant $c > 0$ and all sufficiently large k .

Alteration method

Let us give another argument that slightly improves the previous proof on Ramsey number lower bounds.

Instead of just taking a random coloring and analyzing it, we first randomly color, and then fix some undesirable features. This is called the *alteration method* (sometimes also the *deletion method*).

Theorem 1.1.6 (Ramsey lower bound via alteration)

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

Proof. We construct an edge-coloring of a clique in two steps:

- (1) Randomly color each edge of K_n with red or blue (independently and uniformly at random)
- (2) Delete a vertex from every monochromatic K_k .

The process yields a 2-edge-colored clique with no monochromatic K_k (since the second step destroyed all monochromatic cliques).

Let us now analyze how many vertices we get at the end.

Let X be the number of monochromatic K_k 's in the first step. Since each K_k is monochromatic with probability $2^{1-\binom{k}{2}}$, by the linearity of expectations,

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

In the second step, we delete at most $|X|$ vertices (since we delete one vertex from every clique). Thus final graph has size $\geq n - |X|$, which has expectation $n - \binom{n}{k} 2^{1-\binom{k}{2}}$.

Thus with positive probability, the remaining graph has $\geq n - \binom{n}{k} 2^{1-\binom{k}{2}}$ vertices (and no monochromatic K_k by construction) \square

Remark 1.1.7 (Quantitative bound). By optimizing the choice of n in the theorem, we obtain

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

which improves the previous bound by a constant factor of $\sqrt{2}$.

Lovász local lemma

Often we wish to avoid a set of “bad events” E_1, \dots, E_n . Here are two easy extremes:

- (Union bound) If $\sum_i \mathbb{P}(E_i) < 1$, then union bound tells us that we can avoid all bad events.
- (Independence) If all bad events are independent, then the probability that none of E_i occurs is $\prod_{i=1}^n (1 - \mathbb{P}(E_i)) > 0$ (provided that all $\mathbb{P}(E_i) < 1$).

What if we are in some intermediate situation, where the union bound is not good enough, and the bad events are not independent, but there are only few dependencies? The Lovász local lemma provides us a solution when each event is only independent with all but a small number of other events.

Here is a version of the Lovász local lemma, which we will prove later in Chapter 6.

Theorem 1.1.8 (Lovász local lemma — random variable model)

Let x_1, \dots, x_N be independent random variables. Let $B_1, \dots, B_m \subseteq [N]$. For each i , let E_i be an event that depends only on the variables indexed by B_i (i.e., E_i is allowed to depend only on $\{x_j : j \in B_i\}$).

Suppose, for every $i \in [m]$, B_i has non-empty intersections with at most d other B_j 's, and

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

Then with some positive probability, none of the events E_i occur.

Here $e = 2.71 \dots$ is the base of the natural logarithm. This constant turns out to be optimal in the above theorem.

Using the Lovász local lemma, let us give one more improvement to the Ramsey number lower bounds.

Theorem 1.1.9 (Ramsey lower bound via local lemma; Spencer 1977)

If $\left(\binom{k}{2}\binom{n}{k-2} + 1\right) 2^{1-\binom{k}{2}} < 1/e$, then $R(k, k) > n$.

Proof. Color the edges of K_n with red/blue uniformly and independently at random.

For each k -vertex subset S , let E_S be the event that S induces a monochromatic K_k . So $\mathbb{P}[E_S] = 2^{1-\binom{k}{2}}$.

In the setup of the local lemma, we have one independent random variable corresponding to each edge. Each event E_S depends only on the variables corresponding to the edges in S .

If S and S' are both k -vertex subsets, their cliques share an edge if and only if $|S \cap S'| \geq 2$. So for each S , there are at most $\binom{k}{2}\binom{n}{k-2}$ choices k -vertex sets S' with $|S \cap S'| \geq 2$. So the local lemma applies provided that

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

and in this case, with positive probability none of the events E_S occur, which means an edge-coloring with no monochromatic K_k 's. \square

Remark 1.1.10 (Quantitative lower bounds). By optimizing the choice of n , we obtain

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

once again improving the previous bound by a constant factor of $\sqrt{2}$. This is the best known lower bound to $R(k, k)$ to date.

1.2 Set systems

In *extremal set theory*, we often wish to understand the maximum size of a set system with some given property. A *set system* \mathcal{F} is a collection of subsets of some ground set, usually $[n]$. That is, each element of \mathcal{F} is a subset of $[n]$. We will see some classic results from extremal set theory each with a clever probabilistic proof.

Sperner's theorem

We say that a set family \mathcal{F} is an *antichain* if no element of \mathcal{F} is a subset of another element of \mathcal{F} (i.e., the elements of \mathcal{F} are pairwise incomparable by containment).

Question 1.2.1

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

The set $\mathcal{F} = \binom{[n]}{k}$ (i.e., all k -element subsets of $[n]$) has size $\binom{n}{k}$. It is an antichain (why?). The size $\binom{n}{k}$ is maximized when $k = \lfloor \frac{n}{2} \rfloor$ or $\lceil \frac{n}{2} \rceil$. The next result shows that this is indeed the best we can do.

Theorem 1.2.2 (Sperner's theorem, 1928)

If \mathcal{F} is an antichain of subsets of $\{1, 2, \dots, n\}$, then $|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}$.

In fact, we will show an even stronger result:

Theorem 1.2.3 (LYM inequality; Bollobás 1965, Lubell 1966, Meshalkin 1963, and Yamamoto 1954)

If \mathcal{F} is an antichain of subsets of $[n]$, then

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

Sperner's theorem follows since $\binom{n}{|A|} \leq \binom{n}{\lfloor n/2 \rfloor}$ for all A .

Proof. Let $\sigma(1), \dots, \sigma(n)$ be a permutation of $1, \dots, n$ chosen uniformly at random. Consider the chain:

$$\emptyset, \{\sigma(1)\}, \{\sigma(1), \sigma(2)\}, \{\sigma(1), \sigma(2), \sigma(3)\}, \dots, \{\sigma(1), \dots, \sigma(n)\}.$$

1 Introduction

For each $A \subseteq \{1, 2, \dots, n\}$, let E_A denote the event that A appears in the above chain. Then E_A occurs if and only if all the elements of A appears first in the permutation σ , followed by all the elements of $[n] \setminus A$. The number of such permutations is $|A|!(n - |A|)!$. Hence

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

Since \mathcal{F} is an antichain, if $A, B \in \mathcal{F}$ are distinct, then E_A and E_B cannot both occur. So $\{E_A : A \in \mathcal{F}\}$ is a set of disjoint events, and thus their probabilities sum to at most 1. This gives the desired inequality. \square

Bollobás' two families theorem

Theorem 1.2.4 (Bollobás' two families theorem 1965)

Let A_1, \dots, A_m be r -element sets and B_1, \dots, B_m be s -element sets such that $A_i \cap B_i = \emptyset$ for all i and $A_i \cap B_j \neq \emptyset$ for all $i \neq j$. Then $m \leq \binom{r+s}{r}$.

This bound is sharp: let A_i range over all r -element subsets of $[r + s]$ and set $B_i = [r + s] \setminus A_i$.

Let us give an application/motivation for Bollobás' two families theorem in terms of transversals. Given a set family \mathcal{F} , say that T is a **transversal** for \mathcal{F} if $T \cap S \neq \emptyset$ for all $S \in \mathcal{F}$ (i.e., T hits every element of \mathcal{F}). Let $\tau(\mathcal{F})$, the **transversal number** of \mathcal{F} , be the size of the smallest transversal of \mathcal{F} . Say that \mathcal{F} is **τ -critical** if $\tau(\mathcal{F}') < \tau(\mathcal{F})$ whenever \mathcal{F}' is a proper subset of \mathcal{F} .

Question 1.2.5

What is the maximum size of a τ -critical r -uniform \mathcal{F} with $\tau(\mathcal{F}) = s + 1$?

We claim that the answer is $\binom{r+s}{r}$. Indeed, let $\mathcal{F} = \{A_1, \dots, A_m\}$, and B_i an s -element transversal of $\mathcal{F} \setminus \{A_i\}$ for each i . Then the condition is satisfied. Thus $m \leq \binom{r+s}{r}$.

Conversely, $\mathcal{F} = \binom{[r+s]}{r}$ is τ -critical r -uniform with $\tau(\mathcal{F}) = s + 1$ (why?).

Here is a more general statement of the Bollobás' two-family theorem.

Theorem 1.2.6 (Bollobás' two families theorem 1965)

Let A_1, \dots, A_m and B_1, \dots, B_m be finite sets such that $A_i \cap B_i = \emptyset$ for all i and $A_i \cap B_j \neq \emptyset$ for all $i \neq j$. Then

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

Note that Sperner's theorem and LYM inequality are also special cases, since if $\{A_1, \dots, A_m\}$ is an antichain, then setting $B_i = [n] \setminus A_i$ for all i satisfies the hypothesis.

Proof. The proof is a modification of the proof of the LYM inequality earlier.

Consider a uniform random ordering of all elements that appear in the union of the sets.

Let E_i be the event that all elements of A_i appear before B_i . Then

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

Note that the events E_i are disjoint, since E_i and E_j both occurring would contradict the hypothesis for A_i, B_i, A_j, B_j (why?). Thus $\sum_i \mathbb{P}(E_i) \leq 1$. This yields the claimed inequality. \square

Bollobas' two families theorem has many interesting generalizations that we will not discuss here (e.g., see [Gil Kalai's blog post](#)). There are also beautiful linear algebraic proofs of this theorem and its extensions.

Erdős–Ko–Rado theorem on intersecting families

We say that a family \mathcal{F} is *intersecting* if $A \cap B \neq \emptyset$ for all $A, B \in \mathcal{F}$. That is, no two sets in \mathcal{F} are disjoint.

Here is an easy warm up.

Question 1.2.7 (Intersecting family—unrestricted sizes)

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

Proof. Solution One way to generate a large intersecting family is to include all sets that contain a fixed element (say, the element 1). This family has size 2^{n-1} and is clearly intersecting. (This isn't the only example with size 2^{n-1} ; can you think of other intersecting families with the same size?)

1 Introduction

It turns out that one cannot do better than 2^{n-1} . Since we can pair up each subset of $[n]$ with its complement. At most one of A and $[n] \setminus A$ can be in an intersecting family. And so at most half of all sets can be in an intersecting family. \square

The question becomes much more interesting if we restrict to k -uniform families.

Question 1.2.8 (k -uniform intersecting family)

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

Example: \mathcal{F} = all subsets containing the element 1. Then \mathcal{F} is intersecting and $|\mathcal{F}| = \binom{n-1}{k-1}$.

Theorem 1.2.9 (Erdős–Ko–Rado 1961; proved in 1938)

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

Remark 1.2.10. The assumption $n \geq 2k$ is necessary since if $n < 2k$, then the family of all k -element subsets of $[n]$ is automatically intersecting by pigeonhole.

Proof. Consider a uniform random circular permutation of $1, 2, \dots, n$ (arrange them randomly around a circle)

For each k -element subset A of $[n]$, we say that A is **contiguous** if all the elements of A lie in a contiguous block on the circle.

The probability that A forms a contiguous set on the circle is exactly $n / \binom{n}{k}$.

So the expected number of contiguous sets in \mathcal{F} is exactly $n |\mathcal{F}| / \binom{n}{k}$.

Since \mathcal{F} is intersecting, there are at most k contiguous sets in \mathcal{F} (under every circular ordering of $[n]$). Indeed, suppose that $A \in \mathcal{F}$ is contiguous. Then there are $2(k-1)$ other contiguous sets (not necessarily in \mathcal{F}) that intersect A , but they can be paired off into disjoint pairs. Since \mathcal{F} is intersecting, it follows that it contains at most k contiguous sets.

Combining with result from the previous paragraph, we see that $n |\mathcal{F}| / \binom{n}{k} \leq k$, and hence $|\mathcal{F}| \leq \frac{k}{n} \binom{n}{k} = \binom{n-1}{k-1}$. \square

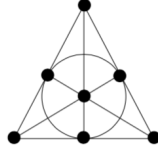
1.3 2-colorable hypergraphs

An **k -uniform hypergraph** (or **k -graph**) is a pair $H = (V, E)$, where V (vertices) is a finite set and E (edges) is a set of k -element subsets of V , i.e., $E \subseteq \binom{V}{k}$. (So hypergraphs and set families are the same concept, just different names.)

We say that H is ***r-colorable*** if the vertices can be colored using r colors so that no edge is monochromatic.

Let $m(k)$ denote the minimum number of edges in a k -uniform hypergraph that is not 2-colorable (elsewhere in the literature, “2-colorable” = “property B”, named after Bernstein who introduced the concept in 1908). Some small values:

- $m(2) = 3$
- $m(3) = 7$. Example: Fano plane (below) is not 2-colorable (the fact there are no 6-edge non-2-colorable 3-graphs is proved by exhaustive search).



- $m(4) = 23$, proved via exhaustive computer search (Östergård 2014)

Exact value of $m(k)$ is unknown for all $k \geq 5$. However, we can get some asymptotic lower and upper bounds using the probability method.

Theorem 1.3.1 (Erdős 1964)

$m(k) \geq 2^{k-1}$ for every $k \geq 2$.

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

Proof. Let there be $m < 2^{k-1}$ edges. In a random 2-coloring, the probability that there is a monochromatic edge is $\leq 2^{-k+1}m < 1$. \square

Remark 1.3.2. Later on we will prove a better lower bound $m(k) \gtrsim 2^k \sqrt{k/\log k}$, which is the best known to date.

Perhaps somewhat surprisingly, the state of the art upper bound is also proved using probabilistic method (random construction).

Theorem 1.3.3 (Erdős 1964)

$m(k) = O(k^2 2^k)$.

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

Proof. Let $|V| = n = k^2$ (this choice is motivated by the displayed inequality below). Let H be the k -uniform hypergraph obtained by choosing m edges S_1, \dots, S_m independently and uniformly at random (i.e., with replacement) among $\binom{V}{k}$.

1 Introduction

Given a coloring $\chi: V \rightarrow [2]$, if χ colors a vertices with one color and b vertices with the other color, then the probability that the (random) edge S_1 is monochromatic under the (non-random) coloring χ is

$$\begin{aligned} \frac{\binom{a}{k} + \binom{b}{k}}{\binom{n}{k}} &\geq \frac{2\binom{n/2}{k}}{\binom{n}{k}} = \frac{2(n/2)(n/2-1)\cdots(n/2-k+1)}{n(n-1)\cdots(n-k+1)} \geq 2\left(\frac{n/2-k+1}{n-k+1}\right)^k \\ &= 2^{-k+1} \left(1 - \frac{k-1}{n-k+1}\right)^k = 2^{-k+1} \left(1 - \frac{k-1}{k^2-k+1}\right)^k \geq c2^{-k} \end{aligned}$$

for some constant $c > 0$.

Since the edges are chosen independently at random, for any coloring χ ,

$$\mathbb{P}(\chi \text{ is a proper coloring}) \leq (1 - c2^{-k})^m \leq e^{-c2^{-k}m}$$

(using $1 + x \leq e^x$ for all real x). By the union bound,

$$\begin{aligned} \mathbb{P}(\text{the random hypergraph has a proper 2-coloring}) &\leq \sum_{\chi} \mathbb{P}(\chi \text{ is a proper coloring}) \\ &\leq 2^n e^{-c2^{-k}m} < 1 \end{aligned}$$

for some $m = O(k^2 2^k)$ (recall $n = k^2$). Thus there exists a non-2-colorable k -uniform hypergraph with m edges. \square

1.4 List chromatic number of $K_{n,n}$

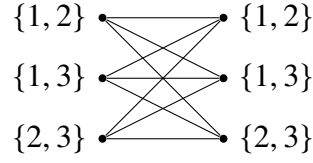
Given a graph G , its **chromatic number** $\chi(G)$ is the minimum number of colors required to proper color its vertices.

In **list coloring**, each vertex of G is assigned a list of allowable colors. We say that G is **k -choosable** (also called **k -list colorable**) if it has a proper coloring no matter how one assigns a list of k colors to each vertex.

We write $\text{ch}(G)$, called the **list chromatic number** (also called: **choice number**, **choosability**, **list colorability**) of G , to be the smallest k so that G is k -choosable.

We have $\chi(G) \leq \text{ch}(G)$ by assigning the same list of colors to each vertex. The inequality may be strict, as we will see below.

For example, while every bipartite graph is 2-colorable, $K_{3,3}$ is not 2-choosable. Indeed, no list coloring of $K_{3,3}$ is possible with color lists (check!):



Exercise: check that $\text{ch}(K_{3,3}) = 3$.

Question 1.4.1

What is the asymptotic behavior of $\text{ch}(K_{n,n})$?

First we prove an upper bound on $\text{ch}(K_{n,n})$.

Theorem 1.4.2

If $n < 2^{k-1}$, then $K_{n,n}$ is k -choosable.

In other words, $\text{ch}(K_{n,n}) \leq \lfloor \log_2(2n) \rfloor + 1$.

Proof. For each color, mark it either L or R independently and uniformly at random.

For any vertex of $K_{n,n}$ on the left part, remove all its colors marked R.

For any vertex of $K_{n,n}$ on the right part, remove all its colors marked L.

The probability that some vertex has no colors remaining is at most $2n2^{-k} < 1$ by the union bound. So with positive probability, every vertex has some color remaining. Assign the colors arbitrarily for a valid coloring. \square

The lower bound on $\text{ch}(K_{n,n})$ turns out to follow from the existence of non-2-colorable k -uniform hypergraph with many edges.

Theorem 1.4.3

If there exists a non-2-colorable k -uniform hypergraph with n edges, then $K_{n,n}$ is not k -choosable.

Proof. Let $H = (V, E)$ be a non-2-colorable k -uniform hypergraph $|E| = n$ edges. Now, view V as colors and assign to the i -th vertex of K_n on both the left and right bipartitions a list of colors given by the i -th edge of H . We leave it as an exercise to check that this $K_{n,n}$ is not list colorable. \square

Recall from Theorem 1.3.3 that there exists a non-2-colorable k -uniform hypergraph with $O(k^2 2^k)$ edges. Thus $\text{ch}(K_{n,n}) > (1 - o(1)) \log_2 n$.

Putting these bounds together:

Corollary 1.4.4 (List chromatic number of a complete bipartite graph)

$$\text{ch}(K_{n,n}) = (1 + o(1)) \log_2 n$$

It turns out that, unlike the chromatic number, the list chromatic number always grows with the average degree. The following result was proved using the method of *hypergraph containers*, a very important modern development in combinatorics that we will see a glimpse of in Chapter 11. It provides the optimal asymptotic dependence (the example of $K_{n,n}$ shows optimality).

Theorem 1.4.5 (Saxton and Thomason 2015)

If a graph G has average degree d , then, as $d \rightarrow \infty$,

$$\text{ch}(G) > (1 + o(1)) \log_2 d.$$

They also proved similar results for the list chromatic number of hypergraphs. For graphs, a slightly weaker result, off by a factor of 2, was proved earlier by [Alon \(2000\)](#).

2 Linearity of expectations

Linearity of expectations refers to the following basic fact about the expectation: given random variables X_1, \dots, X_n and constants c_1, \dots, c_n ,

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

This identity does not require any assumption of independence. On the other hand, generally $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are uncorrelated (independent random variables are always uncorrelated).

Here is a simple application (there are also much more involved solutions via enumeration methods).

Question 2.0.1 (Expected number of fixed points)

What is the average number of fixed points of a uniform random permutation of an n element set?

Solution. Let X_i be the event that element $i \in [n]$ is fixed. Then $\mathbb{E}[X_i] = 1/n$. The expected number of fixed points is

$$\mathbb{E}[X_1 + \dots + X_n] = \mathbb{E}[X_1] + \dots + \mathbb{E}[X_n] = 1. \quad \square$$

2.1 Hamiltonian paths in tournaments

We frequently use the following fact:

with positive probability, $X \geq \mathbb{E}[X]$ (likewise for $X \leq \mathbb{E}[X]$).

A **tournament** is a directed complete graph. A **Hamilton path** in a directed graph is a directed path that contains every vertex exactly once.

Question 2.1.1 (Number of Hamilton paths in a tournament)

What is the maximum (and minimum) number of Hamilton paths in an n -vertex tournament?

2 Linearity of expectations

The minimization problem is easier. The transitive tournament (i.e., respecting a fixed linear ordering of vertices) has exactly one Hamilton path. On the other hand, every tournament has at least one Hamilton path (Exercise: prove this! Hint: consider a longest directed path).

The maximization problem is more difficult and interesting. Here we have some asymptotic results.

Theorem 2.1.2 (Tournaments with many Hamilton paths; Szele 1943)

There is a tournament on n vertices with at least $n!2^{-(n-1)}$ Hamilton paths

Proof. Consider a random tournament where every edge is given a random orientation chosen uniformly and independently. Each of the $n!$ permutations of vertices forms a directed path with probability 2^{-n+1} . So that expected number of Hamilton paths is $n!2^{-n+1}$. Thus, there exists a tournament with at least this many Hamilton paths. \square

This was considered the first use of the probabilistic method. Szele conjectured that the maximum number of Hamilton paths in a tournament on n players is $n!/(2 - o(1))^n$. This was proved by Alon (1990) using the Minc–Brégman theorem on permanents (we will see this later in Chapter 10 on the entropy method).

2.2 Sum-free subset

A subset A in an abelian group is *sum-free* if there do not exist $a, b, c \in A$ with $a + b = c$.

Does every n -element set contain a large sum-free set?

Theorem 2.2.1 (Large sum-free subsets; Erdős 1965)

Every set of n nonzero integers contains a sum-free subset of size $\geq n/3$.

Proof. Let $A \subseteq \mathbb{Z} \setminus \{0\}$ with $|A| = n$. For $\theta \in [0, 1]$, let

$$A_\theta := \{a \in A : \{a\theta\} \in (1/3, 2/3)\}$$

where $\{\cdot\}$ denotes fractional part. Then A_θ is sum-free since $(1/3, 2/3)$ is sum-free in \mathbb{R}/\mathbb{Z} .

For θ uniformly chosen at random, $\{a\theta\}$ is also uniformly random in $[0, 1]$, so $\mathbb{P}(a \in A_\theta) = 1/3$. By linearity of expectations, $\mathbb{E}|A_\theta| = n/3$. \square

Remark 2.2.2 (Additional results). Alon and Kleitman (1990) noted that one can improve the bound to $\geq (n+1)/3$ by noting that $|A_\theta| = 0$ for θ close to zero (say, $|\theta| < (3 \max_{a \in A} |a|)^{-1}$), so that $|A_\theta| < n/3$ with positive probability, and hence $|A_\theta| > n/3$ with positive probability. Note that since $|A_\theta|$ is an integer, being $> n/3$ is the same as being $\geq (n+1)/3$.

Bourgain (1997) improved it to $\geq (n+2)/3$ via a difficult Fourier analytic argument. This is currently the best lower bound known.

It remains an open problem to prove $\geq (n + f(n))/3$ for some function $f(n) \rightarrow \infty$.

In the other direction, Eberhard, Green, and Manners (2014) showed that there exist n -element sets of integers whose largest sum-free subset has size $\leq (1/3 + o(1))n$.

2.3 Turán's theorem and independent sets

Question 2.3.1 (Turán problem)

What is the maximum number of edges in an n -vertex K_k -free graph?

Taking the complement of a graph changes its independent sets to cliques and vice versa. So the problem is equivalent to one about graphs without large independent sets.

The following result, due to Caro (1979) and Wei (1981), shows that a graph with small degrees much contain large independent sets. The probabilistic method proof shown here is due to Alon and Spencer.

Theorem 2.3.2 (Caro 1979, Wei 1981)

Every graph G contains an independent set of size at least

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

where d_v is the degree of vertex v .

Proof. Consider a random ordering (permutation) of the vertices. Let I be the set of vertices that appear before all of its neighbors. Then I is an independent set.

For each $v \in V$, $\mathbb{P}(v \in I) = \frac{1}{1+d_v}$ (this is the probability that v appears first among $\{v\} \cup N(v)$). Thus $\mathbb{E}|I| = \sum_{v \in V(G)} \frac{1}{d_v+1}$. Thus with positive probability, $|I|$ is at least this expectation. \square

Remark 2.3.3. Equality occurs if G is a disjoint union of cliques.

2 Linearity of expectations

Remark 2.3.4 (Derandomization). Here is an alternative “greedy algorithm” proof of the Caro–Wei inequality. At each step, take a vertex of smallest degree, and remove it and all its neighbors. If each vertex v is assigned weight $1/(d_v + 1)$, then the total weight removed at each step is at most 1. Thus there must be at least $\sum_v 1/(d_v + 1)$ steps.

Some probabilistic proofs, especially those involving linearity of expectations, can be derandomized this way into an efficient deterministic algorithm. However, for many other proofs (such as Ramsey lower bounds from Section 1.1), it is not known how to derandomize the proof.

By taking the complement of the graph, independent sets become cliques, and so we obtain the following corollary.

Corollary 2.3.5

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

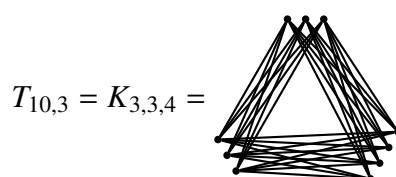
$$\sum_{v \in V(G)} \frac{1}{n - d_v}.$$

Note that equality is attained when G is multipartite.

Now let us answer the earlier question about maximizing the number of edges in a K_{r+1} -free graph.

The **Turán graph** $T_{n,r}$ is the complete multipartite graph formed by partitioning n vertices into r parts with sizes as equal as possible (differing by at most 1).

Example:



It is easy to see that $T_{n,r}$ is K_{r+1} -free.

Turán’s theorem (1941) tells us that $T_{n,r}$ indeed maximizes the number of edges among n -vertex K_{r+1} -free graphs. We will prove a slightly weaker statement, below, which is tight when n is divisible by r .

Theorem 2.3.6 (Turán's theorem 1941)

The number of edges in an n -vertex K_{r+1} -free graph is at most

$$\left(1 - \frac{1}{r}\right) \frac{n^2}{2}.$$

Proof. Let m be the number of edges. Since G is K_{r+1} -free, by Corollary 2.3.5, the size $\omega(G)$ of the largest clique of G satisfies

$$r \geq \omega(G) \geq \sum_{v \in V} \frac{1}{n - d_v} \geq \frac{n}{n - \frac{1}{n} \sum_v d_v} = \frac{n}{n - \frac{2m}{n}}.$$

Rearranging gives $m \leq \left(1 - \frac{1}{r}\right) \frac{n^2}{2}$. □

Remark 2.3.7. By a careful refinement of the above argument, we can deduce Turán's theorem that $T_{n,r}$ maximizes the number of edges in a n -vertex K_{r+1} -free graph, by noting that $\sum_{v \in V} \frac{1}{n - d_v}$ is minimized over fixed $\sum_v d_v$ when the degrees are nearly equal.

Also, Theorem 2.3.6 is asymptotically tight in the sense that the Turán graph $T_{n,r}$, for fixed r and $n \rightarrow \infty$, as $(1 - 1/r - o(1))n^2/2$ edges.

For more on this topic, see Chapter 1 of my textbook *Graph Theory and Additive Combinatorics* and the class with the same title.

2.4 Sampling

By Turán's theorem (actually Mantel's theorem, in this case for triangles, the maximum number of edges in an n -vertex triangle-free graph is $\lfloor n^2/4 \rfloor$).

How about the problem for hypergraphs? A **tetrahedron**, denoted $K_4^{(3)}$, is a complete 3-uniform hypergraph (3-graph) on 4 vertices (think of the faces of a usual 3-dimensional tetrahedron).

Question 2.4.1 (Hypergraph Turán problem for tetrahra)

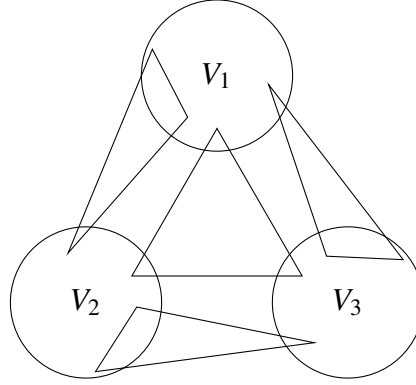
What is the maximum number of edges in an n -vertex 3-uniform hypergraph not containing any tetrahedra?

This turns out to be a notorious open problem. Turán conjectured that the answer is

$$\left(\frac{5}{9} + o(1)\right) \binom{n}{3},$$

2 Linearity of expectations

which can be achieved using the 3-graph illustrated below:



Above, the vertices are partitioned into three nearly equal sets V_1, V_2, V_3 , and all the edges come in two types: (i) with one vertex in each of the three parts, and (ii) two vertices in V_i and one vertex in V_{i+1} , with the indices considered mod 3.

Let us give some easy upper bounds, in order to illustrate a simple yet important technique of **bounding by sampling**.

Proposition 2.4.2 (A cheap sampling bound)

Every tetrahedron-free 3-graph on $n \geq 4$ vertices has at most $\frac{3}{4} \binom{n}{3}$ edges.

Proof. Let S be a 4-vertex subset chosen uniformly at random. If the graph has $p \binom{n}{3}$ edges, then the expected number of edges induced by S is $4p$ by linearity of expectations (why?).

Since the 3-graph is tetrahedron-free, S induces at most 3 edges. Therefore, $4p \leq 3$. Thus the total number of edges is $p \binom{n}{3} \leq \frac{3}{4} \binom{n}{3}$. \square

Why stop at sampling four vertices? Can we do better by sampling five vertices? To run the above argument, we will know how many edges can there be in a 5-vertex tetrahedron-free graph.

Lemma 2.4.3

A 5-vertex tetrahedron-free 3-graph has at most 7 edges.

Proof. We can convert a 5-vertex 3-graph H to a 5-vertex graph G , by replacing each triple by its complement. Then H being tetrahedron-free is equivalent to G not having a vertex of degree 4. The maximum number of edges in a 5-vertex graph with maximum degree at most 3 is $\lfloor 3 \cdot 5/2 \rfloor = 7$ (check this can be achieved). \square

We can improve Proposition 2.4.2 by sampling 5 vertices instead of 4 in its proof. This yields (check):

Proposition 2.4.4

Every tetrahedron-free 3-graph on $n \geq 4$ vertices has at most $\frac{7}{10} \binom{n}{3}$ edges.

By sampling s vertices and using brute-force search to solve the s -vertex problem, we can improve the upper bound by taking larger values of s . In fact in principle, if we had unlimited computational power, we can arbitrarily close to optimum by taking sufficiently large s (why?). However, this is not a practical method due to the cost of the brute-force search. There are more clever ways to get better bounds (also with the help of a computer). The best known upper bound notably via a method known as *flag algebras* (using sums of squares) due to Razborov, which can give $\leq (0.561 \dots) \binom{n}{3}$.

For more on the Hypergraph Turán problem, see the survey by [Keevash \(2011\)](#).

2.5 Unbalancing lights

Consider an $n \times n$ array of light bulbs. Initially some arbitrary subset of the light bulbs are turned on. We are allowed up toggle the lights (on/off) for an entire row or column at a time. How many lights can be guarantee to turn on?

If we flip each row/column independently with probability $1/2$, then on expectation, we get exactly half of the lights to turn on. Can we do better?

In the probabilistic method, not every step has to be random. A better strategy is to first flip all the columns randomly, and then decide what to do with each row greedily based on what has happened so far. This is captured in the following theorem, where the left-hand side represents

$$\# \{\text{bulbs on}\} - \# \{\text{bulbs off}\}.$$

Theorem 2.5.1

Let $a_{ij} \in \{-1, 1\}$ for all $i, j \in [n]$. There exists $x_i, y_j \in \{-1, 1\}$ for all $i, j \in [n]$ such that

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

Proof. Choose $y_1, \dots, y_n \in \{-1, 1\}$ independently and uniformly at random. For each

2 Linearity of expectations

i , let

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

and set $x_i \in \{-1, 1\}$ to be the sign of R_i (arbitrarily choose x_i if $R_i = 0$). Then the LHS sum is

$$\sum_{i=1}^n R_i x_i = \sum_{i=1}^n |R_i|.$$

For each i , R_i has the same distribution as a sum of n i.i.d. uniform $\{-1, 1\}$: $S_n = \varepsilon_1 + \dots + \varepsilon_n$ (note that R_i 's are not independent for different i 's). Thus, for each i

$$\mathbb{E}[|R_i|] = \mathbb{E}[|S_n|] = \left(\sqrt{\frac{2}{\pi}} + o(1) \right) \sqrt{n},$$

since by the central limit theorem

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} \left[\frac{|S_n|}{\sqrt{n}} \right] &= \mathbb{E}[|X|] \quad \text{where } X \sim \text{Normal}(0, 1) \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |x| e^{-x^2/2} dx = \sqrt{\frac{2}{\pi}} \end{aligned}$$

(one can also use binomial sum identities to compute exactly: $\mathbb{E}[|S_n|] = n2^{1-n} \binom{n-1}{\lfloor (n-1)/2 \rfloor}$, though it is rather unnecessary to do so.) Thus

$$\mathbb{E} \sum_{i=1}^n |R_i| = \left(\sqrt{\frac{2}{\pi}} + o(1) \right) n^{3/2}.$$

Thus with positive probability, the sum is $\geq \left(\sqrt{\frac{2}{\pi}} + o(1) \right) n^{3/2}$. □

The next example is tricky. The proof will set up a probabilistic process where the parameters are not given explicitly. A compactness argument will show that a good choice of parameters exists.

Theorem 2.5.2

Let $k \geq 2$. Let $V = V_1 \cup \dots \cup V_k$, where V_1, \dots, V_k are disjoint sets of size n . The edges of the complete k -uniform hypergraph on V are colored with red/blue. Suppose that every edge formed by taking one vertex from each V_1, \dots, V_k is colored blue. Then there exists $S \subseteq V$ such that the number of red edges and blue edges in S differ by more than $c_k n^k$, where $c_k > 0$ is a constant.

Proof. We will write this proof for $k = 3$ for notational simplicity. The same proof

works for any k .

Let p_1, p_2, p_3 be real numbers to be decided. We are going to pick S randomly by including each vertex in V_i with probability p_i , independently. Let

$$a_{i,j,k} = \#\{\text{blue edges in } V_i \times V_j \times V_k\} - \#\{\text{red edges in } V_i \times V_j \times V_k\}.$$

Then

$$\mathbb{E}[\#\{\text{blue edges in } S\} - \#\{\text{red edges in } S\}]$$

equals to some polynomial

$$f(p_1, p_2, p_3) = \sum_{i \leq j \leq k} a_{i,j,k} p_i p_j p_k = n^3 p_1 p_2 p_3 + a_{1,1,1} p_1^3 + a_{1,1,2} p_1^2 p_2 + \cdots.$$

(note that $a_{1,2,3} = n^3$ by hypothesis). We would be done if we can find $p_1, p_2, p_3 \in [0, 1]$ such that $|f(p_1, p_2, p_3)| > c$ for some constant $c > 0$ (not depending on the $a_{i,j,k}$'s). Note that $|a_{i,j,k}| \leq n^3$. We are done after the following lemma

Lemma 2.5.3

Let P_k denote the set of polynomials $g(p_1, \dots, p_k)$ of degree k , whose coefficients have absolute value ≤ 1 , and the coefficient of $p_1 p_2 \cdots p_k$ is 1. Then there is a constant $c_k > 0$ such that for all $g \in P_k$, there is some $p_1, \dots, p_k \in [0, 1]$ with $|g(p_1, \dots, p_k)| \geq c_k$.

Proof of Lemma. Set $M(g) = \sup_{p_1, \dots, p_k \in [0, 1]} |g(p_1, \dots, p_k)|$ (note that sup is achieved as max due to compactness). For $g \in P_k$, since g is nonzero (its coefficient of $p_1 p_2 \cdots p_k$ is 1), we have $M(g) > 0$. As P_k is compact and $M: P_k \rightarrow \mathbb{R}$ is continuous, M attains a minimum value $c = M(g) > 0$ for some $g \in P_k$. ■ □

2.6 Crossing number inequality

Consider drawings of graphs on a plane using continuous curves as edges.

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

A graph is **planar** if $\text{cr}(G) = 0$.

The graphs $K_{3,3}$ and K_5 are non-planar. Furthermore, the following theorem characterizes these two graphs as the only obstructions to planarity:

Kuratowski's theorem (1930). Every non-planar graph contains a subgraph that is topologically homeomorphic to $K_{3,3}$ or K_5 .

2 Linearity of expectations

Wagner's theorem (1937). A graph is planar if and only if it does not have $K_{3,3}$ or K_5 as a minor.

(It is not too hard to show that Wagner's theorem and Kuratowski's theorem are equivalent)

If a graph has a lot of edges, is it guaranteed to have a lot of crossings no matter how it is drawn in the plane?

Question 2.6.1

What is the minimum possible number of crossings that a drawing of:

- K_n ? (Hill's conjecture)
- $K_{n,n}$? (Zarankiewicz conjecture; Turán's brick factory problem)
- a graph on n vertices and $n^2/100$ edges?

The following result, due to [Ajtai–Chvátal–Newborn–Szemerédi \(1982\)](#) and [Leighton \(1984\)](#), lower bounds the number of crossings for graphs with many edges.

Theorem 2.6.2 (Crossing number inequality)

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

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

Remark 2.6.3. The constant 4 in $|E| \geq 4|V|$ can be replaced by any constant greater than 3 (at the cost of changing the constant in the conclusion). On the other hand, by considering a large triangular grid, we get a planar graph with average degree arbitrarily close to 6.

Corollary 2.6.4

In a graph $G = (V, E)$, if $|E| \gtrsim |V|^2$, then $\text{cr}(G) \gtrsim |V|^4$.

Proof. Recall **Euler's formula**: $v - e + f = 2$ for every connected planar drawing of graph. Here v is the number of vertices, e the number of edges, and f the number of faces (connected components of the complement of the drawing, including the outer infinite region).

For every connected planar graph with at least one cycle, $3|F| \leq 2|E|$ since every face is adjacent to ≥ 3 edges, whereas every edge is adjacent to exactly 2 faces. Plugging into Euler's formula, $|E| \leq 3|V| - 6$.

Thus $|E| \leq 3|V|$ for all planar graphs. Hence $\text{cr}(G) > 0$ whenever $|E| > 3|V|$.

The above argument gives us one crossing. Next, we will use it to obtain many crossings.

By deleting one edge for each crossing, we get a planar graph, so $|E| - \text{cr}(G) \leq 3|V|$, that is

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

This is a “cheap bound.” For graphs with $|E| = \Theta(n^2)$, this gives $\text{cr}(G) \gtrsim n^2$. This is not a great bound. We will use the probabilistic method to boost this bound.

Let $p \in [0, 1]$ to be decided. Let $G' = (V', E')$ be obtained from G by randomly keeping each vertex with probability p . Then

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

So

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

We have $\mathbb{E} \text{cr}(G') \leq p^4 \text{cr}(G)$, $\mathbb{E}|E'| = p^2|E|$ and $\mathbb{E}|V'| = p|V|$. So

$$p^4 \text{cr}(G) \geq p^2|E| - 3p|V|.$$

Thus

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

Setting $p = 4|V|/|E| \in [0, 1]$ (here is where we use the hypothesis that $|E| \geq 4|V|$) so that $4p^{-3}|V| = p^{-2}|E|$, we obtain $\text{cr}(G) \gtrsim |E|^3/|V|^2$. \square

Remark 2.6.5. The above idea of boosting a cheap bound to a better bound is an important one. We saw a version of this idea in Section 2.4 where we sampled a constant number of vertices to deduce upper bounds on the hypergraph Turán number. In the above crossing number inequality application, we are also applying some preliminary cheap bound to some sampled induced subgraph, though this time the sampled subgraph has super-constant size.

It is tempting to modify the proof by sampling edges instead of vertices, but this does not work.

2.7 Dense packing of spheres in high dimensions

Question 2.7.1 (Maximum sphere packing density)

What is the maximum density of a packing of non-overlapping unit balls in \mathbb{R}^n for large n ?

Here the **density** is fraction of volume occupied (fraction of the box $[-n, n]^d$ as $n \rightarrow \infty$)

Let Δ_n denote the supremum of unit ball packing densities in \mathbb{R}^n .

Exact maximum is only solved in dimensions 1, 2, 3, 8, 24. Maryna Viazovska was awarded a Fields Medal in 2022 for her breakthrough in solving the problem in dimensions 8 and 24. Dimensions 8 and 24 are special because of the existences of highly symmetric lattices (E_8 lattice in dimension 8 and Leech lattice in dimension 24).

What are examples of dense packings?

We can add balls greedily. Any *maximal* packing has density $\geq 2^{-n}$. Doubling the ball radius would cover space

What about lattices? \mathbb{Z}^n has sphere packing density $\text{vol}(B(1/2)) = \frac{\pi^{n/2}}{(n/2)!2^n} < n^{-cn}$.

Best upper bound: [Kabatiansky–Levenshtein \(1978\)](#): $\Delta_n \leq 2^{-(0.599 \dots + o(1))n}$

Existence of a dense lattice? (Optimal lattices known in dimensions 1–8 and 24)

We will use the probabilistic method to show that a random lattice has high density.

(Aside: is it generally believed, although not rigorously proved in any dimension, that in most high dimensions the optimal packing is non-lattice. In dimension 10, the best known packing is non-lattice, found by Marc Best—sometimes funnily and confusingly referred to as the “Best packing”).

How does one pick a random lattice?

A **lattice** the \mathbb{Z} -span of its basis vectors v_1, \dots, v_n . It’s covolume (volume of its fundamental domain) is given by $|\det(v_1|v_2| \dots |v_n)|$.

So every matrix in $\text{SL}_n(\mathbb{R})$ corresponds to a unimodular lattice (i.e., covolume 1).

Every lattice can be represented in different ways by picking a different basis (e.g., $\{v_1 + v_2, v_2\}$). The matrices $A, A' \in \text{SL}_n(\mathbb{R})$ represent the same lattice if and only if $A' = AU$ for some $U \in \text{SL}_n(\mathbb{Z})$.

So the space of unimodular lattices is $\text{SL}_n(\mathbb{R})/\text{SL}_n(\mathbb{Z})$ (i.e., the space of left cosets), which has a finite Haar measure (even though this space not compact), so can normalize to a probability measure.

2.7 Dense packing of spheres in high dimensions

We can pick a **random unimodular lattice** in \mathbb{R}^n by picking a random point in $\mathrm{SL}_n(\mathbb{R})/\mathrm{SL}_n(\mathbb{Z})$ according to its Haar probability measure.

The following classic result of Siegel acts as like a linearity of expectations statement for random lattices.

Theorem 2.7.2 (Siegel mean value theorem)

Let L be the random lattice in \mathbb{R}^n as above and $S \subseteq \mathbb{R}^n$. Then

$$\mathbb{E}|S \cap L \setminus \{0\}| = \lambda_{\mathrm{Leb}}(S)$$

- Proof sketch.*
1. $\mu(S) = \mathbb{E}|S \cap L \setminus \{0\}|$ defines a measure on \mathbb{R}^n (it is additive by linearity of expectations)
 2. This measure is invariant under $\mathrm{SL}_n(\mathbb{R})$ action (since the random lattice is chosen with respect to Haar measure)
 3. Every $\mathrm{SL}_n(\mathbb{R})$ -invariant measure on \mathbb{R}^n is a constant multiple of the Lebesgue measure.
 4. By considering a large ball S , deduce that $c = 1$. □

Theorem 2.7.3 (Minkowski 1905)

For every n , there exist a lattice sphere packing in \mathbb{R}^n with density $\geq 2^{-n}$.

Proof. Let S be a ball of volume 1 (think $1 - \varepsilon$ for arbitrarily small $\varepsilon > 0$ if you like) centered at the origin. By the Siegel mean value theorem, the random lattice has expected 1 nonzero lattice point in S , so with positive probability it has no nonzero lattice point in S . Putting a copy of $\frac{1}{2}S$ (volume 2^{-n}) at each lattice point then gives a lattice packing of density $\geq 2^{-n}$ □

Here is a factor 2 improvement. Take S to be a ball of volume 2. Note that the number of nonzero lattice points in S must be even (if $x \in S$ then $-x \in S$). So same argument gives lattice packing of density $\geq 2^{-n+1}$.

The above improvement uses 2-fold symmetry of \mathbb{R}^n . Can we do better by introducing more symmetry?

Historically, there were several improvements of the form $\geq cn2^{-n}$ for a sequence of improving constants $c > 0$

Venkatesh (2012) showed that one can get a lattice with a k -fold symmetry by building it using two copies of the cyclotomic lattice $\mathbb{Z}[\omega]$ where $\omega = e^{2\pi/k}$. Every lattice of this form has k -fold symmetry by multiplication by ω .

2 Linearity of expectations

Skipping details, one can extend the earlier idea to choose a random unimodular lattice in dimension $n = 2\phi(k)$ with k -fold length-preserving symmetry (without fixed points). An extension of Siegel mean value theorem also holds in this case.

By apply same argument with S being a ball of volume k , we get a lattice packing of density $\geq k2^{-n}$ in \mathbb{R}^n . This bound can be optimized (in term of asymptotics along a subsequence of n) by taking primorial $k = p_1 p_2 \cdots p_m$ where $p_1 < p_2 < \cdots$ are the prime numbers. This gives the current best known bound:

Theorem 2.7.4 (Venkatesh 2012)

For infinitely many n , there exists a lattice sphere packing in \mathbb{R}^n of density

$$\geq (e^{-\gamma} - o(1))n \log \log n 2^{-n}.$$

Here $\gamma = 0.577 \dots$ is Euler's constant.

Open problem 2.7.5

Do there exist lattices (or sphere packings) in \mathbb{R}^n with sphere packing density $\geq (c + o(1))n$ for some constant $c > 1/2$?

3 Alterations

We saw the alterations method in Section 1.1 to give lower bounds to Ramsey numbers. The basic idea is to first make a random construction, and then fix some blemishes.

3.1 Dominating set in graphs

In a graph $G = (V, E)$, we say that $U \subseteq V$ is **dominating** if every vertex in $V \setminus U$ has a neighbor in U .

Theorem 3.1.1

Every graph on n vertices with minimum degree $\delta > 1$ has a dominating set of size

$$\leq \left(\frac{\log(\delta + 1) + 1}{\delta + 1} \right) n.$$

Naive attempt: take out vertices greedily. The first vertex eliminates $1 + \delta$ vertices, but subsequent vertices eliminate possibly fewer vertices.

Proof. Two-step process (alteration method):

1. Choose a random subset
2. Add enough vertices to make it dominating

Let $p \in [0, 1]$ to be decided later. Let X be a random subset of V where every vertex is included with probability p independently.

Let $Y = V \setminus (X \cup N(X))$. Each $v \in V$ lies in Y with probability $\leq (1 - p)^{1+\delta}$.

Then $X \cup Y$ is dominating, and

$$\mathbb{E}[|X \cup Y|] = \mathbb{E}[|X|] + \mathbb{E}[|Y|] \leq pn + (1 - p)^{1+\delta}n \leq (p + e^{-p(1+\delta)})n$$

using $1 + x \leq e^x$ for all $x \in \mathbb{R}$. Finally, setting $p = \frac{\log(\delta+1)}{\delta+1}$ to minimize $p + e^{-p(1+\delta)}$, we bound the above expression by

$$\leq \left(\frac{1 + \log(\delta + 1)}{\delta + 1} \right). \quad \square$$

3.2 Heilbronn triangle problem

Question 3.2.1

How can one place n points in the unit square so that no three points form a triangle with small area?

Let

$$\Delta(n) = \sup_{\substack{S \subseteq [0,1]^2 \\ |S|=n}} \min_{\substack{p,q,r \in S \\ \text{distinct}}} \text{area}(pqr)$$

Naive constructions fair poorly. E.g., n points around a circle has a triangle of area $\Theta(1/n^3)$ (the triangle formed by three consecutive points has side lengths $\asymp 1/n$ and angle $\theta = (1 - 1/n)2\pi$). Even worse is arranging points on a grid, as you would get triangles of zero area.

Heilbronn conjectured that $\Delta(n) = O(n^{-2})$.

Komlós, Pintz, and Szemerédi (1982) disproved the conjecture, showing $\Delta(n) \gtrsim n^{-2} \log n$. They used an elaborate probabilistic construction. Here we show a much simpler version probabilistic construction that gives a weaker bound $\Delta(n) \gtrsim n^{-2}$.

Remark 3.2.2. The currently best upper bound known is $\Delta(n) \leq n^{-8/7+o(1)}$ (Komlós, Pintz, and Szemerédi 1981)

Theorem 3.2.3 (Many points without small area triangles)

For every positive integer n , there exists a set of n points in $[0, 1]^2$ such that every triple spans a triangle of area $\geq cn^{-2}$, for some absolute constant $c > 0$.

Proof. Choose $2n$ points at random. For every three random points p, q, r , let us estimate

$$\mathbb{P}_{p,q,r}(\text{area}(p, q, r) \leq \varepsilon).$$

By considering the area of a circular annulus around p , with inner and outer radii x and $x + \Delta x$, we find



$$\mathbb{P}_{p,q}(|pq| \in [x, x + \Delta x]) \leq \pi((x + \Delta x)^2 - x^2)$$

So the probability density function satisfies

$$\mathbb{P}_{p,q}(|pq| \in [x, x + dx]) \leq 2\pi x dx$$

For fixed p, q

$$\mathbb{P}_r(\text{area}(pqr) \leq \varepsilon) = \mathbb{P}_r\left(\text{dist}(pq, r) \leq \frac{2\varepsilon}{|pq|}\right) \lesssim \frac{\varepsilon}{|pq|}$$

Thus, with p, q, r at random

$$\mathbb{P}_{p,q,r}(\text{area}(pqr) \leq \varepsilon) \lesssim \int_0^{\sqrt{2}} 2\pi x \frac{\varepsilon}{x} dx \asymp \varepsilon.$$

Given these $2n$ random points, let X be the number of triangles with area $\leq \varepsilon$. Then $\mathbb{E}X = O(\varepsilon n^3)$.

Choose $\varepsilon = c/n^2$ with $c > 0$ small enough so that $\mathbb{E}X \leq n$.

Delete a point from each triangle with area $\leq \varepsilon$.

The expected number of remaining points is $\mathbb{E}[2n - X] \geq n$, and no triangles with area $\leq \varepsilon = c/n^2$.

Thus with positive probability, we end up with $\geq n$ points and no triangle with area $\leq c/n^2$. \square

Algebraic construction. Here is another construction due to Erdős (in appendix of [Roth \(1951\)](#)) also giving $\Delta(n) \gtrsim n^{-2}$:

Let p be a prime. The set $\{(x, x^2) \in \mathbb{F}_p^2 : x \in \mathbb{F}_p\}$ has no 3 points collinear (a parabola meets every line in ≤ 2 points). Take the corresponding set of p points in $[p]^2 \subseteq \mathbb{Z}^2$. Then every triangle has area $\geq 1/2$ due to Pick's theorem. Scale back down to a unit square. (If n is not a prime, then use that there is a prime between n and $2n$.)

3.3 Markov's inequality

We note an important tool that will be used next.

Theorem 3.3.1 (Markov's inequality)

Let $X \geq 0$ be random variable. Then for every $a > 0$,

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

Proof. $\mathbb{E}[X] \geq \mathbb{E}[X1_{X \geq a}] \geq \mathbb{E}[a1_{X \geq a}] = a\mathbb{P}(X \geq a)$ \square

Take-home message: for r.v. $X \geq 0$, if $\mathbb{E}X$ is *very* small, then *typically* X is small.

3.4 High girth and high chromatic number

If a graph has a k -clique, then you know that its chromatic number is at least k .

Conversely, if a graph has high chromatic number, is it always possible to certify this fact from some “local information”?

Surprisingly, the answer is no. The following ingenious construction shows that a graph can be “locally tree-like” while still having high chromatic number.

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

Theorem 3.4.1 (Erdős 1959)

For all k, ℓ , there exists a graph with girth $> \ell$ and chromatic number $> k$.

Proof. Let $G \sim G(n, p)$ with $p = (\log n)^2/n$ (the proof works whenever $\log n/n \ll p \ll n^{-1+1/\ell}$). Here $G(n, p)$ is Erdős–Rényi random graph (n vertices, every edge appearing with probability p independently).

Let X be the number of cycles of length at most ℓ in G . By linearity of expectations, as there are exactly $\binom{n}{i}(i-1)!/2$ cycles of length i in K_n for each $3 \leq i \leq n$, we have (recall that ℓ is a constant)

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

By Markov’s inequality

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

(This allows us to get rid of all short cycles.)

How can we lower bound the chromatic number $\chi(\cdot)$? Note that $\chi(G) \geq |V(G)|/\alpha(G)$, where $\alpha(G)$ is the independence number (the size of the largest independent set).

With $x = (3/p) \log n$,

$$\mathbb{P}(\alpha(G) \geq x) \leq \binom{n}{x} (1-p)^{\binom{x}{2}} < n^x e^{-px(x-1)/2} = (ne^{-p(x-1)/2})^x = o(1).$$

Let n be large enough so that $\mathbb{P}(X \geq n/2) < 1/2$ and $\mathbb{P}(\alpha(G) \geq x) < 1/2$. Then there is some G with fewer than $n/2$ cycles of length $\leq \ell$ and with $\alpha(G) \leq (3/p) \log n$.

Remove a vertex from each cycle to get G' . Then $|V(G')| \geq n/2$, girth $> \ell$, and

$\alpha(G') \leq \alpha(G) \leq (3/p) \log n$, so

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

if n is sufficiently large. □

Remark 3.4.2. Erdős (1962) also showed that in fact one needs to see at least a linear number of vertices to deduce high chromatic number: for all k , there exists $\varepsilon = \varepsilon_k$ such that for all sufficiently large n there exists an n -vertex graph with chromatic number $> k$ but every subgraph on $\lfloor \varepsilon n \rfloor$ vertices is 3-colorable. (In fact, one can take $G \sim G(n, C/n)$; see "Probabilistic Lens: Local coloring" in Alon–Spencer)

3.5 Random greedy coloring

In Section 1.3, we saw a simple argument showing that every k -uniform hypergraph with than 2^{k-1} edges is 2-colorable (meaning that we can color the vertices red/blue without no monochromatic edge). Take a moment to remember the proof.

In this section, we improve this result. The next result gives the current best known bound.

Theorem 3.5.1 (Radhakrishnan and Srinivasan (2000))

There is some constant $c > 0$ so that every k -uniform hypergraph with at most $c \sqrt{\frac{k}{\log k}} 2^k$ edges is 2-colorable.

Recall from Section 1.3 that there exists a non-2-colorable k -uniform hypergraph on k^2 vertices and $O(k^2 2^k)$ edges, via a random construction.

Here we present a simpler proof, based on a **random greedy coloring**, due to Cherkashin and Kozik (2015), following an approach of Pluhaár (2009).

Proof. Consider a k -graph with m edges.

Let us order the vertices using a uniformly random chosen permutation.

Color vertices greedily from left to right: color a vertex blue unless it would create a monochromatic edge, in which case color it red (i.e., every red vertex is the final vertex in an edge with all earlier $k - 1$ vertices have already been colored blue).

The resulting coloring has no blue edges. The greedy coloring succeeds if it does not create a red edge.

3 Alterations

Analyzing a greedy coloring is tricky, since the color of a single vertex may depend on the entire history. Instead, we identify a specific feature that necessarily results from a unsuccessful coloring.

If there is a red edge, then there must be two edges e, f so that the last vertex of e is the first vertex of f . Call such pair (e, f) **conflicting** (note that whether (e, f) is conflicting depends on the random ordering of the vertices, but not on how we assigned colors).

What is the probability of seeing a conflicting pair? Here is the randomness comes from the random ordering of vertices.

Each pair of edges with exactly one vertex in common conflicts with probability $\frac{(k-1)!^2}{(2k-1)!} = \frac{1}{2k-1} \binom{2k-2}{k-1}^{-1} \asymp k^{-1/2} 2^{-2k}$. Summing over all $\leq m^2$ pairs of edges that share a unique vertex, we find that the expected number of conflicting pairs is at most $\lesssim m^2 k^{-1/2} 2^{-2k}$, which is < 1 for some $m \asymp k^{1/4} 2^k$. In this case, there is some ordering of vertices creating no conflicting pairs, in which case the greedy coloring always succeeds.

The above argument, due to [Pluhaár \(2009\)](#), yields $m \lesssim k^{1/4} 2^k$. Next we will refine the argument to obtain a better bound of $\sqrt{\frac{k}{\log k}} 2^k$ as claimed.

Instead of just considering a random permutation, let us map each vertex to $[0, 1]$ independently and uniformly at random. This map induces an ordering of the vertices, but it comes with further information that we will use.

Write $[0, 1] = L \cup M \cup R$ where (p to be decided)

$$L := \left[0, \frac{1-p}{2}\right), \quad M := \left[\frac{1-p}{2}, \frac{1+p}{2}\right], \quad R := \left(\frac{1+p}{2}, 1\right].$$

The probability that a given edge lands entirely in L is $\left(\frac{1-p}{2}\right)^k$, and likewise with R . Taking a union bound over all edges,

$$\mathbb{P}(\text{some edge lies in } L \text{ or } R) \leq 2m \left(\frac{1-p}{2}\right)^k.$$

Suppose that no edge of H lies entirely in L or entirely in R . If (e, f) conflicts, then their unique common vertex $x_v \in e \cap f$ must lie in M . So the probability that (e, f) conflicts is (here we use $x(1-x) \leq 1/4$)

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

Taking a union bound over all $\leq m^2$ pairs of edges, we find that

$$\mathbb{P}(\text{some conflicting pair has the common vertex in } M) \leq m^2 p 4^{-k+1}.$$

Thus

$\mathbb{P}(\text{there is a conflicting pair})$

$$\begin{aligned} &\leq \mathbb{P}(\text{some edge lies in } L \text{ or } R) + \mathbb{P}(\text{some conflicting pair has the common vertex in } M) \\ &\leq 2m \left(\frac{1-p}{2} \right)^k + m^2 p 4^{-k+1} \\ &< 2^{-k+1} m e^{-p^k} + (2^{-k+1} m)^2 p \end{aligned}$$

set $p = \log(2^{k-1} k / m) / k$ to minimize the right-hand side to get

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

which is < 1 for $m = c 2^k \sqrt{k / \log k}$ with $c > 0$ being a sufficiently small constant (we should assume that k is large enough to ensure $p \in [0, 1]$; smaller values of k can be handled in the theorem exceptionally by later reducing the constant c). \square

Food for thought: what is the source of the gain in the $L \cup M \cup R$ argument? The expected number of conflicting pairs is unchanged. It must be that we are somehow clustering the bad events by considering the event when some edge lies in L or R .

It remains an intriguing open problem to improve this bound further.

4 Second moment method

4.1 Does a typical random graph contain a triangle?

We begin with the following motivating question. Recall that the Erdős–Rényi random graph $G(n, p)$ is the n -vertex graph with edge probability p .

Question 4.1.1

For which $p = p_n$ does $G(n, p)$ contain a triangle with probability $1 - o(1)$?

(We sometimes abbreviate “with probability $1 - o(1)$ ” by “with high probability” or simply “whp”. In some literature, this is also called “asymptotically almost surely” or “a.a.s.”)

By computing $\mathbb{E}X$ (also known as the *first moment*), we deduce the following.

Proposition 4.1.2

If $np \rightarrow 0$, then $G(n, p)$ is triangle-free with probability $1 - o(1)$.

Proof. Let X be the number of triangles in $G(n, p)$. We know from linearity of expectations that

$$\mathbb{E}X = \binom{n}{3} p^3 \asymp n^3 p^3 = o(1).$$

Thus, by Markov’s inequality,

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

In other words, $X = 0$ with probability $1 - o(1)$. □

In other words, when $p \ll 1/n$, $G(n, p)$ is triangle-free with high probability (recall that $p \ll 1/n$ means $p = o(1/n)$; see asymptotic notation guide at the beginning of these notes).

What about when $p \gg 1/n$? Can we conclude that $G(n, p)$ contains a triangle with high probability? In this case $\mathbb{E}X \rightarrow \infty$, but this is not enough to conclude that

4 Second moment method

$\mathbb{P}(X \geq 1) = 1 - o(1)$, since we have not ruled out the probability that X is almost always zero but extremely large with some tiny probability.

An important technique in probabilistic combinatorics is to show that some random variable is *concentrated* around its mean. This would then imply that outliers are unlikely.

We will see many methods in this course on proving concentrations of random variables. In this chapter, we begin with the simplest method. It is usually easiest to execute and it requires not much hypotheses. The downside is that it only produces relatively weak (though still useful enough) concentration bounds.

Second moment method: show that a random variable is concentrated near its mean by bounding its variance.

Definition 4.1.3 (Variance)

The **variance** of a random variable X is

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

The **covariance** of two random variables X and Y (jointly distributed) is

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

(Exercise: check the second equality in the definitions of variance and covariance above).

Remark 4.1.4 (Notation convention). It is standard to use the Greek letter μ for the mean, and σ^2 for the variance. Here $\sigma \geq 0$ is the **standard deviation**.

The following basic result provides a concentration bound based on the variance.

Theorem 4.1.5 (Chebyshev's inequality)

Let X be a random variable with mean μ and variance σ^2 . For any $\lambda > 0$

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

Proof. By Markov's inequality,

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

Remark 4.1.6. Concentration bounds that show small probability of deviating from the mean are called **tail bounds** (more precisely: upper tail for $X \geq \mu + a$ and lower tail

4.1 Does a typical random graph contain a triangle?

for $\mathbb{P}(X \leq \mu - a)$). Chebyshev's inequality gives tail bounds that decays quadratically. Later on we will see tools that give much better decay (usually exponential) provided additional assumptions on the random variable (e.g., independence).

We are often interested in upper bounding the probability of non-existence, i.e., $\mathbb{P}(X = 0)$. Chebyshev's inequality yields the following bound.

Corollary 4.1.7 (Chebyshev bound on the probability of non-existence)

For any random variable X ,

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

Proof. By Chebyshev inequality, writing $\mu = \mathbb{E}X$,

$$\mathbb{P}(X = 0) \leq \mathbb{P}(|X - \mu| \geq |\mu|) \leq \frac{\text{Var } X}{\mu^2}. \quad \square$$

Corollary 4.1.8

If $\mathbb{E}X > 0$ and $\text{Var } X = o(\mathbb{E}X)^2$, then $X > 0$ and $X \sim \mathbb{E}X$ with probability $1 - o(1)$.

Remark 4.1.9 (Asymptotic statements). The above statement is really referring to not a single random variable, but a sequence of random variables X_n . It is saying that if $\mathbb{E}X_n > 0$ and $\text{Var } X_n = o(\mathbb{E}X_n)^2$, then $\mathbb{P}(X_n > 0) \rightarrow 1$ as $n \rightarrow \infty$, and for any fixed $\delta > 0$, $\mathbb{P}(|X_n - \mathbb{E}X_n| > \delta \mathbb{E}X_n) \rightarrow 0$ as $n \rightarrow \infty$.

In many situations, it is not too hard to compute the second moment. We have $\text{Var}[X] = \text{Cov}[X, X]$. Also, covariance is bilinear, i.e., for random variables X_1, \dots and Y_1, \dots (no assumptions needed on their independence, etc.) and constants a_1, \dots and b_1, \dots , one has

$$\text{Cov} \left[\sum_i a_i X_i, \sum_j b_j Y_j \right] = \sum_{i,j} a_i b_j \text{Cov}[X_i, Y_j].$$

We are often dealing with X being the cardinality of some random set. We can usually write this as a sum of indicator functions, such as $X = X_1 + \dots + X_n$, so that

$$\text{Var } X = \text{Cov}[X, X] = \sum_{i,j=1}^n \text{Cov}[X_i, X_j] = \sum_{i=1}^n \text{Var } X_i + 2 \sum_{i < j} \text{Cov}[X_i, X_j]$$

4 Second moment method

We have $\text{Cov}[X, Y] = 0$ if X and Y are independent. Thus in the sum we only need to consider dependent pairs (i, j) .

Example 4.1.10 (Sum of independent Bernoulli). Suppose $X = X_1 + \cdots + X_n$ with each X_i being an independent Bernoulli random variables with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$. Then $\mu = np$ and $\sigma^2 = np(1 - p)$ (note that $\text{Var}[X_i] = p - p^2$ and $\text{Cov}[X_i, X_j] = 0$ if $i \neq j$). If $np \rightarrow \infty$, then $\sigma = o(\mu)$, and thus $X = \mu + o(\mu)$ whp.

Note that the above computation remains identical even if we only knew that the X_i 's are *pairwise uncorrelated* (much weaker than assuming full independence).

Here the “tail probability” (the bound hidden in “whp”) decays polynomially in the deviation. Later on we will derive much sharper rates of decay (exponential) using more powerful tools such as the Chernoff bound when the r.v.'s are independent.

Let us now return to the problem of determining when $G(n, p)$ contains a triangle whp.

Theorem 4.1.11

If $np \rightarrow \infty$, then $G(n, p)$ contains a triangle with probability $1 - o(1)$.

Proof. Label the vertices by $[n]$. Let X_{ij} be the indicator random variable of the edge ij , so that $X_{ij} = 1$ if the edge is present, and $X_{ij} = 0$ if the edge is not present in the random graph. Let us write

$$X_{ijk} := X_{ij}X_{ik}X_{jk}.$$

Then the number of triangles in $G(n, p)$ is given by

$$X = \sum_{i < j < k} X_{ij}X_{ik}X_{jk}.$$

Now we compute $\text{Var } X$. Note that the summands of X are not all independent.

If T_1 and T_2 are each 3-vertex subsets, then

$$\begin{aligned} \text{Cov}[X_{T_1}, X_{T_2}] &= \mathbb{E}[X_{T_1}X_{T_2}] - \mathbb{E}[X_{T_1}]\mathbb{E}[X_{T_2}] = p^{e(T_1 \cup T_2)} - p^{e(T_1) + e(T_2)} \\ &= \begin{cases} 0 & \text{if } |T_1 \cap T_2| \leq 1 \\ p^5 - p^6 & \text{if } |T_1 \cap T_2| = 2 \\ p^3 - p^6 & \text{if } T_1 = T_2 \end{cases} \end{aligned}$$

4.1 Does a typical random graph contain a triangle?

Thus

$$\begin{aligned}\text{Var } X &= \sum_{T_1, T_2} \text{Cov}[X_{T_1}, X_{T_2}] = \binom{n}{3}(p^3 - p^6) + \binom{n}{2}n(n-1)(p^5 - p^6) \\ &\lesssim n^3 p^3 + n^4 p^5 = o(n^6 p^6) \quad \text{as } np \rightarrow \infty.\end{aligned}$$

Thus $\text{Var } X = o(\mathbb{E}X)^2$, and hence $X > 0$ whp by Corollary 4.1.8. \square

Here is what we have learned so far: for $p = p_n$ and as $n \rightarrow \infty$,

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

We say that $1/n$ is a **threshold** for containing a triangle, in the sense that if p grows asymptotically faster than this threshold, i.e., $p \gg 1/n$, then the event occurs with probability $1 - o(1)$, while if $p \ll 1/n$, then the event occurs with probability $o(1)$. Note that the definition of a threshold ignores leading constant factors (so that it is also correct to say that $2/n$ is also a threshold for containing a triangle). Determining the thresholds of various properties in random graphs (as well as other random settings) is a central topic in probabilistic combinatorics. We will discuss thresholds in more depth later in this chapter.

What else might you want to know about the probability that $G(n, p)$ contains a triangle?

Remark 4.1.12 (Poisson limit). What if $np \rightarrow c > 0$ for some constant $c > 0$? It turns out in this case that the number of triangles of $G(n, p)$ approaches a Poisson distribution with constant mean. You will show this in the homework. It will be done via the **method of moments**: if Z is some random variable with sufficiently nice properties (known as “determined by moments”, which holds for many common distributions such as the Poisson distribution and the normal distribution), and X_n is a sequence of random variables such that $\mathbb{E}X_n^k \rightarrow \mathbb{E}Z^k$ for all nonnegative integers k , then X_n converges in distribution to Z .

Remark 4.1.13 (Asymptotic normality). Suppose $np \rightarrow \infty$. From the above proof, we also deduce that $X \sim \mathbb{E}X$, i.e., the number of triangles is concentrated around its mean. In fact, we know much more. It turns out that the number X of triangles in $G(n, p)$ is asymptotically normal, meaning that it satisfies a central limit theorem: $(X - \mathbb{E}X)/\sqrt{\text{Var } X}$ converges in distribution to the standard normal $N(0, 1)$ in distribution. This was shown by [Rucinski \(1988\)](#) via the method of moments, by computing the k -th moment of $(X - \mathbb{E}X)/\sqrt{\text{Var } X}$ in the limit, and showing that it agrees with the k -th moment of the standard normal.

4 Second moment method

In the homework, you will prove the asymptotic normality of X using a later-found **method of projections**. The idea is to show that X is close to another random variable that is already known to be asymptotically normal by checking that their difference has negligible variance. For triangle counts, when $p \gg n^{-1/2}$, we can compare the number of triangles to the number of edges after a normalization. The method can be further modified for greater generality. See §6.4 in the book *Random Graphs* by Janson, Luczak, and Rucinski (2000).

Remark 4.1.14 (Better tail bounds). Later on we will use more powerful tools (including martingale methods and Azuma-Hoeffding inequalities, and also Janson inequalities) to prove better tail bounds on triangle (and other subgraph) counts.

4.2 Thresholds for fixed subgraphs

In the last section, we determined the threshold for $G(n, p)$ to contain a triangle. What about other subgraphs instead of a triangle? In this section, we give a complete answer to this question for any fixed subgraph.

Question 4.2.1

What is the threshold for containing a fixed H as a subgraph?

In other words, we wish to find a threshold q_n so that:

- (0-statement) if $p_n/q_n \rightarrow 0$, then $G(n, p_n)$ contains H with probability $o(1)$;
- (1-statement) if $p_n/q_n \rightarrow 1$, then $G(n, p_n)$ contains H with probability $1 - o(1)$.

(It is not a priori clear why a threshold exists in the first place. In fact, threshold always exist for monotone properties, as we will see in the next section.)

Building on our calculations for triangles from previous section, let us consider a more general setup for estimating the variance so that we can be more organized in our calculations.

Setup 4.2.2 (for variance bound with few dependencies)

Suppose $X = X_1 + \cdots + X_m$ where X_i is the indicator random variable for event A_i . Write $i \sim j$ if $i \neq j$ and the pair of events (A_i, A_j) are not independent. Define

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

Remark 4.2.3. (a) For many applications with an underlying symmetry between the events, the sum in the definition of Δ^* does not actually depend on i .

- (b) In the definition of the dependency graph ($i \sim j$) above, we are only considering pairwise dependence. Later on when we study the Lovász Local Lemma, we will need a strong notion of a dependency graph.
- (c) This method is appropriate for a collection of events with few dependencies. It is not appropriate for where there are many weak dependencies (e.g., Section 4.5 on the Hardy–Ramanujan theorem on the number of distinct prime divisors).

We have the bound

$$\text{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] \leq \mathbb{E}[X_i X_j] = \mathbb{P}[A_i A_j] = \mathbb{P}(A_i)\mathbb{P}(A_j|A_i).$$

(Here $A_i A_j$ is the shorthand for $A_i \wedge A_j$, meaning that both events occur.) Also

$$\text{Cov}[X_i, X_j] = 0 \quad \text{if } i \neq j \text{ and } i \not\sim j.$$

Thus

$$\begin{aligned} \text{Var } X &= \sum_{i,j=1}^m \text{Cov}[X_i, X_j] \leq \sum_{i=1}^m \mathbb{P}(A_i) + \sum_{i=1}^m \mathbb{P}(A_i) \sum_{j:j \sim i} \mathbb{P}(A_j|A_i) \\ &\leq \mathbb{E}X + (\mathbb{E}X)\Delta^*. \end{aligned}$$

Recall from Corollary 4.1.8 that $\mathbb{E}X > 0$ and $\text{Var } X = o(\mathbb{E}X)^2$ imply $X > 0$ and $X \sim \mathbb{E}X$ whp. So we have the following.

Lemma 4.2.4

In the above setup, if $\mathbb{E}X \rightarrow \infty$ and $\Delta^* = o(\mathbb{E}X)$, then $X > 0$ and $X \sim \mathbb{E}X$ whp.

Let us now determine the threshold for containing K_4 .

Theorem 4.2.5

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

Proof. Let X denote the number of copies of K_4 in $G(n, p)$. Then

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

If $p \ll n^{-2/3}$ then $\mathbb{E}X = o(1)$, and thus $X = 0$ whp by Markov's inequality.

Now suppose $p \gg n^{-2/3}$, so $\mathbb{E}X \rightarrow \infty$. For each 4-vertex subset S , let A_S be the event that S is a clique in $G(n, p)$.

4 Second moment method

For each fixed S , one has $A_S \sim A_{S'}$ if and only if $|S \cap S'| \geq 2$.

- The number of S' that share exactly 2 vertices with S is $6\binom{n-2}{2} = O(n^2)$, and for each such S' one has $\mathbb{P}(A_{S'}|A_S) = p^5$ (as there are 5 additional edges not in the S -clique that need to appear clique to form the S' -clique).
- The number of S' that share exactly 3 vertices with S is $4(n-4) = O(n)$, and for each such S' one has $\mathbb{P}(A_{S'}|A_S) = p^3$.

Summing over all above S' , we find

$$\Delta^* = \sum_{S': |S' \cap S| \in \{2,3\}} \mathbb{P}(A_{S'}|A_S) \lesssim n^2 p^5 + n p^3 \ll n^4 p^6 \asymp \mathbb{E}X.$$

Thus $X > 0$ whp by Lemma 4.2.4. □

For both K_3 and K_4 , we saw that any choice of $p = p_n$ with $\mathbb{E}X \rightarrow \infty$ one has $X > 0$ whp. Is this generally true?

Example 4.2.6 (First moment is not enough). Let $H = \text{---} \begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \bullet \end{array} \bullet$. We have $\mathbb{E}X_H \asymp n^5 p^7$. If $\mathbb{E}X = o(1)$ then $X = 0$ whp. But what if $\mathbb{E}X \rightarrow \infty$, i.e., $p \gg n^{-5/7}$?

We know that if $n^{-5/7} \ll p \ll n^{-2/3}$, then $X_{K_4} = 0$ whp, so $X_H = 0$ whp since $K_4 \subseteq H$.

On the other hand, if $p \gg n^{-2/3}$, then whp can find K_4 , and pick an arbitrary edge to extend to H (we'll prove this).

Thus the threshold for $H = \text{---} \begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \bullet \end{array} \bullet$ is actually $n^{-2/3}$, and not $n^{-5/7}$ as one might have naively predicted from the first moment alone.

Why didn't $\mathbb{E}X_H \rightarrow \infty$ give $X_H > 0$ whp in our proof strategy? In the calculation of Δ^* , one of the terms is $\asymp np$ (from two copies of H with a K_4 -overlap), and $np \not\ll n^5 p^7 \asymp \mathbb{E}X_H$ if $p \ll n^{-2/3}$.

The above example shows that the threshold is not always necessarily determined by the expectation. For the property of containing H , the example suggests that we should look at the “densest” subgraph of H rather than containing H itself.

Definition 4.2.7

Define the *edge-vertex ratio* of a graph H by

$$\rho(H) := \frac{e_H}{v_H}.$$

(This is the same as half the average degree.)

Define the *maximum edge-vertex ratio of a subgraph* of H :

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

Example 4.2.8. Let $H = \text{---} \bullet \begin{array}{c} \diagup \diagdown \\ \bullet \end{array} \bullet \text{---}$. We have $\rho(H) = 7/5$ whereas $\rho(K_4) = 3/2 > 7/5$. It is not hard to check that $m(H) = \rho(K_4) = 3/2$ as K_4 is the subgraph of H with the maximum edge-vertex ratio.

Remark 4.2.9 (Algorithm). Goldberg (1984) found a polynomial time algorithm for computing $m(H)$ via network flow algorithms.

The next theorem completely determines the threshold for containing some fixed graph H . Basically, it is determined by the expected number of copies of H' , where H' is the “denest” subgraph of H (i.e., with the maximum edge-vertex ratio).

Theorem 4.2.10 (Threshold for containing a fixed graph: Bollobás 1981)

Fix a graph H . Then $p = n^{-1/m(H)}$ is a threshold for containing H has a subgraph.

Proof. Let H' be a subgraph of H achieving the maximum edge-vertex ratio, i.e., $\rho(H') = m(H)$. Let X_H denote the number of copies of H in $G(n, p)$.

If $p \ll n^{-1/m(H)}$, then $\mathbb{E}X_{H'} \asymp n^{v_{H'}} p^{e_{H'}} = o(1)$, so $X_{H'} = 0$ whp, hence $X_H = 0$ whp.

Now suppose $p \gg n^{-1/m(H)}$. Let us count *labeled* copies of the subgraph H in $G(n, p)$. Let J be a labeled copy of H in K_n , and let A_J denote the event that J appears in $G(n, p)$. We have, for fixed J ,

$$\Delta^* = \sum_{J' \sim J} \mathbb{P}(A_{J'} \mid A_J) = \sum_{J' \sim J} p^{|E(J') \setminus E(J)|}$$

For any $J' \sim J$, we have

$$n^{|V(J') \setminus V(J)|} p^{|E(J') \setminus E(J)|} \ll n^{|V(J)|} p^{|E(J)|}$$

4 Second moment method

since

$$p \gg n^{-1/m(H)} \geq n^{-1/\rho(J \cap J')} = n^{-|V(J) \cap V(J')|/|E(J) \cap E(J')|}.$$

It then follows, after considering all the possible ways that J' can overlap with J , that $\Delta^* \ll n^{|V(J)|} p^{|E(J)|} \asymp \mathbb{E}X_H$. So Lemma 4.2.4 yields the result. \square

Remark 4.2.11. The proof also gives that if $p \gg n^{-1/m(H)}$, then the number X_H of copies of H is concentrated near its mean, i.e., with probability $1 - o(1)$,

$$X_H \sim \mathbb{E}X_H = \binom{n}{v_H} \frac{v_H!}{\text{aut}(H)} p^{e_H} \sim \frac{n^{v_H} p^{e_H}}{\text{aut}(H)}.$$

4.3 Thresholds

Previously, we computed the threshold for containing a fixed H as a subgraph. In this section, we take a detour from the discussion of the second moment method and discuss thresholds in more detail.

We begin by discussing the concept more abstractly by first defining the threshold of any monotone property on subsets. Then we show that thresholds always exist.

Thresholds form a central topic in probabilistic combinatorics. For any given property, it is natural to ask the following questions:

1. Where is the threshold?
2. Is the transition sharp? (And more precisely, what is width of the transition window?)

We understand thresholds well for many basic graph properties, but for many others, it can be a difficult problem. Also, one might think that one must first understand the location of the threshold before determining the nature of the phase transition, but surprisingly this is actually not always the case. There are powerful results that can sometimes show a sharp threshold without identifying the location of the threshold.

Here is some general setup, before specializing to graphs.

Let Ω be some finite set (ground set). Let Ω_p be a random subset of Ω where each element is included with probability p independently.

An **increasing property**, also called **monotone property**, on subsets of Ω is some binary property so that if $A \subseteq \Omega$ satisfies the property, any superset of A automatically satisfies the property.

A property is **trivial** if all subsets of Ω satisfy the property, or if all subsets of Ω do not satisfy the property. From now on, we only consider non-trivial monotone properties.

A **graph property** is a property that only depends on isomorphism classes of graphs. Whether the random graph $G(n, p)$ satisfies a given property can be cast in our setup by viewing $G(n, p)$ as Ω_p with $\Omega = \binom{[n]}{2}$.

Here are some examples of increasing properties for subgraphs of a given set of vertices:

- Contains some given subgraph
- Connected
- Has perfect matching
- Hamiltonian
- non-3-colorable

A family $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ of subsets of Ω is called an **up-set** if whenever $A \in \mathcal{F}$ and $A \subseteq B$, then $B \in \mathcal{F}$. Increasing property is the same as being an element of an up-set. We will use these two terms interchangeably.

Definition 4.3.1 (Threshold)

Let $\Omega = \Omega^{(n)}$ be a finite set and $\mathcal{F} = \mathcal{F}^{(n)}$ an monotone property of subsets of Ω . We say that q_n is a **threshold** for \mathcal{F} if,

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

Remark 4.3.2. The above definition is only for increasing properties. We can similarly define the threshold for decreasing properties by an obvious modification. An example of a non-monotone property is containing some H as an induced subgraph. Some (but not all) non-monotone properties also have thresholds, though we will not discuss it here.

Remark 4.3.3. From the definition, we see that if r_n and r'_n are both thresholds of the same property, then they must be within a constant factor of each other (exercise: check this). Thus it makes sense to say “the threshold” rather than “a threshold.”

Existence of threshold

Question 4.3.4 (Existence of threshold)

Does every non-trivial monotone property have a threshold?

4 Second moment method

How would a monotone property not have a threshold? Perhaps one could have $\mathbb{P}(\Omega_{1/n} \in \mathcal{F})$ and $\mathbb{P}(\Omega_{(\log n)/n} \in \mathcal{F}) \in [1/10, 9/10]$ for all sufficiently large n ?

Before answer this question, let us consider an even more elementary claim.

Theorem 4.3.5 (Monotonicity of satisfying probability)

Let Ω be a finite set and \mathcal{F} a non-trivial monotone property of subsets of Ω . Then $p \mapsto \mathbb{P}(\Omega_p \in \mathcal{F})$ is a strictly increasing function of $p \in [0, 1]$.

Let us give two related proofs of this basic fact. Both are quite instructive. Both are based on *coupling* of random processes.

Proof 1. Let $0 \leq p < q \leq 1$. Consider the following process to generate two random subsets of Ω . For each x , generate uniform $t_x \in [0, 1]$ independently at random. Let

$$A = \{x \in \Omega : t_x \leq p\} \quad \text{and} \quad B = \{x \in \Omega : t_x \leq q\}.$$

Then A has the same distribution as Ω_p and B has the same distribution as Ω_q . Furthermore, since $p < q$, we always have $A \subseteq B$. Since \mathcal{F} is monotone, $A \in \mathcal{F}$ implies $B \in \mathcal{F}$. Thus

$$\mathbb{P}(\Omega_p \in \mathcal{F}) = \mathbb{P}(A \in \mathcal{F}) \leq \mathbb{P}(B \in \mathcal{F}) = \mathbb{P}(\Omega_q \in \mathcal{F}).$$

To see that the inequality strict, we simply have to observe that with positive probability, one has $A \notin \mathcal{F}$ and $B \in \mathcal{F}$ (e.g., if all $t_x \in (p, q]$, then $A = \emptyset$ and $B = \Omega$). \square

In the second proof, the idea is to reveal a random subset of Ω in independent random stages.

Proof 2. (By two-round exposure) Let $0 \leq p < q \leq 1$. Note that $B = \Omega_q$ has the same distribution as the union of two independent $A = \Omega_p$ and $A' = \Omega_{p'}$, where p' is chosen to satisfy $1 - q = (1 - p)(1 - p')$ (check that the probability that each element occurs is the same in the two processes). Thus

$$\mathbb{P}(A \in \mathcal{F}) \leq \mathbb{P}(A \cup A' \in \mathcal{F}) = \mathbb{P}(B \in \mathcal{F}).$$

Like earlier, to observe that the inequality is strict, one observes that with positive probability, one has $A \notin \mathcal{F}$ and $A \cup A' \in \mathcal{F}$. \square

The above technique (generalized from two round exposure to multiple round exposures) gives a nice proof of the following theorem (originally proved using the Kruskal–Katona theorem).¹

¹(Thresholds for random subspaces of \mathbb{F}_q^n) The proof of the Bollobás–Thomason paper using the

Theorem 4.3.6 (Existence of thresholds: Bollobás and Thomason 1987)

Every sequence of nontrivial monotone properties has a threshold.

The theorem follows from the next non-asymptotic claim.

Lemma 4.3.7 (Multiple round exposure)

Let Ω be a finite set and \mathcal{F} some non-trivial monotone property. If $p \in [0, 1]$ and m is nonnegative integer. Then

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

Proof. Consider m independent copies of $\Omega_{p/m}$, and let Y be their union. Since \mathcal{F} is monotone increasing, if $Y \notin \mathcal{F}$, then none of the m copies lie in \mathcal{F} . Hence

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

Note that Y has the same distribution as Ω_q for some $q \leq p$. So $\mathbb{P}(\Omega_p \notin \mathcal{F}) \leq \mathbb{P}(\Omega_q \notin \mathcal{F}) = \mathbb{P}(Y \notin \mathcal{F})$ by Theorem 4.3.5. Combining the two inequalities gives the result. \square

Proof of Theorem 4.3.6. Since $p \mapsto \mathbb{P}(\Omega_p \in \mathcal{F})$ is a continuous strictly increasing function from 0 to 1 as p goes from 0 to 1 (in fact it is a polynomial in p), there is some unique “critical probability” p_c so that $\mathbb{P}(\Omega_{p_c} \in \mathcal{F}) = 1/2$.

It remains to check for every $\varepsilon > 0$, there is some $m = m(\varepsilon)$ (not depending on the property) so that

$$\mathbb{P}(\Omega_{p_c/m} \notin \mathcal{F}) \geq 1 - \varepsilon \quad \text{and} \quad \mathbb{P}(\Omega_{mp_c} \notin \mathcal{F}) \leq \varepsilon.$$

Kruskal–Katona theorem is still relevant. For example, there is an interesting analog of this concept for properties of subspaces of \mathbb{F}_q^n , i.e., random linear codes instead of random graphs. The analogue of the Bollobás–Thomason theorem was proved by [Rossman \(2020\)](#) via the the Kruskal–Katona approach. The multiple round exposure proof does not seem to work in the random subspace setting, as one cannot write a subspace as a union of independent copies of smaller subspaces.

As an aside, I disagree with the use of the term “sharp threshold” in Rossman’s paper for describing all thresholds for subspaces—one really should be looking at the cardinality of the subspaces rather than their dimensions. In a related work by [Guruswami, Mosheiff, Resch, Silas, and Wootters \(2022\)](#), they determine thresholds for random linear codes for properties that seem to be analogous to the property that a random graph contains a given fixed subgraph. Here again I disagree with them calling it a “sharp threshold.” It is much more like a coarse threshold once you parameterize by the cardinality of the subspace, which gives you a much better analogy to the random graph setting.

Thresholds for random linear codes seems to an interesting topic that has only recently been studied. I think there is more to be done here.

4 Second moment method

(here we write $\Omega_t = \Omega$ if $t > 1$) Indeed, applying Lemma 4.3.7 with $p = p_c$, we have

$$\mathbb{P}(\Omega_{p_c/m} \notin \mathcal{F}) \geq \mathbb{P}(\Omega_{p_c} \notin \mathcal{F})^{1/m} = 2^{-1/m} \geq 1 - \varepsilon \quad \text{if } m \geq (\log 2)/\varepsilon.$$

Applying Lemma 4.3.7 again with $p = mp_c$, we have

$$\mathbb{P}(\Omega_{mp_c} \notin \mathcal{F}) \leq \mathbb{P}(\Omega_{p_c} \notin \mathcal{F})^m = 2^{-m} \leq \varepsilon \quad \text{if } m \geq \log_2(1/\varepsilon).$$

Thus p_c is a threshold for \mathcal{F} . □

Examples

We will primarily be studying monotone graph properties. In the previous notation, $\Omega = \binom{[n]}{2}$, and we are only considering properties that depend on the isomorphism class of the graph.

Example 4.3.8 (Containing a triangle). We saw earlier in the chapter that the threshold for containing a triangle is $1/n$:

$$\mathbb{P}(G(n, p) \text{ contains a triangle}) \rightarrow \begin{cases} 0 & \text{if } np \rightarrow 0, \\ 1 - e^{-c^3/6} & \text{if } np \rightarrow c \in (0, \infty) \\ 1 & \text{if } np \rightarrow \infty. \end{cases}$$

In this case, the threshold is determined by the expected number of triangles $\Theta(n^3 p^3)$, and whether this quantity goes to zero or infinity (in the latter case, we used a second moment method to show that the number of triangles is positive with high probability).

What if $p = \Theta(1/n)$? If $np \rightarrow c$ for some constant $c > 0$, then (you will show in the homework via the method of moments) that the number of triangles is asymptotically Poisson distributed, and in particular the limit probability of containing a triangle increases from 0 to 1 as a continuous function of $c \in (0, \infty)$. So, in particular, as p increases, it goes through a “window of transition” of width $\Theta(1/n)$ in order for $\mathbb{P}(G(n, p) \text{ contains a triangle})$ to increase from 0.01 to 0.99. The width of this window is on the same order as the threshold. In this case, we call it a **coarse transition**.

Example 4.3.9 (Containing a subgraph). Theorem 4.2.10 tells us that the threshold for containing a fixed subgraph H is $n^{-1/m(H)}$. Here the threshold is not always determined by the expected number of copies of H . Instead, we need to look at the “densest subgraph” $H' \subseteq H$ with the largest edge-vertex ratio (i.e., equivalent to largest average degree). The threshold is determined by whether the expected number of copies of H' goes to zero or infinity.

Similar to the triangle case, we have a coarse threshold.

The analysis can also be generalized to containing one of several fixed subgraphs H_1, \dots, H_k .

Remark 4.3.10 (Monotone graph properties are characterized by subgraph containment). Every monotone graph property can be characterized as containing some element of \mathcal{H} for some \mathcal{H} that could depend on the vertex set n . For example, the property of connectivity corresponds to taking \mathcal{H} to be all spanning trees. More generally, one can take \mathcal{H} to be the set of all minimal graphs satisfying the property. When elements of \mathcal{H} are unbounded in size, the problem of thresholds become quite interesting and sometimes difficult.

The original Erdős–Rényi (1959) paper on random graphs already studied several thresholds, such as the next two examples.

Example 4.3.11 (No isolated vertices). With $p = \frac{\log n + c_n}{n}$,

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

It is a good exercise (and included in the problem set) to check the above claims. The cases $c_n \rightarrow -\infty$ and $c_n \rightarrow \infty$ can be shown using the second moment method. More precisely, when $c_n \rightarrow c$, by comparing moments one can show that the number of isolated vertices is asymptotically Poisson.

In this example, the threshold is at $(\log n)/n$. As we see above, the transition window is $\Theta(1/n)$, much narrower the magnitude of the threshold. In particular, the event probability goes from $o(1)$ to $1 - o(1)$ when p increases from $(1 - \delta)(\log n)/n$ to $(1 + \delta)(\log n)/n$ for any fixed $\delta > 0$. In this case, we say that the property has a **sharp threshold** at $(\log n)/n$ (here the leading constant factor is relevant, unlike the earlier example of a coarse threshold).

Example 4.3.12 (Connectivity). With $p = \frac{\log n + c_n}{n}$

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

In fact, a much stronger statement is true, connecting the above two examples: consider a process where one adds a random edges one at a time, then with probability $1 - o(1)$,

4 Second moment method

the graph becomes connected as soon as there are no more isolated vertices. Such stronger characterization is called a *hitting time* result.

A similar statement is true if we replace “is connected” by “has a perfect matching” (assuming n even).

Example 4.3.13 (Perfect matching in a random hypergraph: Shamir’s problem).

Let $G^{(3)}(n, p)$ be a random hypergraph on n vertices, where each triple of vertices appears as an edge with probability p . Assume that n is divisible by 3. What is the threshold for the existence of a perfect matching (i.e., a set of $n/3$ edges covering all vertices)?

It is easy to check that the property of having no isolated vertices has a sharp threshold at $p = 2n^{-2} \log n$. Is this also a threshold for having a perfect matching? So for smaller p , one cannot have a perfect matching due to having an isolated vertex. What about larger p ? This turns out to be a difficult problem known as “Shamir’s problem”.

A difficult result by [Johansson, Kahn, and Vu \(2008\)](#) (this paper won a Fulkerson Prize) showed that there is some constant $C > 0$ so that if $p \geq Cn^{-2} \log n$ then $G^{(3)}(n, p)$ contains a perfect matching with high probability. They also solved the problem much generally for H -factors in random k -uniform hypergraphs.

Recent exciting breakthroughs on the [Kahn–Kalai conjecture \(2007\)](#) by [Frankston, Kahn, Narayanan, and Park \(2021\)](#) and [Park and Pham \(2022+\)](#) provide new and much shorter proofs of this threshold for Shamir’s problem.

Recently, [Kahn \(2022\)](#) proved a sharp threshold result, and actually an even stronger hitting time version, of Shamir’s problem, showing that with high probability, one has a perfect matching as soon as there are no isolated vertices.

Sharp transition

In some of the examples, the probability that $G(n, p)$ satisfies the property changes quickly and dramatically as p crosses the threshold (physical analogy: similar to how the structure of water changes dramatically as the temperature drops below freezing). For example, while for connectivity, while $p = \log n/n$ is a threshold, we see that $G(n, 0.99 \log n/n)$ is whp not connected and $G(n, 1.01 \log n/n)$ is whp connected, unlike the situation for containing a triangle earlier. We call this the *sharp threshold phenomenon*.

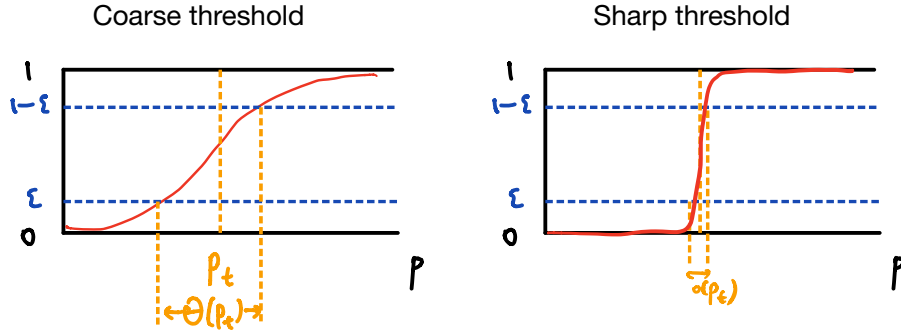


Figure 4.1: Examples of coarse and sharp thresholds. The vertical axis is the probability that $G(n, p)$ satisfies the property.

Definition 4.3.14 (Sharp thresholds)

We say that r_n is a **sharp threshold** for some property \mathcal{F} on subsets of Ω if, for every $\delta > 0$,

$$\mathbb{P}(\Omega_{p_n} \in \mathcal{F}) \rightarrow \begin{cases} 0 & \text{if } p_n/r_n \leq 1 - \delta, \\ 1 & \text{if } p_n/r_n \geq 1 + \delta. \end{cases}$$

On the other hand, if there is some fixed $\varepsilon > 0$ and $0 < c < C$ so that $\mathbb{P}(\Omega_{p_n} \in \mathcal{F}) \in [\varepsilon, 1 - \varepsilon]$ for whenever $c \leq p_n/r_n \leq C$, then we say that r_n is a **coarse threshold**.

As in Figure 4.1, the sharp/coarseness of a thresholds is about how quickly $\mathbb{P}(\Omega_p \in \mathcal{F})$ transitions from ε to $1 - \varepsilon$ as p increases. How wide is the transition window for p ? By the Bollobás–Thomason theorem (Theorem 4.3.6) on the existence of thresholds, this transition window always has width $O(r_n)$. If the transition window has width $\Theta(r_n)$ for some $\varepsilon > 0$, then we have a coarse threshold. On the other hand, if the transition window has width $o(r_n)$ for every $\varepsilon > 0$, then we have a sharp threshold.

From earlier examples, we saw coarse thresholds for the “local” property of containing some given subgraph, as well as sharp thresholds for “global” properties such as connectivity. It turns out that this is a general phenomenon.

Friedgut’s sharp threshold theorem (1999), a deep and important result, completely characterizes when a threshold is coarse versus sharp. We will not state Friedgut’s theorem precisely here since it is rather technical (and actually not always easy to apply). Let us just give a flavor. Roughly speaking, the theorem says that:

All monotone graph properties with a coarse threshold may be approximated by a local property.

In other words, informally, if a monotone graph property \mathcal{P} has a coarse threshold, then there is finite list of graph G_1, \dots, G_m such that \mathcal{P} is “close to” the property of containing one of G_1, \dots, G_m as a subgraph.

4 Second moment method

We need “close to” since the property could be “contains a triangle and has at least $\log n$ edges”, which is not exactly local but it is basically the same as “contains a triangle.”

There is some subtlety here since we can allow very different properties depending on the value of n . E.g., \mathcal{P} could be the set of all n -vertex graphs that contain a K_3 if n is odd and K_4 if n is even. Friedgut’s theorem tells us that if there is a threshold, then there is a partition $\mathbb{N} = \mathbb{N}_1 \cup \dots \cup \mathbb{N}_k$ such that on each \mathbb{N}_i , \mathcal{P} is approximately the form described in the previous paragraph.

In the last section, we derived that the property of containing some fixed H has threshold $n^{-1/m(H)}$ for some rational number $m(H)$. It follows as a corollary of Friedgut’s theorem that every coarse threshold must have this form.

Corollary 4.3.15 (of Friedgut’s sharp threshold theorem)

Suppose $r(n)$ is a coarse threshold of some graph property. Then there is a partition of $\mathbb{N} = \mathbb{N}_1 \cup \dots \cup \mathbb{N}_k$ and rationals $\alpha_1, \dots, \alpha_k > 0$ such that $r(n) \asymp n^{-\alpha_j}$ for every $n \in \mathbb{N}_j$.

In particular, if $(\log n)/n$ is a threshold of some monotone graph property (e.g., this is the case for connectivity), then we automatically know that it must be a sharp threshold, even without knowing anything else about the property. Likewise if the threshold has the form $n^{-\alpha}$ for some irrational α .

The exact statement of Friedgut’s theorem is more cumbersome. We refer those who are interested to Friedgut’s original [1999 paper](#) and his later [survey](#) for details and applications. This topic is connected more generally to an area known as the *analysis of boolean functions*.

Also, it is known that the transition window of every monotone graph property is $(\log n)^{-2+o(1)}$ (Friedgut—Kalai (1996), Bourgain—Kalai (1997)).

Curiously, tools such as Friedgut’s theorem sometimes allow us to prove the existence of a sharp threshold without being able to identify its exact location. For example, it is an important open problem to understand where exactly is the transition for a random graph to be k -colorable.

Conjecture 4.3.16 (k -colorability threshold)

For every $k \geq 3$ there is some real constant $d_k > 0$ such that for any constant $d > 0$,

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

We do know that there *exists* a sharp threshold for k -colorability.

Theorem 4.3.17 (Achlioptas and Friedgut 2000)

For every $k \geq 3$, there exists a function $d_k(n)$ such that for every $\varepsilon > 0$, and sequence $d(n) > 0$,

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

On the other hand, it is not known whether $\lim_{n \rightarrow \infty} d_k(n)$ exists, which would imply Conjecture 4.3.16. Further bounds on $d_k(n)$ are known, e.g. the landmark paper of [Achlioptas and Naor \(2006\)](#) showing that for each fixed $d > 0$, whp $\chi(G(n, d/n)) \in \{k_d, k_d + 1\}$ where $k_d = \min\{k \in \mathbb{N} : 2k \log k > d\}$. Also see the later work of [Coja-Oghlan and Vilenchik \(2013\)](#).

4.4 Clique number of a random graph

The **clique number** $\omega(G)$ of a graph is the maximum number of vertices in a clique of G .

Question 4.4.1

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

Let X be the number of k -cliques of $G(n, 1/2)$. Define

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

Let us first do a rough estimate to see what is the critical k to get $f(n, k)$ large or small. Recall that $\left(\frac{n}{ek}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. We have

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

And so the transition point is at some $k \sim 2 \log_2 n$ in the sense that if $k \geq (2 + \delta) \log_2 n$, then $f(n, k) \rightarrow 0$ while if $k \leq (2 - \delta) \log_2 n$, then $f(n, k) \rightarrow \infty$.

The next result gives us a lower bound on the typical clique number.

Theorem 4.4.2 (Second moment bound for clique number)

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

- (a) If $f(n, k) \rightarrow 0$, then $\omega(G(n, 1/2)) < k$ whp.
- (b) If $f(n, k) \rightarrow \infty$, then $\omega(G(n, 1/2)) \geq k$ whp.

Proof sketch. The first claim follows from Markov's inequality as $\mathbb{P}(X \geq 1) \leq \mathbb{E}X$.

For the second claim, we bound the variance using Setup 4.2.2. For each k -element subset S of vertices, let A_S be the event that S is a clique. Let X_S be the indicator random variable for A_S . Recall

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

For fixed k -set S , consider all k -set T with $|S \cap T| \geq 2$:

$$\Delta^* = \sum_{\substack{T \in \binom{[n]}{k} \\ 2 \leq |S \cap T| \leq k-1}} \mathbb{P}(A_T \mid A_S) = \sum_{i=2}^{k-1} \binom{k}{i} \binom{n-k}{k-i} 2^{\binom{i}{2} - \binom{k}{2}} \overset{\text{omitted}}{\ll} \mathbb{E}X = \binom{n}{k} 2^{-\binom{k}{2}}.$$

It then follows from Lemma 4.2.4 that $X > 0$ (i.e., $\omega(G) \geq k$) whp. □

We can a two-point concentration result for the clique number of $G(n, 1/2)$. This result is due to [Bollobás–Erdős 1976](#) and [Matula 1976](#).

Theorem 4.4.3 (Two-point concentration for clique number)

There exists a $k = k(n) \sim 2 \log_2 n$ such that $\omega(G(n, 1/2)) \in \{k, k+1\}$ whp.

Proof. For $k \sim 2 \log_2 n$,

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

Let $k_0 = k_0(n) \sim 2 \log_2 n$ be the value with

$$f(n, k_0) \geq n^{-1/2} > f(n, k_0 + 1).$$

Then $f(n, k_0 - 1) \rightarrow \infty$ and $f(n, k_0 + 1) = o(1)$. By Theorem 4.4.2, the clique number of $G(n, 1/2)$ is whp in $\{k_0 - 1, k_0\}$. □

Remark 4.4.4. By a more careful analysis, one can show that outside a very sparse

4.5 Hardy–Ramanujan theorem on the number of prime divisors

subset of integers, one has $f(n, k_0) \rightarrow \infty$, in which case one has one-point concentration $\omega(G(n, 1/2)) = k_0$ whp.

By taking the complement of the graph, we also get a two-point concentration result about the independence number of $G(n, 1/2)$. [Bohman and Hofstad \(2022+\)](#) extended the two-point concentration result for the independence number of $G(n, p)$ to all $p \geq n^{-2/3+\varepsilon}$.

Remark 4.4.5 (Chromatic number). Since the chromatic number satisfies $\chi(G) \geq n/\alpha(G)$, we have

$$\chi(G(n, 1/2)) \geq (1 + o(1)) \frac{n}{2 \log_2 n} \quad \text{whp.}$$

Later on, using more advanced methods, we will prove $\chi(G(n, 1/2)) \sim n/(2 \log_2 n)$ whp ([Bollobás 1987](#)).

Also, later, using martingale concentration, we can also show that $\chi(G(n, p))$ is tightly concentrated around its mean without a priori needing to know where the mean is located.

4.5 Hardy–Ramanujan theorem on the number of prime divisors

Let $\nu(n)$ denote the number of distinct primes dividing n (not counting multiplicities).

The next theorem says that “almost all” n have $(1 + o(1)) \log \log n$ prime factors

Theorem 4.5.1 (Hardy and Ramanujan 1917)

For every $\varepsilon > 0$, there exists C such that for all sufficiently large n , all but ε -fraction of $x \in [n]$ satisfy

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

The original proof of Hardy and Ramanujan was quite involved. Here we show a “probabilistic” proof due to [Turán \(1934\)](#), which played a key role in the development of probabilistic methods in number theory.

Proof. Choose $x \in [n]$ uniformly at random. For prime p , let

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

4 Second moment method

Set $M = n^{1/10}$, and (the sum is taken over primes p).

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

We have

$$\nu(x) - 10 \leq X(x) \leq \nu(x)$$

since x cannot have more than 10 prime factors $> n^{1/10}$. So it suffices to analyze X . Since exactly $\lfloor n/p \rfloor$ positive integers $\leq n$ are divisible by p , we have

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

We apply [Merten's theorem](#) from analytic number theory:

$$\sum_{p \leq n} 1/p = \log \log n + O(1)$$

(the $O(1)$ error term converges to the Meissel–Mertens constant). So

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

Next we compute the variance. The intuition is that divisibility by distinct primes should behave somewhat independently. Indeed, if pq divides n , then X_p and X_q are independent (e.g., by the Chinese Remainder Theorem). If pq does not divide n , but n is large enough, then there is some small covariance contribution. (In contrast to the earlier variance calculations in random graphs, here we have many weak dependices.)

If $p \neq q$, then $X_p X_q = 1$ if and only if $pq|x$. Thus

$$\begin{aligned} |\text{Cov}[X_p, X_q]| &= |\mathbb{E}[X_p X_q] - \mathbb{E}[X_p] \mathbb{E}[X_q]| \\ &= \left| \frac{\lfloor n/pq \rfloor}{n} - \frac{\lfloor n/p \rfloor}{n} \frac{\lfloor n/q \rfloor}{n} \right| \\ &= \left| \frac{1}{pq} + O\left(\frac{1}{n}\right) - \left(\frac{1}{p} + O\left(\frac{1}{n}\right) \right) \left(\frac{1}{q} + O\left(\frac{1}{n}\right) \right) \right| \\ &= O\left(\frac{1}{n}\right). \end{aligned}$$

Thus

$$\sum_{p \neq q} |\text{Cov}[X_p, X_q]| \lesssim \frac{M^2}{n} \lesssim n^{-4/5}.$$

4.5 Hardy–Ramanujan theorem on the number of prime divisors

Also, $\text{Var } X_p = \mathbb{E}[X_p] - (\mathbb{E}X_p)^2 = (1/p)(1 - 1/p) + O(1/n)$. Combining, we have

$$\begin{aligned}\text{Var } X &= \sum_{p \leq M} \text{Var } X_p + \sum_{p \neq q} \text{Cov}[X_p, X_q] \\ &= \sum_{p \leq M} \frac{1}{p} + O(1) = \log \log n + O(1) \sim \mathbb{E}X.\end{aligned}$$

Thus by Chebyshev, for every constant $\lambda > 0$

$$\mathbb{P}\left(|X - \log \log n| \geq \lambda \sqrt{\log \log n}\right) \leq \frac{\text{Var } X}{\lambda^2 (\log \log n)} = \frac{1}{\lambda^2} + o(1).$$

Finally, recall that $|X - \nu| \leq 10$, so same asymptotic bound holds with X replaced by ν . \square

The main idea from the above proof is that the number of prime divisors $X = \sum_p X_p$ (from the previous proof) behaves like a sum of independent random variables.

We have the following corollary of the Lindenberg–Feller central limit theorem (see [Durrett, Theorem 3.4.10](#)):

Theorem 4.5.2 (CLT for sums of independent Bernoullis)

If X_n is a sum of independent Bernoulli random variables, and $\text{Var } X_n \rightarrow \infty$ as $n \rightarrow \infty$, then $(X_n - \mathbb{E}X_n)/\sqrt{\text{Var } X}$ converges to the normal distribution.

(Note that the divergent variance hypothesis is necessary and sufficient.)

So it is natural to expect $\nu(x)$ to satisfy a central limit theorem. This is indeed the case, and can be proved by comparing moments.

Theorem 4.5.3 (Asymptotic normality: Erdős and Kac 1940)

With $x \in [n]$ uniformly chosen at random, $\nu(x)$ is asymptotically normal, i.e., for every $\lambda \in \mathbb{R}$,

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

The original proof of Erdős and Kac verifies the above intuition using some more involved results in analytic number theory. Simpler proofs have been subsequently given, and we outline one below, which is based on computing the moments of the distribution. The idea of computing moments for this problem was first used by [Delange \(1953\)](#), who was apparently not aware of the Erdős–Kacs paper. Also see a more modern account by [Granville and Soundararajan \(2007\)](#).

4 Second moment method

The following tool from probability theory allows us to verify asymptotic normality from convergence of moments.

Theorem 4.5.4 (Method of moments)

Let X_n be a sequence of real valued random variables such that for every positive integer k , $\lim_{n \rightarrow \infty} \mathbb{E}[X_n^k]$ equals to the k -th moment of the standard normal distribution. Then X_n converges in distribution to the standard normal, i.e., $\lim_{n \rightarrow \infty} \mathbb{P}(X_n \leq a) = \mathbb{P}(Z \leq a)$ for every $a \in \mathbb{R}$, where Z is a standard normal.

Remark 4.5.5. The same conclusion holds for any probability distribution that is “determined by its moments,” i.e., there are no other distributions sharing the same moments. Many common distributions that arise in practice, e.g., the Poisson distribution, satisfy this property. There are various sufficient conditions for guaranteeing this moments property, e.g., Carleman’s condition tells us that any probability distribution whose moments do not increase too quickly is determined by its moments. (See Durrett §3.3.5).

Proof of Erdős–Kacs Theorem 4.5.3. We compare higher moments of $X = \nu(x)$ with that of an idealized Y treating the prime divisors as truly random variables.

Set $M = n^{1/s(n)}$ where $s(n) \rightarrow \infty$ sufficiently slowly. As earlier, $\nu(x) - s(n) \leq \nu(x) \leq \nu(x)$.

We construct a “model random variable” mimicking X . Let $Y = \sum_{p \leq M} Y_p$, where $Y_p \sim \text{Bernoulli}(1/p)$ independently for all primes $p \leq M$. We can compute:

$$\mu := \mathbb{E}Y \sim \mathbb{E}X \sim \log \log n$$

and

$$\sigma^2 := \text{Var } Y \sim \text{Var } X \sim \log \log n.$$

Let $\tilde{X} = (X - \mu)/\sigma$ and $\tilde{Y} = (Y - \mu)/\sigma$.

A consequence of the Lindeberg–Feller central limit theorem is that a sum of independent Bernoulli random variables with divergent variance satisfies the central limit theorem. So $\tilde{Y} \rightarrow N(0, 1)$ in distribution. In particular, $\mathbb{E}[\tilde{Y}^k] \sim \mathbb{E}[Z^k]$ (asymptotics as $n \rightarrow \infty$) where Z is a standard normal.

Let us compare \tilde{X} and \tilde{Y} . It suffices to show that for every fixed k , $\mathbb{E}[\tilde{X}^k] \sim \mathbb{E}[\tilde{Y}^k]$.

For every set of distinct primes $p_1, \dots, p_r \leq M$,

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

Comparing expansions of \tilde{X}^k in terms of the X_p 's ($n^{o(1)}$ terms), we get

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

It follows that \tilde{X} is asymptotically normal. \square

4.6 Distinct sums

What is the largest subset of $[n]$ all of whose subsets have distinct sums? Equivalently:

Question 4.6.1

For each k , what is the smallest n so that there exists $S \subseteq [n]$ with $|S| = k$ and all 2^k subset sums of S are distinct?

E.g., $S = \{1, 2, 2^2, \dots, 2^{k-1}\}$ (the greedy choice).

We begin with an easy pigeonhole argument. Since all 2^k sums are distinct and are at most kn , we have $2^k \leq kn$. Thus $n \geq 2^k/k$.

Erdős offered \$300 for a proof or disproof of the following. It remains open.

Conjecture 4.6.2 (Erdős)

$$n \gtrsim 2^k$$

Let us use the second moment to give a modest improvement on the earlier pigeonhole argument. The main idea here is that, by second moment, most of the subset sums lie within an $O(\sigma)$ -interval, so that we can improve on the pigeonhole estimate ignoring outlier subset sums.

Theorem 4.6.3

If there is a k -element subset of $[n]$ with distinct subset sums. Then $n \gtrsim 2^k/\sqrt{k}$.

Proof. Let $S = \{x_1, \dots, x_k\}$ be a k -element subset of $[n]$ with distinct subset sums. Set

$$X = \varepsilon_1 x_1 + \dots + \varepsilon_k x_k$$

where $\varepsilon_i \in \{0, 1\}$ are chosen uniformly at random independently. We have

$$\mu := \mathbb{E}X = \frac{x_1 + \dots + x_k}{2}$$

and

$$\sigma^2 := \text{Var } X = \frac{x_1^2 + \dots + x_k^2}{4} \leq \frac{n^2 k}{4}.$$

4 Second moment method

By Chebyshev's inequality,

$$\mathbb{P}(|X - \mu| \geq 2\sigma) \leq \frac{1}{4},$$

and thus

$$\mathbb{P}(|X - \mu| < n\sqrt{k}) = \mathbb{P}(|X - \mu| < 2\sigma) \geq \frac{3}{4}.$$

Since X takes distinct values for every $(\varepsilon_1, \dots, \varepsilon_k) \in \{0, 1\}^k$, we have $\mathbb{P}(X = x) \leq 2^{-k}$ for all x . Since there are $\leq 2n\sqrt{k}$ elements in the interval $(\mu - n\sqrt{k}, \mu + n\sqrt{k})$, we have

$$\mathbb{P}(|X - \mu| < n\sqrt{k}) \leq 2n\sqrt{k}2^{-k}.$$

Putting the upper and lower bounds on $\mathbb{P}(|X - \mu| < n\sqrt{k})$ together, we get

$$2n\sqrt{k}2^{-k} \leq \frac{3}{4}.$$

So $n \gtrsim 2^k / \sqrt{k}$. □

Dubroff, Fox, and Xu (2021) gave another short proof of this result by applying Harper's vertex-isoperimetric inequality on the cube (this is an example of “concentration of measure”, which we will explore more later this course).

Consider the graph representing the n -dimensional boolean cube, with vertex set $\{0, 1\}^n$ with an edge between every pair of n -tuples that differ in exactly one coordinate. Given $A \subseteq \{0, 1\}^n$, write ∂A for the set of all vertices outside A that is adjacent to some vertex of A .

Theorem 4.6.4 (Vertex-isoperimetric inequality on the hypercube: Harper 1966)

Every $A \subseteq \{0, 1\}^k$ with $|A| = 2^{k-1}$ has $|\partial A| \geq \binom{k}{\lfloor k/2 \rfloor}$.

Remark 4.6.5. A stronger form of Harper's theorem gives the precise value of

$$\min_{A \subseteq \{0,1\}^n: |A|=m} |\partial A|$$

for every (n, m) . Basically, the minimum is achieved when A is a Hamming ball (or, if m is not exactly the size of some Hamming ball, then take the first m elements of $\{0, 1\}^n$ when ordered lexicographically).

Theorem 4.6.6 (Dubroff–Fox–Xu 2021)

If there is a k -element subset of $[n]$ with distinct subset sums, then

$$n \geq \binom{k}{\lfloor k/2 \rfloor} = \left(\sqrt{\frac{2}{\pi}} + o(1) \right) \frac{2^k}{\sqrt{k}}.$$

Remark 4.6.7. The above bound has the currently best known leading constant factor, matching an earlier result by Aliev (2009).

Proof. Let $S = \{x_1, \dots, x_k\}$ be a subset of $[n]$ with distinct sums. Let

$$A = \left\{ (\varepsilon_1, \dots, \varepsilon_k) \in \{0, 1\}^k : \varepsilon_1 x_1 + \dots + \varepsilon_k x_k < \frac{x_1 + \dots + x_k}{2} \right\}.$$

Note that due to the distinct sum hypothesis, one can never have $\varepsilon_1 x_1 + \dots + \varepsilon_k x_k = (x_1 + \dots + x_k)/2$. It thus follows by symmetry that $|A| = 2^{k-1}$.

Note that every element of ∂A corresponds to some sum of the form $z + x_i > (x_1 + \dots + x_k)/2$ for some $z < (x_1 + \dots + x_k)/2$, and thus $z + x_i$ lies in the open interval

$$\left(\frac{x_1 + \dots + x_k}{2}, \frac{x_1 + \dots + x_k}{2} + \max S \right).$$

Since all subset sums are distinct, we must have $n \geq |\partial A| \geq \binom{k}{\lfloor k/2 \rfloor}$ by Harper's theorem (Theorem 4.6.4). \square

4.7 Weierstrass approximation theorem

We finish off the chapter with an application to analysis.

The Weierstrass approximation theorem says that every continuous real function on an interval can be uniformly approximated by a polynomial.

Theorem 4.7.1 (Weierstrass approximation theorem 1885)

Let $f: [0, 1] \rightarrow \mathbb{R}$ be a continuous function. Let $\varepsilon > 0$. Then there is a polynomial $p(x)$ such that $|p(x) - f(x)| \leq \varepsilon$ for all $x \in [0, 1]$.

Proof. (Bernstein 1912) The idea is to approximate f by a sum of polynomials that look like “bumps”:

$$P_n(x) = \sum_{i=0}^n E_i(x) f(i/n)$$

4 Second moment method

where

$$E_i(x) = \mathbb{P}(\text{Binomial}(n, x) = i) = \binom{n}{i} x^i (1-x)^{n-i} \quad \text{for } 0 \leq i \leq n$$

is chosen as some polynomials peaks at $x = i/n$ and then decays away from $x = i/n$.

For each $x \in [0, 1]$, the binomial distribution $\text{Binomial}(n, x)$ has mean nx and variance $nx(1-x) \leq n$. By Chebyshev's inequality,

$$\sum_{i: |i-nx| > n^{2/3}} E_i(x) = \mathbb{P}(|\text{Binomial}(n, x) - nx| > n^{2/3}) \leq n^{-1/3}.$$

(In the next chapter, we will see a much better tail bound.)

Since $[0, 1]$ is compact, f is uniformly continuous and bounded. By multiplying by a constant, we assume that $|f(x)| \leq 1$ for all $x \in [0, 1]$. Also there exists $\delta > 0$ such that $|f(x) - f(y)| \leq \varepsilon/2$ for all $x, y \in [0, 1]$ with $|x - y| \leq \delta$.

Take $n > \max\{64\varepsilon^{-3}, \delta^{-3}\}$. Then for every $x \in [0, 1]$ (note that $\sum_{j=0}^n E_j(x) = 1$),

$$\begin{aligned} |P_n(x) - f(x)| &\leq \sum_{i=0}^n E_i(x) |f(i/n) - f(x)| \\ &\leq \sum_{i: |i/n-x| < n^{-1/3} < \delta} E_i(x) |f(i/n) - f(x)| + \sum_{i: |i-nx| > n^{2/3}} 2E_i(x) \\ &\leq \frac{\varepsilon}{2} + 2n^{-1/3} \leq \varepsilon. \end{aligned}$$

□

5 Chernoff bound

Chernoff bounds give us much better tail bounds than the second moment method when applied to sums of independent random variables. This is one of the most useful bounds in probabilistic combinatorics.

The proof technique of bounding the exponential moments is perhaps just as important as the resulting bounds themselves. We will see this proof method come up again later on when we prove martingale concentration inequalities. The method allows us to adapt the proof of the Chernoff bound to other distributions. Let us give the proof in the most basic case for simplicity and clarity.

Theorem 5.0.1 (Chernoff bound)

Let $S_n = X_1 + \dots + X_n$ where $X_i \in \{-1, 1\}$ uniformly iid. Let $\lambda > 0$. Then

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

In contrast, Chebyshev's inequality gives $\mathbb{P}(S_n \geq \lambda\sqrt{n}) \leq 1/\lambda^2$

Proof. Let $t \geq 0$. Consider the **moment generating function**

$$\mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{t \sum_i X_i}] = \mathbb{E}\left[\prod_i e^{tX_i}\right] = \prod_i \mathbb{E}[e^{tX_i}] = \left(\frac{e^{-t} + e^t}{2}\right)^n.$$

By comparing Taylor series,

$$\frac{e^{-t} + e^t}{2} = \sum_{k \geq 0} \frac{x^{2k}}{(2k)!} \leq \sum_{k \geq 0} \frac{x^{2k}}{k!2^k} = e^{t^2/2}.$$

By Markov's inequality,

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

Setting $t = \lambda/\sqrt{n}$ gives the bound. □

Remark 5.0.2. The technique of considering the moment generating function can

5 Chernoff bound

be thought morally as taking an appropriately high moment. Indeed, $\mathbb{E}[e^{tS}] = \sum_{n \geq 0} \mathbb{E}[S^n] t^n / n!$ contains all the moments data of the random variable.

The second moment method (Chebyshev + Markov) can be thought of as the first iteration of this idea. By taking fourth moments (now requiring 4-wise independence of the summands), we can obtain tail bounds of the form $\lesssim \lambda^{-4}$. And similarly with higher moments.

In some applications, where one cannot assume independence, but can estimate some high moments, the above philosophy can allow us to prove good tail bounds as well.

Also by symmetry, $\mathbb{P}(S_n \leq -\lambda\sqrt{n}) \leq e^{-\lambda^2/2}$. Thus we have the following two-sided tail bound.

Corollary 5.0.3

With S_n as before, for any $\lambda \geq 0$,

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

Remark 5.0.4. It is easy to adapt the above proof so that each X_i is a mean-zero random variable taking $[-1, 1]$ -values, and independent (but not necessarily identical) across all i . Indeed, by convexity, we have $e^{tx} \leq \frac{1-x}{2}e^{-t} + \frac{1+x}{2}e^t$ for all $x \in [-1, 1]$ by convexity, so that $\mathbb{E}[e^{tX}] \leq \frac{e^t + e^{-t}}{2}$. In particular, we obtain the following tail bounds on the binomial distribution.

Theorem 5.0.5 (Chernoff bound with bounded variables)

Let each X_i be an independent random variable taking values in $[-1, 1]$ and $\mathbb{E}X_i = 0$. Then $S_n = X_1 + \dots + X_n$ satisfies

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

Corollary 5.0.6

Let X be a sum of n independent Bernoulli's (with not necessarily identical probability). Let $\mu = \mathbb{E}X$ and $\lambda > 0$. Then

$$\mathbb{P}(X \geq \mu + \lambda\sqrt{n}) \leq e^{-\lambda^2/2} \quad \text{and} \quad \mathbb{P}(X \leq \mu - \lambda\sqrt{n}) \leq e^{-\lambda^2/2}$$

The quality the Chernoff compares well to that of the normal distribution. For the standard normal $Z \sim N(0, 1)$, one has $\mathbb{E}[e^{tZ}] = e^{t^2/2}$ and so

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

Set $t = \lambda$ and get

$$\mathbb{P}(Z \geq \lambda) \leq e^{-\lambda^2/2}$$

And this is actually pretty tight, as, for $\lambda \rightarrow \infty$,

$$\mathbb{P}(Z \geq \lambda) = \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-t^2/2} dt \sim \frac{e^{-\lambda^2/2}}{\sqrt{2\pi}\lambda}$$

The same proof method allows you to prove bounds for other sums of random variables, which you can adjust based on the distributions. See Alon–Spencer Appendix A for some calculations.

For example, for a sum of independent Bernoulli's with small means, we can improve on the above estimates as follows

Theorem 5.0.7

Let X be the sum of independent Bernoulli random variables (not necessarily same probability). Let $\mu = \mathbb{E}X$. For all $\varepsilon > 0$,

$$\mathbb{P}(X \geq (1 + \varepsilon)\mu) \leq e^{-((1+\varepsilon) \log(1+\varepsilon) - \varepsilon)\mu} \leq e^{-\frac{\varepsilon^2}{1+\varepsilon}\mu}$$

and

$$\mathbb{P}(X \leq (1 - \varepsilon)\mu) \leq e^{-\varepsilon^2\mu/2}.$$

Remark 5.0.8. The bounds for upper and lower tails are necessarily asymmetric, when the probabilities are small. Why? Think about what happens when $X \sim \text{Bin}(n, c/n)$, which, for a constant $c > 0$, converges as $n \rightarrow \infty$ to a Poisson distribution with mean c , whose value at k is $e^{-c} c^k / k! = e^{-\Theta(k \log k)}$ and not the sub-Gaussian decay $e^{-\Omega(k^2)}$ as one might naively predict by an incorrect application of the Chernoff bound formula.

Nonetheless, both formulas tell us that both tails exponentially decay like ε^2 for small values of ε , say, $\varepsilon \in [0, 1]$.

5.1 Discrepancy

Given a hypergraph (i.e., set family), can we color the vertices red/blue so that every edge has roughly the same number of red versus blue vertices? (Contrast this problem to 2-coloring hypergraphs from Section 1.3.)

Theorem 5.1.1

Let \mathcal{F} be a collection of m subsets of $[n]$. Then there exists some assignment $[n] \rightarrow \{-1, 1\}$ so that the sum on every set in \mathcal{F} is $O(\sqrt{n \log m})$ in absolute value.

Proof. Put ± 1 iid uniformly at random on each vertex. On each edge, the probability that the sum exceeds $2\sqrt{n \log m}$ in absolute value is, by Chernoff bound, less than $2e^{-2 \log m} = 2/m^2$. By union bound over all m edges, with probability greater than $1 - 2/m \geq 0$, no edge has sum exceeding $2\sqrt{n \log m}$. \square

Remark 5.1.2. In a beautiful landmark paper titled *Six standard deviations suffice*, [Spencer \(1985\)](#) showed that one can remove the logarithmic term by a more sophisticated semirandom assignment algorithm.

Theorem 5.1.3 (Six standard deviations suffice: [Spencer 1985](#))

Let \mathcal{F} be a collection of n subsets of $[n]$. Then there exists some assignment $[n] \rightarrow \{-1, 1\}$ so that the sum on every set in \mathcal{F} is at most $6\sqrt{n}$ in absolute value.

More generally, if \mathcal{F} be a collection of $m \geq n$ subsets of $[n]$, then we can replace $6\sqrt{n}$ by $O(\sqrt{n \log(2m/n)})$.

Remark 5.1.4. More generally, Spencer proves that the same holds if vertices have $[0, 1]$ -valued weights.

The idea, very roughly speaking, is to first generalize from $\{-1, 1\}$ -valued assignments to $[-1, 1]$ -valued assignments. Then the all-zero vector is a trivially satisfying assignment. We then randomly, in iterations, alter the values from 0 to other values in $[-1, 1]$, while avoiding potential violations (e.g., edges with sum close to $6\sqrt{n}$ in absolute value), and finalizing a color of a color when its value moves to either -1 and 1 .

Spencer's original proof was not algorithmic, and he suspected that it could not be made efficiently algorithmic. In a breakthrough result, [Bansal \(2010\)](#) gave an efficient algorithm for producing a coloring with small discrepancy. [Lovett and Meka \(2015\)](#) provided another element algorithm with a beautiful proof.

Here is a famous conjecture on discrepancy.

Conjecture 5.1.5 (Komlós)

There exists some absolute constant K so that for any $v_1, \dots, v_m \in \mathbb{R}^n$ all lying in the unit ball, there exist $\varepsilon_1, \dots, \varepsilon_m \in \{-1, 1\}$ such that

$$\varepsilon_1 v_1 + \dots + \varepsilon_m v_m \in [-K, K]^n.$$

Banaszczyk (1998) proved the bound $K = O(\sqrt{\log n})$ in a beautiful paper using deep ideas from convex geometry.

Spencer's theorem implies the special case of Komlós conjecture where all vectors v_i have the form $n^{-1/2}(\pm 1, \dots, \pm 1)$ (or more generally when all coordinates are $O(n^{-1/2})$). The deduction is easy when $m \leq n$. When $m > n$, we use the following observation.

Lemma 5.1.6

Let $v_1, \dots, v_m \in \mathbb{R}^n$. Then there exists $a_1, \dots, a_m \in [-1, 1]^m$ with $|\{i : a_i \notin \{-1, 1\}\}| \leq n$ such that

$$a_1 v_1 + \dots + a_m v_m = 0$$

Proof. Find $(a_1, \dots, a_m) \in [-1, 1]^m$ satisfying and as many $a_i \in \{-1, 1\}$ as possible. Let $I = \{i : a_i \notin \{-1, 1\}\}$. If $|I| > n$, then we can find some nontrivial linear combination of the vectors $v_i, i \in I$, allowing us to move $(a_i)_{i \in I}$'s to new values, while preserving $a_1 v_1 + \dots + a_m v_m = 0$, and end up with at one additional a_i taking $\{-1, 1\}$ -value. \square

Let us explain how to deduce the special cases of Komlós conjecture as stated earlier. Let a_1, \dots, a_m and $I = \{i : a_i \notin \{-1, 1\}\}$ as in the Lemma. Take $\varepsilon_i = a_i$ for all $i \notin I$, and apply a corollary of Spencer's theorem to find $\varepsilon_i \in \{-1, 1\}^n, i \in I$ with

$$\sum_{i \in I} (\varepsilon_i - a_i) v_i \in [-K, K]^n,$$

which would yield the desired result. The above step can be deduced from Spencer's theorem by first assuming that each $a_i \in [-1, 1]$ has finite binary length (a compactness argument), and then rounding it off one digit at a time during Spencer's theorem, starting from the least significant bit (see Corollary 8 in Spencer's paper for details).

5.2 Nearly equiangular vectors

How many vectors can one place in \mathbb{R}^d so that pairwise make equal angles?

5 Chernoff bound

Let $S = \{v_1, \dots, v_m\}$ be a set of unit vectors in \mathbb{R}^n whose pairwise inner products all equal to some $\alpha \in [-1, 1)$. How large can S be?

The Gram matrix of S , defined as the matrix of pairwise inner products, has 1's on the diagonal and α off diagonal. So

$$\begin{pmatrix} | & \cdots & | \\ v_1 & \ddots & v_m \\ | & \cdots & | \end{pmatrix}^T \begin{pmatrix} | & \cdots & | \\ v_1 & \ddots & v_m \\ | & \cdots & | \end{pmatrix} = \begin{pmatrix} v_1 \cdot v_1 & \cdots & v_1 \cdot v_m \\ \vdots & \ddots & \vdots \\ v_m \cdot v_1 & \cdots & v_m \cdot v_m \end{pmatrix} = (1 - \alpha)I_m + \alpha J_m$$

(here I_m and J_m are the $m \times m$ identity and all-ones matrix respectively). Since the eigenvalues of J_m are m (once) and 0 (repeated $m - 1$ times), the eigenvalues of $I_m + (\alpha - 1)J_m$ are $(m - 1)\alpha + 1$ (once) and $1 - \alpha$ ($m - 1$ times). Since the Gram matrix is positive semidefinite, all its eigenvalues are nonnegative, and so $\alpha \geq -1/(m - 1)$.

- If $\alpha \neq -1/(m - 1)$, then this $m \times m$ matrix is non-singular, and since its rank is at most n (as $v_i \in \mathbb{R}^n$), we have $m \leq n$.
- If $\alpha = -1/(m - 1)$, then this matrix has rank $m - 1$, and we conclude that $m \leq n + 1$.

It is left as an exercise to check all these bounds are tight.

Exercise: given m unit vectors in \mathbb{R}^n whose pairwise inner products are all $\leq -\beta$, one has $m \leq 1 + \lfloor 1/\beta \rfloor$. (A bit more difficult: show that for $\beta = 0$, one has $m \leq 2n$).

What if instead of asking for exactly equal angles, we ask for approximately the same angle. It turns out that we can get many more vectors.

Theorem 5.2.1 (Exponentially many approximately equiangular vectors)

For every $\alpha \in (0, 1)$ and $\varepsilon > 0$, there exists $c > 0$ so that for every n , one can find at least 2^{cn} unit vectors in \mathbb{R}^n whose pairwise inner products all lie in $[\alpha - \varepsilon, \alpha + \varepsilon]$.

Remark 5.2.2. Such a collection of vectors is a type of “spherical code.” Also, by examining the volume of spherical caps, one can prove an upper bound of the form $2^{C_{\alpha, \varepsilon} n}$.

Proof. Let $x_1, \dots, x_m \in \{0, 1\}^n$ (for some m to be decided), where each x_i is obtained independently at random by choosing each coordinate to be 1 with probability α and 0 with probability $1 - \alpha$ independently. For each $i \in [m]$, since

$$x_i \cdot x_i \sim \text{Binomial}(n, \alpha),$$

by the Chernoff bound (in the form Theorem 5.0.5 after subtracting the mean), for any

$t \geq 0$,

$$\mathbb{P}(|x_i \cdot x_i - n\alpha| \geq nt) \leq 2e^{-nt^2/2}.$$

Also, for any distinct $i, j \in [m]$, since

$$x_i \cdot x_j \sim \text{Binomial}(n, \alpha^2),$$

by the Chernoff bound, for any $t \geq 0$,

$$\mathbb{P}(|x_i \cdot x_j - n\alpha^2| \geq nt) \leq 2e^{-nt^2/2}.$$

Let $t > 0$ be some quantity to be decided. Let

$$Y = \{i \in [m] : |x_i \cdot x_i - n\alpha| \geq nt\}$$

and

$$Z = \{(i, j) \in [m]^2 : i \neq j \text{ and } |x_i \cdot x_j - n\alpha^2| \geq nt\}.$$

Let $Z' = \{i \in [m] : (i, j) \in Z \text{ for some } j\}$. Then

$$\mathbb{E}|Y| + \mathbb{E}|Z'| \leq \mathbb{E}|Y| + \mathbb{E}|Z| \leq 2m^2 e^{-nt^2/2}.$$

Let $X = [m] \setminus (Y \cup Z')$. Then for any distinct $i, j \in X$,

$$\frac{\alpha^2 - t}{\alpha + t} \leq \frac{x_i \cdot x_j}{\sqrt{x_i \cdot x_i} \sqrt{x_j \cdot x_j}} \leq \frac{\alpha^2 + t}{\alpha - t}$$

By choosing $t > 0$ a sufficiently small constant (depending only on α and ε), can ensure that

$$\alpha - \varepsilon \leq \frac{x_i \cdot x_j}{\sqrt{x_i \cdot x_i} \sqrt{x_j \cdot x_j}} \leq \alpha + \varepsilon.$$

Thus the unit vectors in $\{x_i/|x_i| : i \in X\}$ all have pairwise inner products in $[\alpha - \varepsilon, \alpha + \varepsilon]$. Furthermore, the expected size of this set is

$$\geq m - \mathbb{E}|Y| - \mathbb{E}|Z'| \geq m - 2m^2 e^{-nt^2/2}.$$

By choosing $m = e^{nt^2/2}/4$, the above expectation is $\geq m/2$. Thus there exists a set of vectors of size $\geq e^{nt^2/2}/4$ with the desired property. \square

Remark 5.2.3 (Equiangular lines with a fixed angle). Given a fixed angle θ , for large n , how many lines in \mathbb{R}^n through the origin can one place whose pairwise angles are all exactly θ ? This problem was solved by Jiang, Tidor, Yao, Zhang, Zhao (2021). This is the same as asking for a set of unit vectors in \mathbb{R}^n whose pairwise inner products are $\pm\alpha$. It turns out that for fixed α , the maximum number of lines grows linearly with the dimension n , and the rate of growth depends on properties of α in relation to spectral

graph theory. We refer to the cited paper for details.

5.3 Hajós conjecture counterexample

We begin by reviewing some classic result from graph theory. Recall some definitions:

- H is an **induced subgraph** of G if H can be obtained from G by removing vertices;
- H is a **subgraph** of G if H can be obtained from G by removing vertices and edges;
- H is a **subdivision** of G if H can be obtained from a subgraph of G by contracting induced paths to edges;
- H is a **minor** of G if H can be obtained from a subgraph of G by contracting edges to vertices.

Kuratowski’s theorem (1930). Every graph without $K_{3,3}$ and K_5 as subdivisions as subdivision is planar.

Wagner’s theorem (1937). Every graph free of $K_{3,3}$ and K_5 as minors is planar.

(There is a short argument shows that Kuratowski and Wagner’s theorems are equivalent.)

Four color theorem (Appel and Haken 1977) Every planar graph is 4-colorable.

Corollary: Every graph without $K_{3,3}$ and K_5 as minors is 4-colorable.

The condition on K_5 is clearly necessary, but what about $K_{3,3}$? What is the “real” reason for 4-colorability?

Hadwiger’s conjecture, below, remains a major conjectures in graph theory.

Conjecture 5.3.1 (Hadwiger 1936)

For every $t \geq 1$, every graph without a K_{t+1} minor is t -colorable.

- $t = 1$ trivial
- $t = 2$ nearly trivial (if G is K_3 -minor-free, then it’s a tree)
- $t = 3$ elementary graph theoretic arguments
- $t = 4$ is equivalent to the 4-color theorem (Wagner 1937)
- $t = 5$ is equivalent to the 4-color theorem (Robertson–Seymour–Thomas 1994; this work won a Fulkerson Prize)

- $t \geq 6$ remains open

Let us explore a variation of Hadwiger's conjecture:

Hajós conjecture. (1961) Every graph without a K_{t+1} -subdivision is t -colorable.

Hajós conjecture is true for $t \leq 3$. However, it turns out to be false in general. [Catlin \(1979\)](#) constructed counterexamples for all $t \geq 6$ ($t = 4, 5$ are still open).

It turns out that Hajós conjecture is not just false, but very false.

[Erdős–Fajtlowicz \(1981\)](#) showed that almost every graph is a counterexample (it's a good idea to check for potential counterexamples among random graphs!)

Theorem 5.3.2

With probability $1 - o(1)$, $G(n, 1/2)$ has no K_t -subdivision with $t = \lceil 10\sqrt{n} \rceil$.

From Theorem 4.4.3 we know that, with high probability, $G(n, 1/2)$ has independence number $\sim 2 \log_2 n$ and hence chromatic number $\geq (1 + o(1)) \frac{n}{2 \log_2 n}$. Thus the above result shows that $G(n, 1/2)$ is whp a counterexample to Hajós conjecture.

Proof. If G had a K_t -subdivision, say with $S \subseteq V$, $|S| = t$. Each pair of vertices of S are connected via a path, whose intermediate vertices are outside S , and distinct for different pairs of vertices.

At most n of the $\binom{t}{2}$ pairs of vertices in S can be joined this way using a path of at least two edges, since each uses up a vertex outside S . Thus at $\geq \binom{t}{2} - n$ of the pairs of vertices of S form edges.

By Chernoff bound, for fixed t -vertex subset S

$$\mathbb{P} \left(e(S) \geq \binom{t}{2} - n \right) \leq \mathbb{P} \left(e(S) \geq \frac{3}{4} \binom{t}{2} \right) \leq e^{-t^2/10}.$$

Taking a union bound over all t -vertex subsets S , and noting that

$$\binom{n}{t} e^{-t^2/10} < n^t e^{-t^2/10} \leq e^{-10n + O(\sqrt{n} \log n)} = o(1)$$

we see that whp no such S exists, so that this $G(n, 1/2)$ whp has no K_t -subdivision \square

Remark 5.3.3 (Quantitative question). One can ask the following quantitative question regarding Hadwiger's conjecture:

Can every graph without a K_{t+1} -minor can be properly colored with a small number of colors?

[Wagner \(1964\)](#) showed that every graph without K_{t+1} -minor is 2^{t-1} colorable.

5 Chernoff bound

Here is the proof: assume that the graph is connected. Take a vertex v and let L_i be the set of vertices with distance exactly i from v . The subgraph induced on L_i has no K_t -minor, since otherwise such a K_t -minor would extend to a K_{t+1} -minor with v . Then by induction L_i is 2^{t-2} -colorable (check base cases), and using alternating colors for even and odd layers L_i yields a proper coloring of G .

This bound has been improved over time. [Delcourt and Postle \(2021+\)](#) showed that every graph with no K_t -minor is $O(t \log \log t)$ -colorable.

For more on Hadwiger's conjecture, see [Seymour's survey \(2016\)](#).

6 Lovász local lemma

The Lovász local lemma (LLL) was introduced in the paper of [Erdős and Lovász \(1975\)](#). It is a powerful tool in the probabilistic method.

Usually, we have a collection of “bad events” we wish to simultaneously avoid. Here are two extreme scenerios, both of which are easy to handle:

- (Complete independence) All the bad events are independent and have probability less than 1.
- (Union bound) The sum of the bad event probabilities is less than 1.

The local lemma deals with an intermediate situation where there is a small amount of local dependencies.

We saw an application of the Lovász local lemma back in Section 1.1, where we used it to lower bound Ramsey numbers. This chapter we will explore the local lemma and its applications in depth.

6.1 Statement and proof

Definition 6.1.1 (Independence from a set of events)

Here we say that an event A_0 is **independent** from events A_1, \dots, A_m if A_0 is independent of every event of the form $B_1 \wedge \dots \wedge B_m$ (we sometimes omit the “logical and” symbol \wedge) where each B_i is either A_i or $\overline{A_i}$, i.e.,

$$\mathbb{P}(A_0 B_1 \dots B_m) = \mathbb{P}(A_0) \mathbb{P}(B_1 \dots B_m),$$

or, equivalently, using Bayes’s rule:

$$\mathbb{P}(A_0 | B_1 \dots B_m) = \mathbb{P}(A_0).$$

Given a collection of events, we can associate to it a dependency graph. This is a slightly subtle notion, as we will explain. Technically speaking, the graph can be a directed graph (=digraph), but for most applications, it will be sufficient (and easier) to use undirected graphs.

Definition 6.1.2 (Dependency (di)graph)

Let A_1, \dots, A_n be events (the “bad events” we wish to avoid). Let G be a (directed) graph with vertex set $[n]$. We say that G is a **dependency (di)graph** for the events A_1, \dots, A_n if, for every i , A_i is independent from all $\{A_j : j \notin N(i) \cup \{i\}\}$ ($N(i)$ is the set of (out)neighbors of i in G).

Remark 6.1.3 (Non-uniqueness). Given a collection of events, there can be more than one valid dependency graphs. For example, the complete graph is always a valid dependency graph.

Remark 6.1.4 (Important!). Independence \neq pairwise independence

The dependency graph is *not* made by joining $i \sim j$ whenever A_i and A_j are not independent (i.e., $\mathbb{P}(A_i A_j) \neq \mathbb{P}(A_i)\mathbb{P}(A_j)$).

Example: suppose one picks $x_1, x_2, x_3 \in \mathbb{Z}/2\mathbb{Z}$ uniformly and independently at random and set, for each $i = 1, 2, 3$ (indices taken mod 3), A_i the event that $x_{i+1} + x_{i+2} = 0$. Then these events are pairwise independent but not independent. So the empty graph on three vertices is not a valid dependency graph (on the other hand, having at least two edges makes it a valid dependency graph).

In practice, it is not too hard to construct a valid dependency graph, since most applications of the Lovász local lemma use the following setup (which we saw in Section 1.1).

Setup 6.1.5 (Random variable model / hypergraph coloring)

Let $\{x_i : i \in I\}$ be a collection of independent random variables. Let E_1, \dots, E_n be events where each E_i depends only on the variables indexed by some subset $B_i \subseteq I$ of variables. A **canonical dependency graph** for E_1, \dots, E_n has vertex set $[n]$ and an edge ij whenever $E_i \cap E_j \neq \emptyset$.

It is easy to check that the canonical dependency graph above is indeed a valid dependency graph.

Example 6.1.6 (Boolean satisfiability problem (SAT)). Given a **CNF formula** (conjunctive normal form, i.e., *and-of-or*’s), e.g., $(\wedge = \text{and}; \vee = \text{or})$

$$(x_1 \vee x_2 \vee x_3) \wedge (\overline{x_1} \vee x_2 \vee x_4) \wedge (\overline{x_2} \vee x_4 \vee x_5) \wedge \dots$$

the problem is to find a satisfying assignment with boolean variables x_1, x_2, \dots . Many problems in computer science can be modeled using this way. This problem can be viewed as in Setup 6.1.5, where A_i is the event that the i -th clause is violated.

The following formulation of the local lemma is easiest to apply and is the most commonly used. It applies to settings where the dependency graph has small maximum degree.

Theorem 6.1.7 (Lovász local lemma; symmetric form)

Let A_1, \dots, A_n be events, with $\mathbb{P}[A_i] \leq p$ for all i . Suppose that each A_i is independent from a set of all other A_j except for at most d of them. If

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

then with some positive probability, none of the events A_i occur.

Remark 6.1.8. The constant e is best possible (Shearer 1985). In most applications, the precise value of the constant is unimportant.

Theorem 6.1.9 (Lovász local lemma; general form)

Let A_1, \dots, A_n be events. For each $i \in [n]$, let $N(i)$ be such that A_i is independent from $\{A_j : j \notin \{i\} \cup N(i)\}$. If $x_1, \dots, x_n \in [0, 1)$ satisfy

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

then

$$\mathbb{P}(\text{none of the events } A_i \text{ occur}) \geq \prod_{i=1}^n (1 - x_i).$$

Proof that the general form implies the symmetric form. Set $x_i = 1/(d+1) < 1$ for all i . Then

$$x_i \prod_{j \in N(i)} (1 - x_j) \geq \frac{1}{d+1} \left(1 - \frac{1}{d+1}\right)^d > \frac{1}{(d+1)e} \geq p$$

so the hypothesis of general local lemma holds. \square

Here is another corollary of the general form local lemma, which applies if the total probability of any neighborhood in a dependency graph is small.

Corollary 6.1.10

In the setup of Theorem 6.1.9, if $\mathbb{P}(A_i) < 1/2$ and $\sum_{j \in N(i)} \mathbb{P}(A_j) \leq 1/4$ for all i , then with positive probability none of the events A_i occur.

6 Lovász local lemma

Proof. In Theorem 6.1.9, set $x_i = 2\mathbb{P}(A_i)$ for each i . Then

$$x_i \prod_{j \in N(i)} (1 - x_j) \geq x_i \left(1 - \sum_{j \in N(i)} x_j\right) = 2\mathbb{P}(A_i) \left(1 - \sum_{j \in N(i)} 2\mathbb{P}(A_j)\right) \geq \mathbb{P}(A_i).$$

(The first inequality is by “union bound.”) □

In some applications, one may need to apply the general form local lemma with carefully chosen values for x_i .

Proof of Lovász local lemma (general case). We will prove that

$$\mathbb{P}\left(A_i \mid \bigwedge_{j \in S} \bar{A}_j\right) \leq x_i \quad \text{whenever } i \notin S \subseteq [n]. \quad (6.1)$$

Once (6.1) has been established, we then deduce that

$$\begin{aligned} \mathbb{P}(\bar{A}_1 \cdots \bar{A}_n) &= \mathbb{P}(\bar{A}_1) \mathbb{P}(\bar{A}_2 \mid \bar{A}_1) \mathbb{P}(\bar{A}_3 \mid \bar{A}_1 \bar{A}_2) \cdots \mathbb{P}(\bar{A}_n \mid \bar{A}_1 \cdots \bar{A}_{n-1}) \\ &\geq (1 - x_1)(1 - x_2) \cdots (1 - x_n), \end{aligned}$$

which is the conclusion of the local lemma.

Now we prove (6.1) by induction on $|S|$. The base case $|S| = 0$ is trivial.

Let $i \notin S$. Let $S_1 = S \cap N(i)$ and $S_2 = S \setminus S_1$. We have

$$\mathbb{P}\left(A_i \mid \bigwedge_{j \in S} \bar{A}_j\right) = \frac{\mathbb{P}\left(A_i \wedge_{j \in S_1} \bar{A}_j \mid \bigwedge_{j \in S_2} \bar{A}_j\right)}{\mathbb{P}\left(\bigwedge_{j \in S_1} \bar{A}_j \mid \bigwedge_{j \in S_2} \bar{A}_j\right)} \quad (6.2)$$

For the RHS of (6.2), using that A_i is independent of $\{j \in S_2 : A_j\}$,

$$\text{numerator} \leq \mathbb{P}\left(A_i \mid \bigwedge_{j \in S_2} \bar{A}_j\right) = \mathbb{P}(A_i) \leq x_i \prod_{j \in N(i)} (1 - x_j), \quad (6.3)$$

and, denoting the elements of S_1 by $S_1 = \{j_1, \dots, j_r\}$,

$$\begin{aligned} \text{denominator} &= \mathbb{P}\left(\bar{A}_{j_1} \mid \bigwedge_{j \in S_2} \bar{A}_j\right) \mathbb{P}\left(\bar{A}_{j_2} \mid \bar{A}_{j_1} \bigwedge_{j \in S_2} \bar{A}_j\right) \cdots \mathbb{P}\left(\bar{A}_{j_r} \mid \bar{A}_{j_1} \cdots \bar{A}_{j_{r-1}} \bigwedge_{j \in S_2} \bar{A}_j\right) \\ &\geq (1 - x_{j_1}) \cdots (1 - x_{j_r}) \quad [\text{by induction hypothesis}] \\ &\geq \prod_{j \in N(i)} (1 - x_i) \end{aligned}$$

Thus (6.2) $\leq x_i$, thereby finishing the induction proof of (6.1). \square

Remark 6.1.11. We used the independence assumption only at step (6.3) of the proof. Upon a closer examination, we see that we only need to know correlation inequalities of the form $\mathbb{P}\left(A_i \mid \bigwedge_{j \in S_2} \bar{A}_j\right) \leq \mathbb{P}(A_i)$ for $S_2 \subseteq N(i)$, rather than independence. This observation allows a strengthening of the local lemma, known as a lopsided local lemma, that we will explore later in the chapter.

6.2 Coloring hypergraphs

Previously, in Theorem 1.3.1, we saw that every k -uniform hypergraph with fewer than 2^{k-1} edges is 2-colorable. The next theorem gives a sufficient local condition for 2-colorability.

Theorem 6.2.1

A k -uniform hypergraph is 2-colorable if every edge intersects at most $e^{-1}2^{k-1} - 1$ other edges

Proof. For each edge f , let A_f be the event that f is monochromatic. Then $\mathbb{P}(A_f) = p := 2^{-k+1}$. Each A_f is independent from all $A_{f'}$ where f' is disjoint from f . Since at most $d := e^{-1}2^{k-1} - 1$ edges intersect every edge, and $e(d+1)p \leq 1$, so the local lemma implies that with positive probability, none of the events A_f occur. \square

Corollary 6.2.2

For $k \geq 9$, every k -uniform k -regular hypergraph is 2-colorable. (Here k -regular means that every vertex lies in exactly k edges)

Proof. Every edge intersects $\leq d = k(k-1)$ other edges. And $e(k(k-1)+1)2^{-k+1} < 1$ for $k \geq 9$. \square

Remark 6.2.3. The statement is false for $k = 2$ (triangle) and $k = 3$ (Fano plane) but actually true for all $k \geq 4$ (Thomassen 1992).

Here is an example where the symmetric form of the local lemma is insufficient (why?).

Theorem 6.2.4

Let H be a (non-uniform) hypergraph where every edge has size ≥ 3 . Suppose

$$\sum_{f \in E(H) \setminus \{e\} : e \cap f \neq \emptyset} 2^{-|f|} \leq \frac{1}{8}, \quad \text{for each edge } e,$$

then H is 2-colorable.

Proof. Consider a uniform random 2-coloring of the vertices. Let A_e be the event that edge e is monochromatic. Then $\mathbb{P}(A_e) = 2^{-|e|+1} \leq 1/4$ since $|e| \geq 3$. Also,

$$\sum_{f \in E(H) \setminus \{e\} : e \cap f \neq \emptyset} \mathbb{P}(A_f) = \sum_{f \in E(H) \setminus \{e\} : e \cap f \neq \emptyset} 2^{-|f|+1} \leq 1/4.$$

Thus by Corollary 6.1.10 one can avoid all events A_e , and hence H is 2-colorable. \square

Remark 6.2.5. A sign to look beyond the symmetric local lemma is when there are bad events of very different nature (e.g., having very different probabilities).

Compactness argument

Now we highlight an important *compactness argument* that allows us to deduce the existence of an infinite object, even though the local lemma itself is only applicable to finite systems.

Theorem 6.2.6

Let H be a (non-uniform) hypergraph on a possibly infinite vertex set, such that each edge is finite, has at least k vertices, and intersect at most d other edges. If $e 2^{-k+1} (d+1) \leq 1$, then H has a proper 2-coloring.

Proof. From a vanilla application of the symmetric local lemma, we deduce that for any finite subset X of vertices, there exists a 2-coloring X so that no edge contained in X is monochromatic (color each vertex iid uniformly, and consider the bad event A_e that the edge $e \subseteq X$ is monochromatic).

Next we extend the coloring to the entire vertex set V by a compactness argument. The set of all colorings is $[2]^V$. By Tikhonov's theorem (which says a product of a possibly

infinite collection of compact topological spaces is compact), $[2]^V$ is compact under the product topology.

For each finite subset X , let $C_X \subseteq [2]^V$ be the subset of colorings where no edge contained in X is monochromatic. Earlier from the local lemma we saw that $C_X \neq \emptyset$. If $Y \subseteq X$, then $C_Y \supseteq C_X$. Thus

$$C_{X_1} \cap \cdots \cap C_{X_\ell} \supseteq C_{X_1 \cup \cdots \cup X_\ell},$$

so $\{C_X : |X| < \infty\}$ is a collection of closed subsets of $[2]^V$ with the finite intersection property (i.e., the intersection of any finite subcollection is nonempty).

Recall from point-set topology the following basic fact (a defining property): a space is compact if and only if every family of closed subsets having the finite intersection property has non-empty intersection.

Hence by compactness of $[2]^V$, the intersection of C_X taken over all finite X is non-empty. Any element of this intersection corresponds to a valid coloring of the hypergraph. \square

More generally, the above compactness argument yields the following.

Lemma 6.2.7 (Compactness argument)

Consider a variation of the random variable model (Setup 6.1.5) where each variable has only finitely many choices but there can be infinitely many events (each event depends on a finite subset of variables). If it is possible to avoid any finite subset of events, then it is possible to avoid all the events. \square

Remark 6.2.8. Note the conclusion may be false if we do not assume the random variable model (why?).

The next application appears in the paper of [Erdős and Lovász \(1975\)](#) where the local lemma originally appears.

Consider k -coloring the real numbers, i.e., a function $c : \mathbb{R} \rightarrow [k]$. We say that $T \subseteq \mathbb{R}$ is **multicolored** with respect to c if all k colors appear in T .

Question 6.2.9

For each k is there an m so that for every $S \subseteq \mathbb{R}$ with $|S| = m$, one can k -color \mathbb{R} so that every translate of S is multicolored?

The following theorem shows that this can be done whenever $m > (3 + \varepsilon)k \log k$ and $k > k_0(\varepsilon)$ sufficiently large.

Theorem 6.2.10

The answer to the above equation is yes if

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

Proof. Each translate of S is not multicolored with probability $p \leq k(1 - 1/k)^m$, and each translate of S intersects at most $m(m-1)$ other translates. Consider a bad event for each translate of S contained in X . The symmetric local lemma tells us that it is possible to avoid any finite collection of bad events. By the compactness argument, it is possible to avoid all the bad events. \square

Coloring arithmetic progressions

Here is an application where we need to apply the asymmetric local lemma.

Theorem 6.2.11 (Beck 1980)

For every $\varepsilon > 0$, there exists k_0 and a 2-coloring of \mathbb{Z} with no monochromatic k -term arithmetic progressions with $k \geq k_0$ and common difference less than $2^{(1-\varepsilon)k}$.

Proof. We pick a uniform random color for each element of \mathbb{Z} . For each k -term arithmetic progression in \mathbb{Z} with $k \geq k_0$ and common difference less than $2^{(1-\varepsilon)k}$, consider the event that this k -AP is monochromatic. By the compactness argument, it suffices to check that we can avoid any finite subset of events.

The event that a particular k -AP is monochromatic has probability exactly 2^{-k+1} . (Since this number depends on k , we should use the asymmetric local lemma.)

Recall that in the asymmetric local lemma (Theorem 6.1.9), we need to select $x_i \in [0, 1)$ for each bad event A_i so that

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

It is usually a good idea to select x_i to be somewhat similar to $\mathbb{P}(A_i)$. In this case, if A_i is the event corresponding to a k -AP, then we take

$$x_i = 2^{-(1-\varepsilon/2)k} = \left(\frac{\mathbb{P}(A_i)}{2}\right)^{1-\varepsilon/2}$$

(with the same ε as in the statement of the theorem).

Fix a k -AP P in \mathbb{Z} with $k \geq k_0$. The number of ℓ -APs with $\ell \geq k_0$ and common difference less than $2^{(1-\varepsilon)\ell}$ that intersects P is at most $k\ell 2^{(1-\varepsilon)\ell}$ (one choice for the element of k , a choice of the position of the ℓ -AP, and at most $2^{(1-\varepsilon)\ell}$ choices for the common difference). So to apply the local lemma, it suffices to check that

$$2^{-\varepsilon k/2+1} \leq \prod_{\ell \geq k_0} \left(1 - 2^{-(1-\varepsilon/2)\ell}\right)^{k\ell 2^{(1-\varepsilon)\ell}}.$$

Note that $1 - x \geq e^{-2x}$ for $x \in [0, 1/2]$. So

$$RHS \geq \exp\left(-\sum_{\ell \geq k_0} 2^{1-(1-\varepsilon/2)\ell} \cdot k\ell 2^{(1-\varepsilon)\ell}\right) = \exp\left(-k \sum_{\ell \geq k_0} \ell 2^{1-\varepsilon\ell/2}\right)$$

By making $k_0 = k_0(\varepsilon)$ large enough, we can ensure that $\sum_{\ell \geq k_0} \ell 2^{1-\varepsilon\ell/2} < \varepsilon/4$, and so continuing,

$$\dots \geq e^{-\varepsilon k/4} \geq 2^{-\varepsilon k/2+1}$$

provided that $k \geq k_0(\varepsilon)$. So we can apply the local lemma to conclude. \square

Decomposing coverings

We say that a collection of disks in \mathbb{R}^d is a **covering** if their union is \mathbb{R}^d . We say that it is a **k -fold covering** if every point of \mathbb{R}^d is contained in at least k disks (so 1-fold covering is the same as a covering).

We say that a k -fold covering is **decomposable** if it can be partitioned into two coverings.

In \mathbb{R}^d , is every k -fold covering by unit balls decomposable if k is sufficiently large?

A fun exercise: in \mathbb{R}^1 , every k -fold covering by intervals can be partitioned into k coverings.

[Mani-Levitska and Pach \(1986\)](#) showed that every 33-fold covering of \mathbb{R}^2 is decomposable.

What about higher dimensions?

Surprising, they also showed that for every k , there exists a k -fold indecomposable covering of \mathbb{R}^3 (and similarly for \mathbb{R}^d for $d \geq 3$).

However, it turns out that indecomposable coverings must cover the space quite unevenly:

Theorem 6.2.12 (Mani-Levitska and Pach 1986)

Every k -fold nondecomposable covering of \mathbb{R}^3 by open unit balls must cover some point $\gtrsim 2^{k/3}$ times.

Remark 6.2.13. In \mathbb{R}^d , the same proof gives $\geq c_d 2^{k/d}$.

We will need the following combinatorial geometric fact:

Lemma 6.2.14

A set of $n \geq 2$ spheres in \mathbb{R}^3 cut \mathbb{R}^3 into at most n^3 connected components.

Proof. Let us first consider the problem in one dimension lower. Let $f(m)$ be the maximum number of connected regions that m circles on a sphere in \mathbb{R}^3 can cut the sphere into.

We have $f(m+1) \leq f(m) + 2m$ for all $m \geq 1$ since adding a new circle to a set of m circles creates at most $2m$ intersection points, so that the new circle is divided into at most $2m$ arcs, and hence its addition creates at most $2m$ new regions.

Combined with $f(1) = 2$, we deduce $f(m) \leq m(m-1) + 2$ for all $m \geq 1$.

Now let $g(m)$ be the maximum number of connected regions that m spheres in \mathbb{R}^3 can cut \mathbb{R}^3 into. We have $g(1) = 2$, and $g(m+1) \leq g(m) + f(m) \leq g(m) + m(m-1) + 2$ by a similar argument as earlier. So $g(m) \leq f(m-1) + f(m-2) + \dots + f(1) + g(0) \leq m^3$. \square

Proof. Suppose for contradiction that every point in \mathbb{R}^3 is covered by at most $t \leq c2^{k/3}$ unit balls from F (for some sufficiently small c that we will pick later).

Construct an infinite hypergraph H with vertex set being the set of balls and edges having the form $E_x = \{\text{balls containing } x\}$ for some $x \in \mathbb{R}^3$. Note that $|E_x| \geq k$ since we have a k -fold covering.

Also, note that if $x, y \in \mathbb{R}^3$ lie in the same connected component in the complement of the union of all the unit spheres, then $E_x = E_y$ (i.e., the same edge).

Claim: every edge of intersects at most $d = O(t^3)$ other edges

Proof of claim: Let $x \in \mathbb{R}^3$. If $E_x \cap E_y \neq \emptyset$, then $|x - y| \leq 2$, so all the balls in E_y lie in the radius-4 ball centered at x . The volume of the radius-4 ball is 4^3 times the unit ball. Since every point lies in at most t balls, there are at most $4^3 t$ balls appearing among those E_y intersecting x , and these balls cut the radius-2 centered at x into $O(t^3)$ connected regions by the earlier lemma, and two different y 's in the same region produce the same E_y . So E_x intersects $O(t^3)$ other edges. \blacksquare

With $t \leq c2^{k/3}$ and c sufficiently small, and knowing $d = O(t^3)$ from the claim, we have $e2^{-k+1}(d+1) \leq 1$. It then follows by Theorem 6.2.6 (local lemma + compactness argument) that this hypergraph is 2-colorable, which corresponds to a decomposition of the covering, a contradiction. \square

6.3 Independent transversal

The application of the local lemma in this section is instructive in that it is not obvious at first what to choose as bad events (even if you are already told to apply the local lemma). It is worth trying different possibilities.

Every graph with maximum degree Δ contains an independent set of size $\geq |V|/(\Delta+1)$ (choose the independent set greedily). The following lemma shows that by decreasing the desired size of the independent set by a constant factor, we can guarantee an independent set that is also a transversal to a vertex set partition.

Theorem 6.3.1

Let $G = (V, E)$ be a graph with maximum degree Δ and let $V = V_1 \cup \dots \cup V_r$ be a partition, where each $|V_i| \geq 2e\Delta$. Then there is an independent set in G containing one vertex from each V_i .

Proof. The first step in the proof is simple yet subtle: we may assume that $|V_i| = k := \lceil 2e\Delta \rceil$ for each i , or else we can remove some vertices from V_i (if we do not trim the vertex sets now, we will run into difficulties later).

Pick $v_i \in V_i$ uniformly at random, independently for each i .

This is an instance of the random variable model (Setup 6.1.5), where v_1, \dots, v_r are the random variables.

We would like to design a collection of “bad events” so that if we avoid all of them, then $\{v_1, \dots, v_r\}$ is guaranteed to be independent set.

What do we choose as bad events? It turns out that some choices work better than others.

Attempt 1:

For each $1 \leq i < j \leq r$ where there exists an edge between V_i and V_j , let $A_{i,j}$ be the event that v_i is adjacent to v_j .

We find that $\mathbb{P}(A_{i,j}) \leq \Delta/k$.

The canonical dependency graph has $A_{i,j} \sim A_{i',j'}$ if and only if the two sets $\{i, j\}$ and $\{i', j'\}$ intersect. This dependency graph has max degree $\leq 2\Delta k$ (starting from (i, j)),

look at the neighbors of all vertices in $V_i \cup V_j$). The max degree is too large compared to the bad event probabilities.

Attempt 2:

For each edge $e \in E$, let A_e be the event that both endpoints of e are picked.

We have $\mathbb{P}(A_e) = 1/k^2$.

The canonical dependency graph has $A_e \sim A_f$ if some V_i intersects both e and f .

This dependency graph has max degree $\leq 2k\Delta$ (if e is between V_i and V_j , then f must be incident to $V_i \cup V_j$).

We have $e(1/k^2)(2k\Delta + 1) \leq 1$, so the local lemma implies the with probability no bad event occurs, in which case $\{v_1, \dots, v_r\}$ is an independent set. \square

Remark 6.3.2. Alon (1988) introduced the above result as lemma in his near resolution of the still-open *linear arboricity conjecture* (see the Alon–Spencer textbook §5.5). Alon’s approach makes heavy use of the local lemma.

Haxell (1995, 2001) relaxed the hypothesis to $|V_i| \geq 2\Delta$ for each i . The statement becomes false if 2Δ is replaced by $2\Delta - 1$ (Szabó and Tardos 2006).

6.4 Directed cycles of length divisible by k

A directed graph is ***d-regular*** if every vertex has in-degree d and out-degree d .

Theorem 6.4.1 (Alon and Linial 1989)

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

Corollary 6.4.2

For every k there exists d so that every $2d$ -regular graph has a cycle of length divisible by k .

Proof. Every $2d$ -regular graph can be made into a d -regular digraph by orientating its edges according to an Eulerian tour. And then we can apply the previous theorem. \square

We will prove the following more general statement.

Theorem 6.4.3 (Alon and Linial 1989)

Every directed graph with min out-degree δ and max in-degree Δ contains a cycle of length divisible by $k \in \mathbb{N}$ as long as

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

Proof. By deleting edges, can assume that every vertex has out-degree exactly δ .

Assign every vertex v an element $x_v \in \mathbb{Z}/k\mathbb{Z}$ iid uniformly at random.

We will look for directed cycles where the labels increase by 1 (mod k) at each step. These cycles all have length divisible by k .

For each vertex v , let A_v be the event that there is nowhere to go from v (i.e., if no outneighbor is labeled $x_v + 1 \pmod{k}$). We have

$$\mathbb{P}(A_v) = (1 - 1/k)^\delta \leq e^{-\delta/k}.$$

Since A_v depends only on $\{x_w : w \in \{v\} \cup N^+(v)\}$, where $N^+(v)$ denotes the out-neighbors of v and $N^-(v)$ the in-neighbors of v , the canonical dependency graph has

$$A_v \sim A_w \text{ if } \{v\} \cup N^+(v) \text{ intersects } \{w\} \cup N^+(w).$$

The maximum degree in the dependency graph is at most $\Delta + \delta\Delta$ (starting from v , there are

- (1) at most Δ choices stepping backward
- (2) δ choices stepping forward, and
- (3) at most $\delta(\Delta - 1)$ choices stepping forward and then backward to land somewhere other than v).

So an application of the local lemma shows that we are done as long as $e^{1-\delta/k}(1+\Delta+\delta\Delta)$, i.e.,

$$k \leq \delta/(1 + \log(1 + \Delta + \delta\Delta)).$$

This is almost, but not quite the result (though, for most applications, we would be perfectly happy with such a bound).

The final trick is to notice that we actually have an even smaller valid dependency digraph:

$$A_v \text{ is independent of all } A_w \text{ where } N^+(v) \text{ is disjoint from } N^+(w) \cup \{w\}.$$

Indeed, even if we fix the colors of all vertices outside $N^+(v)$, the conditional probability that A_v is still $(1 - 1/k)^\delta$.

The number of w such that $N^+(v)$ intersects $N^+(w) \cup \{w\}$ is at most $\delta\Delta$ (no longer need to consider (1) in the previous count). And we have

$$ep(\delta\Delta + 1) \leq e^{1-\delta/k}(\delta\Delta + 1) \leq 1.$$

So we are done by the local lemma. □

6.5 Lopsided local lemma

Let us move beyond the random variable model, and consider a collection of bad events in the general setup of the local lemma. Instead of requiring that each event is independent of its non-neighbors (in the dependency graph), what if we assume that avoiding some bad events make it easier to avoid some others? Intuitively, it seems that it would only make it easier to avoid bad events.

We can make this notion precise by re-examining the proof of the local lemma. Where did we actually use the independence assumption in the hypothesis of the local lemma? It was in the following step, Equation (6.3):

$$\text{numerator} \leq \mathbb{P}\left(A_i \mid \bigwedge_{j \in S_2} \bar{A}_j\right) = \mathbb{P}(A_i) \leq x_i \prod_{j \in N(i)} (1 - x_j).$$

If we had changed the middle $=$ to \leq , the whole proof would remain valid. This observation allows us to weaken the independence assumption. Therefore we have the following theorem, which was used by [Erdős and Spencer \(1991\)](#) to give an application to Latin transversals that we will see shortly.

Theorem 6.5.1 (Lopsided local lemma)

Let A_1, \dots, A_n be events. For each i , let $N(i) \subseteq [n]$ be such that

$$\mathbb{P}\left(A_i \mid \bigwedge_{j \in S} \overline{A_j}\right) \leq \mathbb{P}(A_i) \quad \text{for all } i \in [n] \text{ and } S \subseteq [n] \setminus (N(i) \cup \{i\}) \quad (6.1)$$

Suppose there exist $x_1, \dots, x_n \in [0, 1)$ such that

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

Then

$$\mathbb{P}(\text{none of the events } A_i \text{ occur}) \geq \prod_{i=1}^n (1 - x_i).$$

Like earlier, by setting $x_i = 1/(d+1)$, we deduce a symmetric version that is easier to apply.

Corollary 6.5.2 (Lopsided local lemma; symmetric version)

In the previous theorem, if $|N(i)| \leq d$ and $\mathbb{P}(A_i) \leq p$ for every $i \in [n]$, and $ep(d+1) \leq 1$, then with positive probability none of the events A_i occur.

The (di)graph where $N(i)$ is the set of (out-)neighbors of i is called a **negative dependency (di)graph**.

Remark 6.5.3 (Important!). Just as with the usual local lemma, the negative dependency graph is **not** constructed by simply checking pairs of events.

The hypothesis of Theorem 6.5.1 seems annoying to check. Fortunately, many applications of lopsided local lemma fall within a model that we will soon describe, where there is a canonical negative dependency graph that is straightforward to construct. This is analogous to the random variable model for the usual local lemma, where the canonical dependence graph has two events adjacency if they share variables.

Random injection model

We describe a random injection model where there is an easy-to-construct canonical negative dependency graph (Lu and Székely 2007).

Recall that a **matching** in a graph is a subset of edges with no two sharing a vertex. In a bipartite graph with vertex parts X and Y , a **complete matching** from X to Y is a matching where every vertex of X belongs to an edge of the matching.

Setup 6.5.4 (Random injection model)

Let X and Y be finite sets with $|X| \leq |Y|$.

Let $f: X \rightarrow Y$ be an injection chosen uniformly at random. We can also represent f by a complete matching M from X to Y in $K_{X,Y}$ (the complete bipartite graph between X and Y). We will speak interchangeably of the injection f and matching M .

For a given matching F (not necessarily complete) in $K_{X,Y}$, let A_F denote the event that $F \subseteq M$.

Let F_1, \dots, F_n be matchings in $K_{X,Y}$. The **canonical negative dependency graph** for the vents A_{F_1}, \dots, A_{F_n} has one vertex for each event, and an edge between the events A_{F_i} and A_{F_j} ($i \neq j$) if F_i and F_j are not vertex disjoint.

The following result shows that the above canonical negative dependency graph is a valid for the lopsided local lemma (Theorem 6.5.1).

Theorem 6.5.5 (Nonnegative dependence for random injections)

In Setup 6.5.4, let F_0 be a matching in $K_{X,Y}$ such that F_0 is vertex disjoint from $F_1 \cup \dots \cup F_k$. Then

$$\mathbb{P}(A_{F_0} \mid \overline{A_{F_1}} \cdots \overline{A_{F_k}}) \leq \mathbb{P}(A_{F_0}).$$

Proof. Let $X_0 \subseteq X$ and $Y_0 \subseteq Y$ be the set of endpoints of F_0 .

For each matching T in $K_{X,Y}$, let

$$\mathcal{M}_T = \{\text{complete matchings from } X \text{ to } Y \text{ containing } T \text{ but not containing any of } F_1, \dots, F_k\}.$$

For the desired inequality, note that

$$LHS = \mathbb{P}(A_{F_0} \mid \overline{A_{F_1}} \cdots \overline{A_{F_k}}) = \frac{|\mathcal{M}_{F_0}|}{|\mathcal{M}_\emptyset|} = \frac{|\mathcal{M}_{F_0}|}{\sum_{T: X_0 \hookrightarrow Y} |\mathcal{M}_T|}$$

where the sum is taken over all $|Y|(|Y| - 1) \cdots (|Y| - |X| + 1)$ complete matchings T from X_0 to Y (which we denote by $T: X_0 \hookrightarrow Y$), and

$$RHS = \mathbb{P}(A_{F_0}) = \frac{1}{|\{T: X_0 \hookrightarrow Y\}|}.$$

Thus to show that $LHS \leq RHS$, it suffices to prove

$$|\mathcal{M}_{F_0}| \leq |\mathcal{M}_T| \quad \text{for every } T: X_0 \hookrightarrow Y.$$

It suffices to construct an injection $\mathcal{M}_{F_0} \hookrightarrow \mathcal{M}_T$. Let Y_1 be the set of endpoints of T in Y . Fix a permutation σ of Y such that

- σ fixes all elements of Y outside $Y_0 \cup Y_1$; and
- σ sends F_0 to T .

Then σ induces a permutation on the set of complete matchings from X to Y . It remains to show that if we start with a matching in \mathcal{M}_{F_0} , so that it avoids F_i for all $i \geq 1$, then it is sent to a matching that also avoids F_i for all $i \geq 1$ (and hence lies in \mathcal{M}_T). Indeed, this follows from the hypothesis that none of the edges in F_i use any vertex from X_0 or Y_0 . \square

As an example, here is a quick application.

Corollary 6.5.6 (Derangement lower bound)

The probability that a uniform random permutation of $[n]$ has no fixed points is at least $(1 - 1/n)^n$.

Proof. In the random injection model, let $X = Y = [n]$. Let $f: X \rightarrow Y$ be a uniform random permutation. For each $i \in [n]$, let F_i be the single edge (i, i) , i.e., A_{F_i} is the event that $f(i) = i$. Note that the canonical negative dependency graph is empty since no two F_i 's share a vertex. Since $\mathbb{P}(A_i) = 1 - 1/n$, we can set $x_i = 1 - 1/n$ for each i in the lopsided local lemma to obtain the conclusion

$$\mathbb{P}(f \text{ has no fixed points}) = \mathbb{P}(\overline{A_1} \cdots \overline{A_n}) \geq \left(1 - \frac{1}{n}\right)^n. \quad \square$$

Remark 6.5.7. A fixed-point free permutation is called a **derangement**. Using inclusion-exclusion, one can deduce an exact answer to the above question: $\sum_{i=0}^n (-1)^i / i!$. This quantity converges to $1/e$ as $k \rightarrow \infty$, and the above lower bound $(1 - 1/n)^n$ also converges to $1/e$ and so is asymptotically optimal.

Latin transversal

A **Latin square** of order n is an $n \times n$ array filled with n symbols so that every symbol appears exactly once in every row and column. Example:

$$\begin{array}{ccc} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{array}$$

These objects are called Latin squares because they were studied by Euler (1707–1783) who used Latin symbols to fill the arrays.

Given an $n \times n$ array, a **transversal** is a set of n entries with one in every row and column. A **Latin transversal** is a transversal with distinct entries. Example:

1	2	3
2	3	1
3	1	2

Here is a famous open conjecture about Latin transversals.¹ (Do you see why the hypothesis on parity is necessary?)

Conjecture 6.5.8 (Ryser 1967)

Every odd order Latin square has a transversal.

The conjecture should be modified for even order Latin squares.

Conjecture 6.5.9 (Ryser-Brauer-Stein conjecture)

Every even order Latin square has a transversal containing all but at most one symbol.

Remark 6.5.10. Keevash, Pokrovskiy, Sudakov and Yepremyan (2022) proved that every order n Latin square contains a transversal containing all but $O(\log n / \log \log n)$ symbols, improving an earlier bound of $O(\log^2 n)$ by Hatami and Shor (2008).

Recently, Montgomery announced a proof of the conjecture for all sufficiently large even n . The proof uses sophisticated techniques combining the semi-random method and the absorption method.

The next result is the original application of the lopsided local lemma.

Theorem 6.5.11 (Erdős and Spencer 1991)

Every $n \times n$ array where every entry appears at most $n/(4e)$ times has a Latin transversal.

Proof. Pick a transversal uniformly at random. This is the same as picking a permutation $f: [n] \rightarrow [n]$ uniformly at random. In Setup 6.5.4, the random injection model, transversals correspond to perfect matchings.

For each pair of equal entries in the array not both lying in the same row or column, consider the bad event that the transversal contains both entries.

The canonical negative dependency graph is obtained by joining an edge between two bad events if the four entries involved share some row or column.

¹Not to be confused with another conjecture also known as Ryser's conjecture concerning an inequality between the covering number and the matching number of multipartite hypergraphs, as a generalization of König's theorem. See Best and Wanless (2018) for a historical commentary and a translation of Ryser's 1967 paper.

Let us count neighbors in this negative dependency graph. Fix a pair of equal entries in the array. Their rows and columns span fewer than $4n$ entries, and for each such entry z , there are at most $n/(4e) - 1$ choices for another entry equal to z . Thus the maximum degree in the canonical negative dependence graph is

$$\leq (4n - 4) \left(\frac{n}{4e} - 1 \right) \leq \frac{n(n-1)}{e} - 1.$$

We can now apply the symmetric lopsided local lemma to conclude that with positive probability, none of the bad events occur. \square

6.6 Algorithmic local lemma

Consider an instance of a problem in the random variable setting (e.g., k -CNF) for which the local lemma guarantees a solution. Can one find a satisfying assignment efficiently?

The local lemma tells you that some good configuration exists, but the proof is non-constructive. The probability that a random sample avoids all the bad events is often very small (usually exponentially small, e.g., in the case of a set of independent bad events). It had been an open problem for a long time whether the local lemma can be made algorithmic.

[Moser \(2009\)](#), during his PhD, achieved a breakthrough by coming up with the first efficient algorithmic version of the local lemma for finding a satisfying assignment for k -CNF formulas. [Moser and Tardos \(2010\)](#) later extended the algorithm for the general local lemma in the random variable model.

Remark 6.6.1 (Too hard in general). The Moser–Tardos algorithm works in the random variable model (there are subsequent work that concern other models such as the random injection model). Some assumption on the model is necessary since the problem can be computationally hard in general.

For example, let $q = 2^k$, and $f: [q] \rightarrow [q]$ be some fixed bijection (with an explicit description and easy to compute). Consider the computational task of inverting f : given $y \in [q]$, find x such that $f(x) = y$ (we would like an algorithm with running time polynomial in k).

If $x \in [q]$ is chosen uniformly, then $f(x) \in [q]$ is also uniform. For each $i \in [k]$, let A_i be the event that $f(x)$ and y disagree on i -th bit. Then A_1, \dots, A_k are independent events. Also, $f(x) = y$ if and only if no event A_i occurs. So a trivial version of the local lemma (with empty dependency graph) implies the existence of some x such that $f(x) = y$.

On the other hand, it is believed that there exist functions f that is easy to compute but hard to invert. Such functions are called **one-way functions**, and they are a fundamental building block in cryptography. For example, let g be a multiplicative generator of \mathbb{F}_q , and let $f: \mathbb{F}_q \rightarrow \mathbb{F}_q$ be given by $f(0) = 0$ and $f(x) = g^x$ and for $x \neq 0$. Then inverting f is the **discrete logarithm problem**, which is believed to be computationally difficult. The computational difficulty of this problem is the basis for the security of important public key cryptography schemes, such as the Diffie–Hellman key exchange.

Moser–Tardos algorithm

The Moser–Tardos algorithm considers problems in the random variable model (Setup 6.1.5). The algorithm is surprisingly simple.

Algorithm 6.6.2 (Moser–Tardos “fix-it”)

input : a set of variables and events in the random variable model

output : an assignment of variables avoiding all bad events

Initialize by setting all variables to arbitrary values;

while *there is some violated event* **do**

 Pick an arbitrary violated event and uniformly resample its variables;

(We can make the algorithm more well defined by specifying a way to pick an “arbitrary” choice, e.g., the lexicographically first choice.)

Theorem 6.6.3 (Moser and Tardos 2010)

In Algorithm 6.6.2, letting A_1, \dots, A_n denote the bad events, suppose there are $x_1, \dots, x_n \in [0, 1)$ such that

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

then for each i ,

$$\mathbb{E}[\text{number of times that } A_i \text{ is chosen for resampling}] \leq \frac{x_i}{1 - x_i}.$$

We won’t prove the general theorem here. The proof in Moser and Tardos (2010) is beautifully written and not too long. I highly recommend it reading it. In the next subsection, we will prove the correctness of the algorithm in a special case using a neat idea known as entropy compression.

Remark 6.6.4 (Las Vegas versus Monte Carlo). Here are some important classes of randomized algorithms:

- **Monte Carlo algorithm** (MC): a randomized algorithm that terminates with an output, but there is a small probability that the output is incorrect;
- **Las Vegas algorithm** (LV): a randomized algorithm that always returns a correct answer, but may run for a long time (or possibly forever).

The Moser–Tardos algorithm is a LV algorithm whose expected runtime is bounded by $\sum_i x_i / (1 - x_i)$, which is usually at most polynomial in the parameters of the problem.

We are usually interested in randomized algorithms whose running time is small (e.g., at most a polynomial of the input size).

We can convert an efficient LV algorithm into an efficient MC algorithm as follows: suppose the LV algorithm has expected running time T , and now we run the algorithm but if it takes more than CT time, then halt and declare a failure. Markov’s inequality then shows that the algorithm fails with probability $\leq 1/C$.

However, it is not always possible to convert an efficient MC algorithm into an efficient LV algorithm. Starting with an MC algorithm, one might hope to repeatedly run it until a correct answer has been found. However, there might not be an efficient way to **check the answer**.

For example, consider the problem of finding a **Ramsey coloring**, specifically, 2-edge-coloring of K_n without a monochromatic clique of size $\geq 100 \log_2 n$. A uniform random coloring works with overwhelming probability, as can be checked by a simple union bound (see Theorem 1.1.2). However, we do not have an efficient way to check whether the random edge-coloring indeed has the desired property. It is a major open problem to find an LV algorithm for finding such an edge-coloring.

Entropy compression argument

We now give a simple and elegant proof for a special case of the above algorithm, due to Moser (2009). Actually, the argument in his paper is quite a bit more complicated. Moser presented a version of the proof below in a conference, and his ideas were popularized by Fortnow and Tao. (Fortnow called Moser’s talk “one of the best STOC talks ever”). Tao introduced the phrase **entropy compression argument** to describe Moser’s influential idea. (We won’t use the language of entropy here, and instead use a more elementary argument involving counting and the pigeonhole principle. We will discuss entropy in Chapter 10.)

To keep the argument simple, we work in the setting of k -CNFs. Recall from Example 6.1.6 that a **k -CNF formula** (conjunctive normal form) consist of a logical conjunction (i.e., and, \wedge) of clauses, where each **clause** is a disjunction (i.e., or, \vee) of exactly k literals. We shall require that the k literals of each clause use distinct

6 Lovász local lemma

variables (x_1, \dots, x_N) , and each variable appears either in its positive x_i or negative form \bar{x}_i . For example, here is a 3-CNF with 4 clauses on 6 variables:

$$(x_1 \vee x_2 \vee x_3) \wedge (\bar{x}_1 \vee x_2 \vee x_4) \wedge (\bar{x}_2 \vee x_4 \vee x_5) \wedge (x_3 \vee \bar{x}_5 \vee \bar{x}_6).$$

The problem is to find a satisfying assignment with boolean variables so that the expression output to TRUE.

Algorithm 6.6.5 (Moser “fix-it”)

input : a k -CNF

output : a satisfying assignment

```

1 Initialize by setting all variables to arbitrary values;
2 while there is some violated clause  $C$  do
3    $\lfloor$  fix ( $C$ );
4 Subroutine fix (clause  $C$ ) :
5   Resample the variables in  $C$  uniformly at random;
6   while there is some violated clause  $D$  that shares a variable with  $C$  do
7      $\lfloor$  fix ( $D$ );

```

(We can make the algorithm more well defined by specifying a way to pick an “arbitrary” choice, e.g., the lexicographically first choice. Also, in Line 6, we allow taking $D = C$.)

Theorem 6.6.6 (Correctness of Moser’s algorithm)

Given a k -CNF where every clause shares variables with at most 2^{k-3} other clauses, Algorithm 6.6.5 output a satisfying assignment with expected running time at most polynomial in the number of variables and clauses.

Note that the Lovász local lemma guarantees the existence of a solution if each clause shares variables with at most $2^k/e - 1$ clauses (each clause is violated with probability exactly 2^{-k} in a uniform random assignment of variables). So the theorem above is tight up to an unimportant constant factor.

Lemma 6.6.7 (Outer while loop)

Each clause of the k -CNF appears at most once as a violated clause in the outer while loop (Line 2).

Proof. Given an assignment of variables, by calling $\text{fix}(C)$ for any clause C , any clause that was previously satisfied remains satisfied after the completion of the execution of $\text{fix}(C)$. Furthermore, C becomes satisfied after the function call. Thus, once $\text{fix}(C)$ is called, C can never show up again as a violated clause in Line 2. \square

Lemma 6.6.8 (The number of recursive calls to `fix`)

Fix a k -CNF on n variables where every clause shares variables with at most 2^{k-3} other clauses. Also fix a clause C_0 and some assignment of variables. Then, in an execution of `fix`(C_0), for any positive integer ℓ ,

$$\mathbb{P}(\text{there are at least } \ell \text{ recursive calls to } \text{fix} \text{ in Line 7}) \leq 2^{-\ell+n+1}.$$

It follows that the expected number of recursive calls to `fix` is $n + O(1)$. Thus, in the Moser algorithm (Algorithm 6.6.5), the expected total number of calls to `fix` is $mn + O(m)$, where n is the number of variables and m is the number of clauses. This proves the correctness of the algorithm (Theorem 6.6.6).

Proof. Let us formalize the randomness in the algorithm by first initializing a random string of bits. Specifically, let $x \in \{0, 1\}^{k\ell}$ be generated uniformly at random. Whenever a clause is resampled in Line 5, one replaces the variables in the clause by the next k bits from x . Furthermore, if the line Line 7 is called for the ℓ -th time, we halt the algorithm and declare a failure (as we would have run out of random bits to resample had we continued).

At the same time, we keep an *execution trace* which keeps track of which clauses got called `fix`, and also when the inner while loop Line 6 ends. Note that the very first call to `fix`(C_0) is not included in the execution trace since it is already given as fixed and so we don't need to include this information. Here is an example of an execution trace, writing C7 for the 7th clause in the k -CNF:

```
fix(C7) called
fix(C4) called
fix(C7) called
while loop ended
fix(C2) called
while loop ended
while loop ended
...
```

For illustration, here is the example of how clause variables could intersect:

```
C2: ****
C4:  ****
C7:      ****
```

It is straightforward to deduce which `while loop ended` corresponds to which `fix` call by reading the execution trace and keeping track of a first-in-first-out stack.

Encoding the execution trace as a bit string. We fix at the beginning some canonical order of all clauses (e.g., lexicographic). It would be too expensive to refer to each clause in its absolute position in this order (this is an important point!). Instead, we note that every clause shares variables with at most 2^{k-3} other clauses, and only these $\leq 2^{k-3}$ could be called in the inner while loop in Line 6. So we can record which one got called using a $k - 3$ bit string.

- **fix(D) called:** suppose this was called inside an execution of **fix(C)**, and D is the j -th clause among all clauses sharing a variable with C , then record in the execution trace bit string \emptyset followed by exactly $\ell - 3$ bits giving the binary representation of j (prepended by zeros to get exactly $\ell - 3$ bits).
- **while loop ended:** record 1 in the execution trace bit string.

Note that one can recover the execution trace from the above bit string encoding.

Now, suppose the algorithm terminates as a failure due to **fix** being called the ℓ -th time. Here is the key claim.

Key claim (recovering randomness). At the moment right before the ℓ -th recursive call to **fix** on Line 7, we can completely recover x from the current variable assignments and the execution trace.

Note that all ℓk random bits in x have been used up at this point.

To see the key claim, note that from the execution trace, we can determine which clauses were resampled and in what order. Furthermore, if **fix(D)** was called on Line 7, then D must have been violated right before the call, and there is a unique possibility for the violating assignment to D right before the call (e.g., if $D = x_1 \vee x_2 \vee \overline{x_3}$, then the only violating assignment is $(x_1, x_2, x_3) = (0, 0, 1)$). We can then rewind history, and put the reassigned values to D back into the random bit string x to complete recover x .

How long can the execution bit string be? It has length $\leq \ell(k - 1)$. Indeed, each of the $\leq \ell$ recursive calls to **fix** produces $k - 2$ bits for the call to **fix** and 1 bit for ending the while loop. So the total number of possible execution strings is $\leq 2^{\ell(k-1)+1}$ (the $+1$ accounts for variable lengths, though it can be removed with a more careful analysis).

Thus, the key claim implies that each $x \in \{0, 1\}^{\ell k}$ that leads to a failed execution produces a unique pair (variable assignment, execution bit string). Thus

$$\mathbb{P}(\geq \ell \text{ recursive calls to fix}) 2^{\ell k} = |\{x \in \{0, 1\}^n \text{ leading to failure}\}| \leq 2^n 2^{\ell(k-1)+1}.$$

Therefore, the failure probability is $\leq 2^{-\ell k + n + 1}$. \square

Remark 6.6.9 (Entropy compression). Tao use the phrase “entropy compression” to describe this argument. The intuition is that the recoverability of the random bit string

x means that we are somehow “compressing” a ℓk -bit random string into a shorter length losslessly, but that would be impossible. Each call to `fix` uses up k random bits and converts it to $k - 1$ bits to the execute trace (plus at most n bits of extra information, namely the current variables assignment, and this is viewed as a constant amount of information), and this conversion is reversible. So we are “compressing entropy.” The conservation of information tells us that we cannot losslessly compress k random bits to $k - 1$ bits for very long.

Remark 6.6.10 (Relationship between the two proofs of the local lemma?). The above proof, along with extensions of these ideas in [Moser and Tardos \(2010\)](#), seems to give a completely different proof of the local lemma than the one we saw at the beginning of the chapter. Is there some way to relate these seemingly completely different proofs? Are they secretly the same proof? We do not know. This is an interesting open-ended research problem.

7 Correlation inequalities

7.1 Harris–FKG inequality

Recall that $A \subseteq \{0, 1\}^n$ is called an *increasing event* (also: *increasing property*, *up-set*) if A is upwards-closed, that is,

$$x \in A \text{ and } x \leq y \text{ (coordinatewise)} \implies y \in A.$$

Similarly, a *decreasing event* is defined by a downward closed collection of subset of $\{0, 1\}^n$. A subset $A \subseteq \{0, 1\}^n$ is increasing if and only if its complement $\bar{A} \subseteq \{0, 1\}^n$ is decreasing.

The main theorem of this chapter tells us that

increasing events of independent variables are positively correlated .

Theorem 7.1.1 (Harris 1960)

If A and B are increasing events of independent boolean random variables, then

$$\mathbb{P}(AB) \geq \mathbb{P}(A)\mathbb{P}(B).$$

Equivalently, we can write $\mathbb{P}(A \mid B) \geq \mathbb{P}(A)$.

Remark 7.1.2 (Independence assumption). It is important the boolean random variables are independent, also they do not have to be identically distributed.

There are other important settings where the independence assumption can be relaxed. This is important for certain statistical physics models, where much of this theory originally arose. Indeed, the above inequality is often called the *FKG inequality*, attributed to [Fortuin, Kasteleyn, Ginibre \(1971\)](#), who proved a more general result in the setting of distributive lattices, which we will not discuss here (see Alon–Spencer).

Remark 7.1.3 (Percolation). Many of such inequalities were initially introduced for the study of *percolations*. A classic setting of this problem takes place in infinite grid with vertices \mathbb{Z}^2 with edges connecting adjacent vertices at distance 1. Suppose we keep each edge of this infinite grid with probability p independently, what is the

7 Correlation inequalities

probability that the origin is part of an infinite component (in which case we say that there is “percolation”)? This is supposed to an idealized mathematical model of how a fluid permeates through a medium. Harris showed that with probability 1, percolation does not occur for $p \leq 1/2$. A later breakthrough of [Kesten \(1980\)](#) shows that percolation occurs with probability 1 for all $p > 1/2$. Thus the “bond percolation threshold” for \mathbb{Z}^2 is exactly $1/2$. Such exact results are extremely rare.

Example 7.1.4. Here is a quick application of Harris’ inequality to a random graph $G(n, p)$:

$$\mathbb{P}(\text{planar} \mid \text{connected}) \leq \mathbb{P}(\text{planar}).$$

Indeed, being planar is a decreasing property, whereas being connected is an increasing property.

We state and prove a more general result, which says that independent random variables possess **positive association**.

Let each Ω_i be a linearly ordered set (i.e., $\{0, 1\}, \mathbb{R}$) and $x_i \in \Omega_i$ with respect to some probability distribution independent for each i . We say that a function $f(x_1, \dots, x_n)$ is **monotone increasing** if

$$x \leq y \text{ (coordinatewise)} \implies f(x) \leq f(y).$$

Theorem 7.1.5 (Harris)

If f and g are monotone increasing functions of independent random variables, then

$$\mathbb{E}[fg] \geq (\mathbb{E}f)(\mathbb{E}g).$$

This version of Harris inequality implies the earlier version by setting $f = 1_A$ and $g = 1_B$.

Proof. We use induction on n .

For $n = 1$, for independent $x, y \in \Omega_1$, we have

$$0 \leq \mathbb{E}[(f(x) - f(y))(g(x) - g(y))] = 2\mathbb{E}[fg] - 2(\mathbb{E}f)(\mathbb{E}g).$$

So $\mathbb{E}[fg] \geq (\mathbb{E}f)(\mathbb{E}g)$. (The one-variable case is sometimes called Chebyshev’s inequality. It can also be deduced using the rearrangement inequality).

Now assume $n \geq 2$. Let $h = fg: \Omega_1 \times \dots \times \Omega_n \rightarrow \mathbb{R}$. Define marginals $f_1, g_1, h_1: \Omega_1 \rightarrow$

\mathbb{R} by

$$\begin{aligned} f_1(y_1) &= \mathbb{E}[f|x_1 = y_1] = \mathbb{E}_{(x_2, \dots, x_n) \in \Omega_2 \times \dots \times \Omega_n} [f(y_1, x_2, \dots, x_n)], \\ g_1(y_1) &= \mathbb{E}[g|x_1 = y_1] = \mathbb{E}_{(x_2, \dots, x_n) \in \Omega_2 \times \dots \times \Omega_n} [g(y_1, x_2, \dots, x_n)], \\ h_1(y_1) &= \mathbb{E}[h|x_1 = y_1] = \mathbb{E}_{(x_2, \dots, x_n) \in \Omega_2 \times \dots \times \Omega_n} [h(y_1, x_2, \dots, x_n)]. \end{aligned}$$

Note that f_1 and g_1 are 1-variable monotone increasing functions on Ω_1 .

For every fixed $y_1 \in \Omega_1$, the function $(x_2, \dots, x_n) \mapsto f(y_1, x_2, \dots, x_n)$ is monotone increasing, and likewise with g . So applying the induction hypothesis for $n - 1$, we have

$$h_1(y_1) \geq f_1(y_1)g_1(y_1). \quad (7.1)$$

Thus

$$\begin{aligned} \mathbb{E}[fg] &= \mathbb{E}[h] = \mathbb{E}[h_1] \geq \mathbb{E}[f_1g_1] && \text{[by (7.1)]} \\ &\geq (\mathbb{E}f_1)(\mathbb{E}g_1) && \text{[by the } n = 1 \text{ case]} \\ &= (\mathbb{E}f)(\mathbb{E}g). && \square \end{aligned}$$

Corollary 7.1.6 (Decreasing events and multiple events)

Let A and B be events on independent random variables.

- (a) If A and B are decreasing, then $\mathbb{P}(A \wedge B) \geq \mathbb{P}(A)\mathbb{P}(B)$.
- (b) If A is increasing and B is decreasing, then $\mathbb{P}(A \wedge B) \leq \mathbb{P}(A)\mathbb{P}(B)$.

If A_1, \dots, A_k are all increasing (or all decreasing) events on independent random variables, then

$$\mathbb{P}(A_1 \cdots A_k) \geq \mathbb{P}(A_1) \cdots \mathbb{P}(A_k).$$

Proof. For the second inequality, note that the complement \overline{B} is increasing, so

$$\mathbb{P}(AB) = \mathbb{P}(A) - \mathbb{P}(A\overline{B}) \stackrel{\text{Harris}}{\leq} \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(\overline{B}) = \mathbb{P}(A)\mathbb{P}(B).$$

The proof of the first inequality is similar. For the last inequality we apply the Harris inequality repeatedly. \square

7.2 Applications to random graphs

Triangle-free probability

Question 7.2.1

What's the probability that $G(n, p)$ is triangle-free?

Harris inequality will allow us to prove a lower bound. In the next chapter, we will use Janson inequalities to derive upper bounds.

Theorem 7.2.2

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \geq (1 - p^3)^{\binom{n}{3}}$$

Proof. For each triple of distinct vertices $i, j, k \in [n]$, the event that ijk does not form a triangle is a decreasing event (here the ground set is the set of edges of the complete graph on n). So by Harris' inequality,

$$\begin{aligned} \mathbb{P}(G(n, p) \text{ is triangle-free}) &\geq \mathbb{P}\left(\bigwedge_{i < j < k} \{ijk \text{ not a triangle}\}\right) \\ &\geq \prod_{i < j < k} \mathbb{P}(ijk \text{ not a triangle}) = (1 - p^3)^{\binom{n}{3}}. \quad \square \end{aligned}$$

Remark 7.2.3. How good is this bound? For $p \leq 0.99$, we have $1 - p^3 = e^{-\Theta(p^3)}$, so the above bound gives

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \geq e^{-\Theta(n^3 p^3)}.$$

Here is another lower bound

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \geq \mathbb{P}(G(n, p) \text{ is empty}) = (1 - p)^{\binom{n}{2}} = e^{-\Theta(n^2 p)}.$$

The bound from Harris is better when $p \ll n^{-1/2}$. Putting them together, we obtain

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \gtrsim \begin{cases} e^{-\Theta(n^3 p^3)} & \text{if } p \lesssim n^{-1/2} \\ e^{-\Theta(n^2 p)} & \text{if } n^{-1/2} \lesssim p \leq 0.99 \end{cases}$$

(note that the asymptotics agree at the boundary $p \asymp n^{-1/2}$). In the next chapter, we will prove matching upper bounds using Janson inequalities.

Maximum degree

Question 7.2.4

What's the probability that the maximum degree of $G(n, 1/2)$ is at most $n/2$?

For each vertex v , $\deg(v) \leq n/2$ is a decreasing event with probability just slightly over $1/2$. So by Harris inequality, the probability that every v has $\deg(v) \leq n/2$ is at least $\geq 2^{-n}$.

It turns out that the appearance of high degree vertices is much more correlated than the independent case. The truth is exponentially more than the above bound.

Theorem 7.2.5 (Riordan and Selby 2000)

$$\mathbb{P}(\max \deg G(n, 1/2) \leq n/2) = (0.6102 \cdots + o(1))^n$$

Instead of giving a proof, we consider an easier continuous model of the problem that motivates the numerical answer. Building on this intuition, [Riordan and Selby \(2000\)](#) proved the result in the random graph setting, although this is beyond the scope of this class.

In a random graphs, we assign independent Bernoulli random variables on edges of a complete graph. Instead, let us assign independent standard normal random variables to each edge of the complete graph.

Proposition 7.2.6 (Max degree with normal random edge labels)

Assign an independent standard normal random variable Z_{uv} to each edge of K_n . Let $W_v = \sum_{u \neq v} Z_{uv}$ be the sum of the labels of the edges incident to a vertex v . Then

$$\mathbb{P}(W_v \leq 0 \ \forall v) = (0.6102 \cdots + o(1))^n$$

The event $W_v \leq 0$ is supposed to model the event that the degree at vertex v is less than $n/2$. Of course, other than intuition, there is no justification here that these two models should behave similarly

We have $\mathbb{P}(W_v \leq 0) = 1/2$. Since each $\{W_v \leq 0\}$ is a decreasing event of the independent edge labels, Harris' inequality tells us that

$$\mathbb{P}(W_v \leq 0 \ \forall v) \geq 2^{-n}.$$

The truth turns out to be significantly greater.

Proof sketch of Proposition 7.2.6. The tuple $(W_v)_{v \in [n]}$ has a joint normal distribution, with each coordinate variance $n - 1$ and pairwise covariance 1. So $(W_v)_{v \in [n]}$ has the

7 Correlation inequalities

same distribution as

$$\sqrt{n-2}(Z'_1, Z'_2, \dots, Z'_n) + Z'_0(1, 1, \dots, 1)$$

where Z'_0, \dots, Z'_n are iid standard normals.

Let Φ be the pdf and cdf of the standard normal $N(0, 1)$.

Thus

$$\mathbb{P}(W_v \leq 0 \ \forall v) = \mathbb{P}\left(Z'_i \leq -\frac{Z'_0}{\sqrt{n-2}} \ \forall i \in [n]\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \Phi\left(\frac{-z}{\sqrt{n-2}}\right)^n dz$$

where the final step is obtained by conditioning on Z'_0 . Substituting $z = y\sqrt{n}$, the above quantity equals to

$$= \sqrt{\frac{n}{2\pi}} \int_{-\infty}^{\infty} e^{nf(y)} dy \quad \text{where} \quad f(y) = -\frac{y^2}{2} + \log \Phi\left(y\sqrt{\frac{n}{n-2}}\right).$$

We can estimate the above integral for large n using the *Laplace method* (which can be justified rigorously by considering Taylor expansion around the maximum of f). We have

$$f(y) \approx g(y) := -\frac{y^2}{2} + \log \Phi(y)$$

and we can deduce that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\max_{v \in [n]} W_v \leq 0) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \int e^{nf(y)} dy = \max g = \log 0.6102 \dots \quad \square$$

8 Janson inequalities

We present a collection of inequalities, known collectively as Janson inequalities (Janson 1990). These tools allow us to estimate **lower tail** large deviation probabilities.

A typical application of Janson's inequality allows us to upper bound the probability that a random graph $G(n, p)$ does not contain any copy of some subgraph. Compared to the second moment method from Chapter 4, Janson inequalities (which is applicable in more limited setups) gives much better bounds, usually with exponential decays.

8.1 Probability of non-existence

The following setup should be a reminiscent of both the second moment method as well as Lovász local lemma (the random variable model).

Setup 8.1.1 (for Janson's inequality: counting containments)

Let R be a random subset of $[N]$ with each element included independently (possibly with different probabilities).

Let $S_1, \dots, S_k \subseteq [N]$. Let A_i be the event that $S_i \subseteq R$. Let

$$X = \sum_i 1_{A_i}$$

be the number of sets S_i contained in the same set R . Let

$$\mu = \mathbb{E}[X] = \sum_i \mathbb{P}(A_i).$$

Write $i \sim j$ if $i \neq j$ and $S_i \cap S_j \neq \emptyset$. Let (as in the second moment method)

$$\Delta = \sum_{(i,j): i \sim j} \mathbb{P}(A_i A_j) = \sum_{(i,j): i \sim j} \mathbb{P}(S_i \cup S_j \subseteq R)$$

(note that (i, j) and (j, i) is each counted once).

The following inequality appeared in Janson, Łuczak, and Ruciński (1990).

Theorem 8.1.2 (Janson inequality I)

Assuming Setup 8.1.1,

$$\mathbb{P}(X = 0) \leq e^{-\mu + \Delta/2}.$$

This inequality is most useful when $\Delta = o(\mu)$.

Remark 8.1.3. When $\mathbb{P}(A_i) = o(1)$ (which is the case in a typical application), Harris' inequality gives us

$$\begin{aligned} \mathbb{P}(X = 0) &= \mathbb{P}(\overline{A_1} \cdots \overline{A_k}) \geq \prod_{i=1}^k \mathbb{P}(\overline{A_i}) \\ &= \prod_{i=1}^k (1 - \mathbb{P}(A_i)) = \exp\left(-\left(1 + o(1)\right) \sum_{i=1}^k \mathbb{P}(A_i)\right) = e^{-(1+o(1))\mu}. \end{aligned}$$

In the setting where $\Delta = o(\mu)$, two bounds match to give $\mathbb{P}(X = 0) = e^{-(1+o(1))\mu}$.

Proof. Let

$$r_i = \mathbb{P}(A_i | \overline{A_1} \cdots \overline{A_{i-1}}).$$

We have

$$\begin{aligned} \mathbb{P}(X = 0) &= \mathbb{P}(\overline{A_1} \cdots \overline{A_k}) \\ &= \mathbb{P}(\overline{A_1}) \mathbb{P}(\overline{A_2} | \overline{A_1}) \cdots \mathbb{P}(\overline{A_k} | \overline{A_1} \cdots \overline{A_{k-1}}) \\ &= (1 - r_1) \cdots (1 - r_k) \\ &\leq e^{-r_1 - \cdots - r_k} \end{aligned}$$

It suffices now to prove that:

Claim. For each $i \in [k]$

$$r_i \geq \mathbb{P}(A_i) - \sum_{j < i: j \sim i} \mathbb{P}(A_i A_j).$$

Summing the claim over $i \in [k]$ would then yield

$$\sum_{i=1}^k r_i \geq \sum_i \mathbb{P}(A_i) - \frac{1}{2} \sum_i \sum_{j \sim i} \mathbb{P}(A_i A_j) = \mu - \frac{\Delta}{2}$$

and thus

$$\mathbb{P}(X = 0) \leq \exp\left(-\sum_i r_i\right) \leq \exp\left(-\mu + \frac{\Delta}{2}\right)$$

Proof of claim. Recall that i is given and fixed. Let

$$D_0 = \bigwedge_{j < i: j \not\sim i} \bar{A}_j \quad \text{and} \quad D_1 = \bigwedge_{j < i: j \sim i} \bar{A}_j$$

Then

$$\begin{aligned} r_i &= \mathbb{P}(A_i | \bar{A}_1 \cdots \bar{A}_{i-1}) = \mathbb{P}(A_i | D_0 D_1) = \frac{\mathbb{P}(A_i D_0 D_1)}{\mathbb{P}(D_0 D_1)} \geq \frac{\mathbb{P}(A_i D_0 D_1)}{\mathbb{P}(D_0)} \\ &= \mathbb{P}(A_i D_1 | D_0) = \mathbb{P}(A_i | D_0) - \mathbb{P}(A_i \bar{D}_1 | D_0) \\ &= \mathbb{P}(A_i) - \mathbb{P}(A_i \bar{D}_1 | D_0) \quad [\text{by independence}] \end{aligned}$$

Since A_i and \bar{D}_1 are both increasing events, and D_0 is a decreasing event, by Harris' inequality (Corollary 7.1.6),

$$\mathbb{P}(A_i \bar{D}_1 | D_0) \leq \mathbb{P}(A_i \bar{D}_1) = \mathbb{P}\left(A_i \wedge \bigvee_{j < i: j \sim i} A_j\right) \leq \sum_{j < i: j \sim i} \mathbb{P}(A_i A_j)$$

This concludes the proof of the claim, and thus the proof of the theorem. \square

Remark 8.1.4 (History). Janson's original proof was via analytic interpolation. The above proof is based on [Boppana and Spencer \(1989\)](#) with a modification by Warnke (personal communication). It has some similarities to the proof of Lovász local lemma from Section 6.1. The above proof incorporates ideas from [Riordan and Warnke \(2015\)](#), who extended Janson's inequality from principal up-set to general up-sets. Indeed, the above proof only requires that the events A_i are increasing, whereas earlier proofs of the result (e.g., the proof in Alon–Spencer) requires the full assumption of Setup 8.1.1, namely that each A_i is an event of the form $S_i \subseteq R_i$ (i.e., a **principal up-set**).

Question 8.1.5

What is the probability that $G(n, p)$ is triangle-free?

In Setup 8.1.1, let $[N]$ with $N = \binom{n}{2}$ be the set of edges of K_n , and let $S_1, \dots, S_{\binom{n}{3}}$ be 3-element sets where each S_i is the edge-set of a triangle. As in the second moment calculation in Section 4.2, we have

$$\mu = \binom{n}{3} p^3 \asymp n^3 p^3 \quad \text{and} \quad \Delta \asymp n^4 p^5.$$

(where Δ is obtained by considering all appearances of a pair of triangles glued along an edge).

8 Janson inequalities

If $p \ll n^{-1/2}$, then $\Delta = o(\mu)$, in which case Janson inequality I (Theorem 8.1.2 and Remark 8.1.3) gives the following.

Theorem 8.1.6

If $p = o(n^{-1/2})$, then

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) = e^{-(1+o(1))\mu} = e^{-(1+o(1))n^3 p^3/6}.$$

Corollary 8.1.7

For a constant $c > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(G(n, c/n) \text{ is triangle-free}) = e^{-c^3/6}.$$

In fact, the number of triangles in $G(n, c/n)$ converges to a Poisson distribution with mean $c^3/6$. On the other hand, when $p \gg 1/n$, the number of triangles is asymptotically normal.

What about if $p \gg n^{-1/2}$, so that $\Delta \gg \mu$. Janson inequality I does not tell us anything nontrivial. Do we still expect the triangle-free probability to be $e^{-(1+o(1))\mu}$, or even $\leq e^{-c\mu}$?

As noted earlier in Remark 7.2.3, another way to obtain a lower bound on the probability triangle-freeness is to consider the probability the $G(n, p)$ is empty (or contained in some fixed complete bipartite graph), in which case we obtain

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \geq (1 - p)^{\Theta(n^2)} = e^{-\Theta(n^2 p)}$$

(the second step assumes that p is bounded away from 1. If $p \gg n^{-1/2}$, so the above lower bound better than the previous one: $e^{-\Theta(n^2 p)} \gg e^{-(1+o(1))\mu}$).

Nevertheless, we'll still use Janson to bootstrap an upper bound on the triangle-free probability. More generally, the next theorem works in the complement region of the Janson inequality I, where now $\Delta \geq \mu$.

Theorem 8.1.8 (Janson inequality II)

Assuming Setup 8.1.1, if $\Delta \geq \mu$, then

$$\mathbb{P}(X = 0) \leq e^{-\mu^2/(2\Delta)}.$$

The proof idea is to applying the first Janson inequality on a randomly sampled subset of events. This sampling technique might remind you of some earlier proofs, e.g., the proof of the crossing number inequality (Theorem 2.6.2), where we first proved a

“cheap bound” that worked in a more limited range, and then used sampling to obtain a better bound.

Proof. For each $T \subseteq [k]$, let $X_T := \sum_{i \in T} 1_{A_i}$ denote the number of occurring events in T . We have

$$\mathbb{P}(X = 0) \leq \mathbb{P}(X_T = 0) \leq e^{-\mu_T + \Delta_T/2}$$

where

$$\mu_T = \sum_{i \in T} \mathbb{P}(A_i)$$

and

$$\Delta_T = \sum_{(i,j) \in T^2: i \sim j} \mathbb{P}(A_i A_j)$$

Choose $T \subseteq [k]$ randomly by including every element with probability $q \in [0, 1]$ independently. We have

$$\mathbb{E}\mu_T = q\mu \quad \text{and} \quad \mathbb{E}\Delta_T = q^2\Delta$$

and so

$$\mathbb{E}(-\mu_T + \Delta_T/2) = -q\mu + q^2\Delta/2.$$

By linearity of expectations, thus there is some choice of $T \subseteq [k]$ so that

$$-\mu_T + \Delta_T/2 \leq -q\mu + q^2\Delta/2$$

so that

$$\mathbb{P}(X = 0) \leq e^{-q\mu + q^2\Delta/2}$$

for every $q \in [0, 1]$. Since $\Delta \geq \mu$, we can set $q = \mu/\Delta \in [0, 1]$ to get the result. \square

To summarize, the first two Janson inequalities tell us that

$$\mathbb{P}(X = 0) \leq \begin{cases} e^{-\mu + \Delta/2} & \text{if } \Delta < \mu \\ e^{-\mu^2/(2\Delta)} & \text{if } \Delta \geq \mu. \end{cases}$$

Remark 8.1.9. If $\mu \rightarrow \infty$ and $\Delta \ll \mu^2$, then Janson inequality II implies $\mathbb{P}(X = 0) = o(1)$, which we knew from second moment method. However Janson’s inequality gives an exponentially decaying tail bound, compared to only a polynomially decaying tail via the second moment method. The exponential tail will be important in an application below to determining the chromatic number of $G(n, 1/2)$.

Let us revisit the example of estimating the probability that $G(n, p)$ is triangle-free,

8 Janson inequalities

now in the regime $p \gg n^{-1/2}$. We have

$$n^3 p^3 \asymp \mu \ll \Delta \asymp n^4 p^5.$$

So so for large enough n , Janson inequality II tells us

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \leq e^{-\mu^2/(2\Delta)} = e^{-\Theta(n^2 p)}$$

Since

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \geq \mathbb{P}(G(n, p) \text{ is empty}) \geq (1 - p)^{\binom{n}{2}} = e^{-\Theta(n^2 p)}$$

where the final step assumes that p is bounded away from 1, we conclude that

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) = e^{-\Theta(n^2 p)}$$

We summarize the results below (strictly speaking we have not yet checked the case $p \asymp n^{-1/2}$, which we can verify by applying Janson inequalities; note that the two regimes below match at the boundary).

Theorem 8.1.10

Suppose $p = p_n \leq 0.99$. Then

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) = \begin{cases} \exp(-\Theta(n^2 p)) & \text{if } p \gtrsim n^{-1/2} \\ \exp(-\Theta(n^3 p^3)) & \text{if } p \lesssim n^{-1/2} \end{cases}$$

Remark 8.1.11. Sharper results are known. Here are some highlights.

1. The number of triangle-free graphs on n vertices is $2^{(1+o(1))n^2/4}$. In fact, an even stronger statement is true: almost all (i.e., $1 - o(1)$ fraction) n -vertex triangle-free graphs are bipartite (Erdős, Kleitman, and Rothschild 1976).
2. If $m \geq Cn^{3/2}\sqrt{\log n}$ for any constant $C > \sqrt{3}/4$ (and this is best possible), then almost all all n -vertex m -edge triangle-free graphs are bipartite (Osthus, Prömel, and Taraz 2003). This result has been extended to K_r -free graphs for every fixed r (Balogh, Morris, Samotij, and Warnke 2016).
3. For $n^{-1/2} \ll p \ll 1$, (Łuczak 2000)

$$-\log \mathbb{P}(G(n, p) \text{ is triangle-free}) \sim -\log \mathbb{P}(G(n, p) \text{ is bipartite}) \sim n^2 p / 4.$$

This result was generalized to general H -free graphs using the powerful recent method of hypergraph containers (Balogh, Morris, and Samotij 2015).

8.2 Lower tails

Previously we looked at the probability of non-existence. Now we would like to estimate lower tail probabilities. Here is a model problem.

Question 8.2.1

Fix a constant $0 < \delta \leq 1$. Let X be the number of triangles of $G(n, p)$. Estimate

$$\mathbb{P}(X \leq (1 - \delta)\mathbb{E}X).$$

We will bootstrap Janson inequality I, $\mathbb{P}(X = 0) \leq \exp(-\mu + \Delta/2)$, to an upper bound on lower tail probabilities.

Theorem 8.2.2 (Janson inequality III)

Assume Setup 8.1.1. For any $0 \leq t \leq \mu$,

$$\mathbb{P}(X \leq \mu - t) \leq \exp\left(\frac{-t^2}{2(\mu + \Delta)}\right)$$

Note that setting $t = \mu$ we basically recover the first two Janson inequalities (up to an unimportant constant factor in the exponent):

$$\mathbb{P}(X = 0) \leq \exp\left(\frac{-\mu^2}{2(\mu + \Delta)}\right). \quad (8.1)$$

(Note that this form of the inequality conveniently captures Janson inequalities I & II.)

Proof. (by Lutz Warnke¹) We start the proof similarly to the proof of the Chernoff bound, by applying Markov's inequality on the moment generating function. To that end, let $\lambda \geq 0$ to be optimized later. Let

$$q = 1 - e^{-\lambda}.$$

By Markov's inequality,

$$\begin{aligned} \mathbb{P}(X \leq \mu - t) &= \mathbb{P}\left(e^{-\lambda X} \geq e^{-\lambda(\mu - t)}\right) \\ &\leq e^{\lambda(\mu - t)} \mathbb{E} e^{-\lambda X} \\ &\leq e^{\lambda(\mu - t)} \mathbb{E}[(1 - q)^X]. \end{aligned}$$

¹Personal communication

8 Janson inequalities

For each $i \in [k]$, let $W_i \sim \text{Bernoulli}(q)$ independently. Consider the random variable

$$Y = \sum_{i=1}^k 1_{A_i} W_i.$$

Conditioned on the value of X , the probability that $Y = 0$ is $(1-q)^X$ (i.e., the probability that $W_i = 0$ for each of the X events A_i that occurred). Taking expectation over X , we have

$$\mathbb{P}(Y = 0) = \mathbb{E}[\mathbb{P}(Y = 0|X)] = \mathbb{E}[(1-q)^X].$$

Note that Y fits within Setup 8.1.1 by introducing k new elements to the ground set $[N]$, where each new element is included according to W_i , and enlarging each S_i to include this new element. The relevant parameters of Y are

$$\mu_Y := \mathbb{E}Y = q\mu$$

and

$$\Delta_Y := \sum_{(i,j): i \sim j} \mathbb{E}[1_{A_i} W_i 1_{A_j} W_j] = q^2 \Delta.$$

Then Janson inequality I applied to Y gives

$$\mathbb{P}(Y = 0) \leq e^{-\mu_Y + \Delta_Y/2} = e^{-q\mu + q^2\Delta/2}.$$

Therefore,

$$\mathbb{E}[(1-q)^X] = \mathbb{P}(Y = 0) \leq e^{-q\mu + q^2\Delta/2}.$$

Continuing the moment calculation at the beginning of the proof, and using that

$$\lambda - \frac{\lambda^2}{2} \leq q \leq \lambda,$$

we have

$$\begin{aligned} \mathbb{P}(X \leq -\mu + t) &\leq e^{\lambda(\mu-t)} \mathbb{E}[(1-q)^X] \\ &\leq \exp\left(\lambda(\mu-t) - q\mu + q^2\Delta/2\right) \\ &\leq \exp\left(\lambda(\mu-t) - \left(\lambda - \frac{\lambda^2}{2}\right)\mu + \lambda^2\frac{\Delta}{2}\right) \\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2}(\mu + \Delta)\right) \end{aligned}$$

We optimize by setting $\lambda = t/(\mu + \Delta)$ to obtain $\leq \exp\left(\frac{-t^2}{2(\mu + \Delta)}\right)$. \square

Example 8.2.3 (Lower tails for triangle counts). Let X be the number of triangles in

$G(n, p)$. We have $\mu \asymp n^3 p^3$ and $\Delta \asymp n^4 p^5$. Fix a constant $\delta \in (0, 1]$. Let $t = \delta \mathbb{E}X$. We have

$$\mathbb{P}(X \leq (1 - \delta)\mathbb{E}X) \leq \exp\left(-\Theta\left(\frac{-\delta^2 n^6 p^6}{n^3 p^3 + n^4 p^5}\right)\right) = \begin{cases} \exp(-\Theta_\delta(n^2 p)) & \text{if } p \gtrsim n^{-1/2}, \\ \exp(-\Theta_\delta(n^3 p^3)) & \text{if } p \lesssim n^{-1/2}. \end{cases}$$

The bounds are tight up to a constant in the exponent, since

$$\mathbb{P}(X \leq (1 - \delta)\mathbb{E}X) \geq \mathbb{P}(X = 0) = \begin{cases} \exp(-\Theta(n^2 p)) & \text{if } p \gtrsim n^{-1/2}, \\ \exp(-\Theta(n^3 p^3)) & \text{if } p \lesssim n^{-1/2}. \end{cases}$$

Example 8.2.4 (No corresponding Janson inequality for upper tails). Continuing with X being the number of triangles of $G(n, p)$, from on the above lower tail results, we might expect $\mathbb{P}(X \geq (1 + \delta)\mathbb{E}X) \leq \exp(-\Theta_\delta(n^2 p))$, but actually this is false!

By planting a clique of size $\Theta(np)$, we can force $X \geq (1 + \delta)\mathbb{E}X$. Thus

$$\mathbb{P}(X \geq (1 + \delta)\mathbb{E}X) \geq p^{\Theta_\delta(n^2 p^2)}$$

which is much bigger than $\exp(-\Theta(n^2 p))$. The above is actually the truth ([Kahn–DeMarco 2012](#) and [Chatterjee 2012](#)):

$$\mathbb{P}(X \geq (1 + \delta)\mathbb{E}X) = p^{\Theta_\delta(n^2 p^2)} \quad \text{if } p \gtrsim \frac{\log n}{n},$$

but the proof is much more intricate. Recent results allow us to understand the exact constant in the exponent though new developments in large deviation theory. The current state of knowledge is summarized below.

Theorem 8.2.5 ([Harel, Mousset, Samotij 2022](#))

Let X be the number of triangles in $G(n, p)$ with $p = p_n$ satisfying $n^{-1/2} \ll p \ll 1$,

$$-\log \mathbb{P}(X \geq (1 + \delta)\mathbb{E}X) \sim \min\left\{\frac{\delta}{3}, \frac{\delta^{2/3}}{2}\right\} n^2 p^2 \log(1/p),$$

and for $n^{-1} \log n \ll p \ll n^{-1/2}$,

$$-\log \mathbb{P}(X \geq (1 + \delta)\mathbb{E}X) \sim \frac{\delta^{2/3}}{2} n^2 p^2 \log(1/p).$$

Remark 8.2.6. The leading constants were determined by [Lubetzky and Zhao \(2017\)](#) by solving an associated variational problem. Earlier results, starting with [Chatterjee and Varadhan \(2011\)](#) and [Chatterjee and Dembo \(2016\)](#) prove large deviation

frameworks that gave the above theorem for sufficiently slowly decaying $p \geq n^{-c}$.

For the corresponding problem for lower tails, see [Kozma and Samotij \(2021+\)](#) for an approach using relative entropy that reduces the rate problem to a variational problem. The exact leading constant is known only for sufficiently small $\delta > 0$, where the answer is given by “replica symmetry”, meaning that the exponential rate is given by a uniform decrement in edge densities for the random graph. In contrast, for δ close to 1, we expect (though cannot prove) that the typical structure of a conditioned random graph is close to a two-block model ([Zhao 2017](#)).

8.3 Chromatic number of a random graph

Question 8.3.1

What is the chromatic number of $G(n, 1/2)$?

In Section 4.4, we used the second moment method to find the clique number ω of $G(n, 1/2)$. We saw that, with probability $1 - o(1)$, the clique number is concentrated on two values, and in particular,

$$\omega(G(n, 1/2)) \sim 2 \log_2 n \quad \text{whp.}$$

The **independence number** $\alpha(G)$ is the size of the largest independent set in G . The independence number $\alpha(G)$ is equal to the clique number of the complement of G . Since $G(n, 1/2)$ and its graph complement have the same distribution, we have $\alpha(G(n, 1/2)) \sim 2 \log_2 n$ whp as well.

Using the following lower bound on the chromatic number $\chi(G)$:

$$\chi(G) \geq \frac{|V(G)|}{\alpha(G)}$$

(since each color class is an independent set), we obtain that

$$\chi(G(n, 1/2)) \geq \frac{(1 + o(1))n}{\log_2 n} \quad \text{whp.}$$

The following landmark theorem shows that the above lower bound on $\chi(G(n, 1/2))$ is asymptotically tight.

Theorem 8.3.2 (Chromatic number of a random graph — Bollobás 1988)

With probability $1 - o(1)$,

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

Recall that $\omega(G(n, 1/2))$ is typically concentrated around the point k where the expected number of k -cliques $\binom{n}{k} 2^{-\binom{k}{2}}$ is neither too large nor too close to zero. The next lemma show that this probability drops very quickly when we decrease k even by a constant.

Lemma 8.3.3

Let $k_0 = k_0(n)$ be the largest possible integer k so that $\binom{n}{k} 2^{-\binom{k}{2}} \geq 1$. Then

$$\mathbb{P}(\alpha(G(n, 1/2)) < k_0 - 3) \leq e^{-n^{2-o(1)}}$$

Note that there is a trivial lower bound of $2^{-\binom{n}{2}}$ coming from an empty graph.

Proof. Let us prove the equivalent claim

$$\mathbb{P}(\omega(G(n, 1/2)) < k_0 - 3) \leq e^{-n^{2-o(1)}}.$$

Let $\mu_k := \binom{n}{k} 2^{-\binom{k}{2}}$. For $k \sim k_0(n) \sim 2 \log_2 n$, we have

$$\frac{\mu_{k+1}}{\mu_k} = \frac{\binom{n}{k+1}}{\binom{n}{k}} 2^{-k} \sim \frac{n}{k} 2^{-(2+o(1)) \log_2 n} = \frac{1}{n^{1-o(1)}}.$$

We have $\mu_{k+1}/\mu_k = n^{-1+o(1)}$ whenever .

Let $k = k_0 - 3$ and applying Setup 8.1.1 for Janson inequality with X being the number of k -cliques, we have

$$\mu = \mu_k > n^{3-o(1)}$$

and (details of the computation omitted)

$$\Delta \sim \mu^2 \frac{k^4}{n^2} = n^{4-o(1)}.$$

So $\Delta > \mu$ for sufficiently large n , and we can apply Janson inequality II:

$$\mathbb{P}(\omega(G(n, 1/2)) < k) = \mathbb{P}(X = 0) \leq e^{-n^{2-o(1)}}.$$

□

Proof of Theorem 8.3.2. The lower bound proof was discussed before the theorem statement. For the upper bound we will give a strategy to properly color the random

8 Janson inequalities

graph with $(2 + o(1)) \log_2 n$ colors. We will proceed by taking out independent sets of size $\sim 2 \log_2 n$ iteratively until $o(n/\log n)$ vertices remain, at which point we can use a different color for each remaining vertex.

Note that after taking out the first independent set of size $\sim 2 \log_2 n$, we cannot claim that the remaining graph is still distributed as $G(n, 1/2)$. It is not. Our selection of the vertices was dependent on the random graph. We are not allowed to “resample” the edges after the first selection.

The strategy is to apply the previous lemma to see that every large enough subset of vertices has an independent set of size $\sim 2 \log_2 n$.

Let $G \sim G(n, 1/2)$. Let $m = \lfloor n/(\log n)^2 \rfloor$, say. For any set S of m vertices, the induced subgraph $G[S]$ has the distribution of $G(m, 1/2)$. By Lemma 8.3.3, for

$$k = k_0(m) - 3 \sim 2 \log_2 m \sim 2 \log_2 n,$$

we have

$$\mathbb{P}(\alpha(G[S]) < k) = e^{-m^{2-o(1)}} = e^{-n^{2-o(1)}}.$$

Taking a union bound over all $\binom{n}{m} < 2^n$ such sets S ,

$$\mathbb{P}(\text{there is an } m\text{-vertex subset } S \text{ with } \alpha(G[S]) < k) < 2^n e^{-n^{2-o(1)}} = o(1).$$

So the following statement is true in $G(n, 1/2)$ with probability $1 - o(1)$:

(*) Every m -vertex subset contains a k -vertex independent set.

Assume that G has property (*). Now we execute our strategy at the beginning of the proof:

1. While $\geq m$ vertices remain:
 - i. Find an independent set of size k , and let it form its own color class
 - ii. Remove these k vertices
2. Color the remaining $< m$ vertices each with a new color.

The result is a proper coloring. The number of colors used is

$$\frac{n}{k} + m \sim \frac{n}{2 \log_2 n}. \quad \square$$

9 Concentration of measure

9.1 Bounded differences inequality

Recall that the Chernoff bound allows to prove exponential tail bounds for sums of **independent** random variables. For example, if Z is a sum of n independent Bernoulli random variables, then

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-2t^2/n}.$$

In this chapter, we develop tools for proving similar tail bounds for other random variables that do not necessarily arise as a sum of independent random variables.

The next theorem says:

A Lipschitz function of many *independent* random variables is concentrated.

We will prove the following important and useful result, known by several names: **McDiarmid's inequality**, **Azuma–Hoeffding inequality**, and **bounded differences inequality**.

Theorem 9.1.1 (Bounded differences inequality)

Let $X_1 \in \Omega_1, \dots, X_n \in \Omega_n$ be **independent** random variables. Suppose $f: \Omega_1 \times \dots \times \Omega_n \rightarrow \mathbb{R}$ satisfies

$$|f(x_1, \dots, x_n) - f(x'_1, \dots, x'_n)| \leq 1 \quad (9.1)$$

whenever (x_1, \dots, x_n) and (x'_1, \dots, x'_n) differ on exactly one coordinate. Then the random variable $Z = f(X_1, \dots, X_n)$ satisfies, for every $\lambda \geq 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq \lambda) \leq e^{-2\lambda^2/n} \quad \text{and} \quad \mathbb{P}(Z - \mathbb{E}Z \leq -\lambda) \leq e^{-2\lambda^2/n}.$$

In particular, we can apply the above inequality to $f(x_1, \dots, x_n) = x_1 + \dots + x_n$ to recover the Chernoff bound. The theorem tells us that the window of fluctuation of Z has length $O(\sqrt{n})$.

Example 9.1.2 (Coupon collector). Let $s_1, \dots, s_n \in [n]$ chosen uniformly and inde-

9 Concentration of measure

pendently at random. Denote the number of “missing” elements by

$$Z = |[n] \setminus \{s_1, \dots, s_n\}|.$$

Note that changing one of the s_1, \dots, s_n changes Z by at most 1, so we have

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \lambda) \leq 2e^{-2\lambda^2/n},$$

with

$$\mathbb{E}Z = n \left(1 - \frac{1}{n}\right)^n \in \left[\frac{n-1}{e}, \frac{n}{e}\right].$$

Theorem 9.1.1 holds more generally allowing the bounded difference to depend on the coordinate.

Theorem 9.1.3 (Bounded differences inequality)

Let $X_1 \in \Omega_1, \dots, X_n \in \Omega_n$ be **independent** random variables. Suppose $f: \Omega_1 \times \dots \times \Omega_n \rightarrow \mathbb{R}$ satisfies

$$|f(x_1, \dots, x_n) - f(x'_1, \dots, x'_n)| \leq c_i \quad (9.2)$$

whenever (x_1, \dots, x_n) and (x'_1, \dots, x'_n) differ only on the i -th coordinate. Here c_1, \dots, c_n are constants. Then the random variable $Z = f(X_1, \dots, X_n)$ satisfies, for every $\lambda \geq 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq \lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \dots + c_n^2}\right)$$

and

$$\mathbb{P}(Z - \mathbb{E}Z \leq -\lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \dots + c_n^2}\right).$$

We will prove these inequality using martingales.

9.2 Martingales concentration inequalities

Definition 9.2.1

A **martingale** is a random real sequence Z_0, Z_1, \dots such that for every Z_n , $\mathbb{E}|Z_n| < \infty$ and

$$\mathbb{E}[Z_{n+1}|Z_0, \dots, Z_n] = Z_n.$$

(To be more formal, we should talk about filtrations of a probability space ...)

Example 9.2.2 (Random walks with independent steps). If $(X_i)_{i \geq 0}$ is a sequence of

independent random variables with $\mathbb{E}X_i = 0$ for all i , then the partial sums $Z_n = \sum_{i \leq n} X_i$ is a Martingale.

Example 9.2.3 (Betting strategy). Betting on a sequence of fair coin tosses. After round, you are allow to change your bet. Let Z_n be your balance after the n -th round. Then Z_n is always a martingale regardless of your strategy.

Originally, the term “martingale” referred to the betting strategy where one doubles the bet each time until the first win and then stop betting. Then, with probability 1, $Z_n = 1$ for all sufficiently large n . (Why does this “free money” strategy not actually work?)

The next example is especially important to us.

Example 9.2.4 (Doob martingale). Let X_1, \dots, X_n be a random sequence (not necessarily independent, though they often are independent in practice). Consider a function $f(X_1, \dots, X_n)$. Let Z_i be the expected value of f after “revealing” (exposing) X_1, \dots, X_i , i.e.,

$$Z_i = \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_i].$$

So Z_i is the expected value of the random variable $Z = f(X_1, \dots, X_n)$ after seeing the first i arguments, and letting the remaining arguments be random. Then Z_0, \dots, Z_n is a martingale (why?). It satisfies $Z_0 = \mathbb{E}Z$ (a non-random quantity) and $Z_n = Z$ (the random variable that we care about), and thereby offering a way to interpolate between the two.

Example 9.2.5 (Edge-exposure martingale). We can reveal the random graph $G(n, p)$ by first fixing an order on all unordered pairs of $[n]$ and then revealing in order whether each pair is an edge. For any graph parameter $f(G)$ we can produce a martingale $X_0, X_1, \dots, X_{\binom{n}{2}}$ where Z_i is the conditional expectation of $f(G(n, p))$ after revealing whether there are edges for first i pairs of vertices. See Figure 9.1 for an example.

Example 9.2.6 (Vertex-exposure martingale). Similar to the previous example, except that we now first fix an order on the vertex set, and, at the i -th step, with $0 \leq i \leq n$, we reveal all edges whose endpoints are contained in the first i vertices. See Figure 9.1 for an example.

Sometimes it is better to use the edge-exposure martingale and sometimes it is better to use the vertex-exposure martingale. It depends on the application. There is a trade-off between the length of the martingale and the control on the bounded differences.

The main result is that a martingale with *bounded differences* must be concentrated. The following fundamental result is called Azuma’s inequality or the Azuma–Hoeffding inequality.

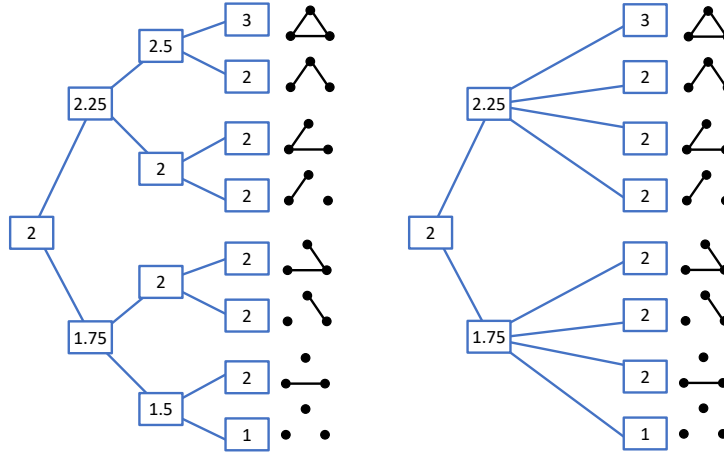


Figure 9.1: The edge-exposure martingale (left) and vertex-exposure martingale (right) for the chromatic number of $G(n, 1/2)$ with $n = 3$. The martingale is obtained by starting at the leftmost point, and splitting at each branch with equal probability.

Theorem 9.2.7 (Azuma's inequality)

Let Z_0, Z_1, \dots, Z_n be a martingale satisfying

$$|Z_i - Z_{i-1}| \leq 1 \quad \text{for each } i \in [n].$$

Then for every $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \geq \lambda\sqrt{n}) \leq e^{-\lambda^2/2}.$$

Note that this is the same bound that we derived in Chapter 5 for $Z_n = X_1 + \dots + X_n$ where $X_i \in \{-1, 1\}$ uniform and iid.

More generally, allowing different bounds on different steps of the martingale, we have the following.

Theorem 9.2.8 (Azuma's inequality)

Let Z_0, Z_1, \dots, Z_n be a martingale satisfying

$$|Z_i - Z_{i-1}| \leq c_i \quad \text{for each } i \in [n].$$

For any $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-\lambda^2}{2(c_1^2 + \dots + c_n^2)}\right).$$

The above formulations of Azuma's inequality can be used to recover the bounded differences inequality (Theorems 9.1.1 and 9.1.3) up to a usually unimportant constant in the exponent. To obtain the exact statement of Theorem 9.1.3, we state the following

strengthening of Azuma's inequality. (You are welcome to ignore the next statement if you do not care about the constant factor in the exponent — and really, you should not care.)

Theorem 9.2.9 (Azuma's inequality for Doob martingales)

Consider a Doob martingale $Z_i = \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_i]$ as in Example 9.2.4. Suppose, conditioned on any value of (X_1, \dots, X_{i-1}) , the possibilities for Z_i lies in an interval of length c_i (here c_i is non-random, but the location of the interval may depend on X_1, \dots, X_{i-1}). Then for any $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \dots + c_n^2}\right).$$

Remark 9.2.10. Applying the inequality to the martingale with terms $-Z_n$, we obtain the following lower tail bound:

$$\mathbb{P}(Z_n - Z_0 \leq -\lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \dots + c_n^2}\right).$$

And we can put them together as

$$\mathbb{P}(|Z_n - Z_0| \geq \lambda) \leq 2 \exp\left(\frac{-2\lambda^2}{c_1^2 + \dots + c_n^2}\right).$$

Remark 9.2.11. Theorem 9.2.8 is a special case of Theorem 9.2.9, since we can take $(X_1, \dots, X_n) = (Z_1, \dots, Z_n)$ and $f(X_1, \dots, X_n) = X_n$. Note that the $|Z_i - Z_{i-1}| \leq c_i$ condition in Theorem 9.2.8 implies that Z_i lies in an interval of length $2c_i$ if we condition on (X_1, \dots, X_{i-1}) .

Lemma 9.2.12 (Hoeffding's lemma)

Let X be a real random variable contained in an interval of length ℓ . Suppose $\mathbb{E}X = 0$. Then

$$\mathbb{E}[e^X] \leq e^{\ell^2/8}.$$

Proof. Suppose $X \in [a, b]$ with $a \leq 0 \leq b$ and $b - a = \ell$. Then since e^x is convex, using a linear upper bound on the interval $[a, b]$, we have (note that RHS below is linear in x)

$$e^x \leq \frac{b-x}{b-a}e^a + \frac{x-a}{b-a}e^b, \quad \text{for all } x \in [a, b].$$

9 Concentration of measure

Since $\mathbb{E}X = 0$, we obtain

$$\mathbb{E}e^X \leq \frac{b}{b-a}e^a + \frac{-a}{b-a}e^b.$$

Let $p = -a/(b-a)$. Then $a = -p\ell$ and $b = (1-p)\ell$. So

$$\log \mathbb{E}e^X \leq \log \left((1-p)e^{-p\ell} + pe^{(1-p)\ell} \right) = -p\ell + \log(1-p+pe^\ell).$$

Fix $p \in [0, 1]$. Let

$$\varphi(\ell) := -p\ell + \log(1-p+pe^\ell).$$

It remains to show that $\varphi(\ell) \leq \ell^2/8$ for all $\ell \geq 0$, which follows from $\varphi(0) = \varphi'(0) = 0$ and $\varphi''(\ell) \leq 1/4$ for all $\ell \geq 0$, as

$$\varphi''(\ell) = \left(\frac{p}{(1-p)e^{-p\ell} + p} \right) \left(1 - \frac{p}{(1-p)e^{-p\ell} + p} \right) \leq \frac{1}{4},$$

since $t(1-t) \leq 1/4$ for all $t \in [0, 1]$. □

Proof of Theorem 9.2.9. Let $t \geq 0$ be some constant to be decided later. Conditional on any values of (X_1, \dots, X_{i-1}) , the random variable $Z_i - Z_{i-1}$ has mean zero and lies in an interval of length c_i . So Lemma 9.2.12 gives

$$\mathbb{E}[e^{t(Z_i - Z_{i-1})} | X_1, \dots, X_{i-1}] \leq e^{t^2 c_i^2 / 8}.$$

Then the moment generating function satisfies

$$\begin{aligned} \mathbb{E}[e^{t(Z_n - Z_0)}] &= \mathbb{E} \left[e^{t(Z_i - Z_{i-1})} e^{t(Z_{i-1} - Z_0)} \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[e^{t(Z_i - Z_{i-1})} \mid X_1, \dots, X_{i-1} \right] e^{t(Z_{i-1} - Z_0)} \right] \\ &= e^{t^2 c_i^2 / 8} \mathbb{E} \left[e^{t(Z_{i-1} - Z_0)} \right]. \end{aligned}$$

Iterating, we obtain

$$\mathbb{E} \left[e^{t(Z_n - Z_0)} \right] \leq e^{t^2 (c_1^2 + \dots + c_n^2) / 8}.$$

By Markov,

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) \leq e^{-t\lambda} \mathbb{E} \left[e^{t(Z_n - Z_0)} \right] \leq e^{-t\lambda + \frac{t^2}{8} (c_1^2 + \dots + c_n^2)}.$$

Setting $t = 4\lambda / (c_1^2 + \dots + c_n^2)$ yields the theorem. □

Now we apply Azuma's inequality to deduce the bounded differences inequality.

Proof of the bounded differences inequality (Theorem 9.1.3). Consider the Doob mar-

tingale $Z_i = \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_i]$. The hypothesis of Theorem 9.1.3 implies that the hypothesis of Theorem 9.2.9 is satisfied. The same conclusion then follows. \square

Remark 9.2.13. Azuma's inequality (Theorem 9.2.9) is more versatile than (Theorem 9.1.3). For example, while changing X_i might change $f(X_1, \dots, X_n)$ by a lot in the worst case over all possible (X_1, \dots, X_n) , it might not change it by much in expectation over random choices of (X_{i+1}, \dots, X_n) . And so the c_i in Theorem 9.2.9 could potentially be smaller than in Theorem 9.1.3. This will be useful in some applications, including one that we will see later in the chapter.

9.3 Chromatic number of random graphs

Concentration of the chromatic number

Even before Bollobás (1988) showed that $\chi(G(n, 1/2)) \sim \frac{n}{2 \log_2 n}$ whp (Theorem 8.3.2), using the bounded difference inequality, it was already known that the chromatic number of a random graph must be concentrated in a $O(\sqrt{n})$ window around its mean. The following application shows that one can prove concentration around the mean without even knowing where is the mean!

Theorem 9.3.1 (Shamir and Spencer 1987)

For every $\lambda \geq 0$, $Z = \chi(G(n, p))$ satisfies

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq \lambda \sqrt{n-1}) \leq 2e^{-2\lambda^2}.$$

Proof. Let $V = [n]$, and consider each vertex labeled graph as an element of $\Omega_2 \times \dots \times \Omega_n$ where $\Omega_i = \{0, 1\}^{i-1}$ and its coordinates correspond to edges whose larger coordinate is i (cf. the vertex-exposure martingale Example 9.2.6). If two graphs G and G' differ only in edges incident to one vertex v , then $|\chi(G) - \chi(G')| \leq 1$ since, given a proper coloring of G using $\chi(G)$ colors, one can obtain a proper coloring of G' using $\chi(G) + 1$ colors by using a new color for v . Theorem 9.1.3 implies the result. \square

Remark 9.3.2 (Non-concentration of the chromatic number). Heckel (2021) showed that the $\chi(G(n, 1/2))$ is *not* concentrated on any interval of length n^c for any constant $c < 1/4$. This was the opposite of what most experts believed in. It has been conjectured that width of the window of concentrations fluctuates between $n^{1/4+o(1)}$ to $n^{1/2+o(1)}$ depending on n .

Clique number, again

Previously in Section 8.3, we used Janson inequalities to prove the following exponentially small bound on the probability that $G(n, 1/2)$ has small clique number. This was a crucial step in the proof of Bollobás' theorem (Theorem 8.3.2) that $\chi(G(n, 1/2)) \sim n/(2 \log_2 n)$ whp. Here we give a different proof using the bounded difference inequality instead of Janson inequalities. The proof below in fact was the original approach of Bollobás (1988).

Lemma 9.3.3 (Same as Lemma 8.3.3)

Let $k_0 = k_0(n) \sim 2 \log_2 n$ be the largest positive integer so that $\binom{n}{k_0} 2^{-\binom{k_0}{2}} \geq 1$. Then

$$\mathbb{P}(\omega(G(n, 1/2)) < k_0 - 3) = e^{-n^{2-o(1)}}.$$

A naive approach might be to estimate the number of k -cliques in G (this is the approach taken with Janson inequalities. The issue is that this quantity can change too much when we modify one edge of G . We will use a more subtle function on graphs. Note that we only care about whether there exists a k -clique or not.

Proof. Let $k = k_0 - 3$. Let $Y = Y(G)$ be the maximum number of edge-disjoint set of k -cliques in G . Then as a function of G , Y changes by at most 1 if we change G by one edge. (Note that the same does not hold if we change G by one vertex, e.g., when G consists of many k -cliques glued along a common vertex.)

So by the bounded differences inequality, for $G \sim G(n, 1/2)$,

$$\mathbb{P}(\omega(G) < k) = \mathbb{P}(Y = 0) \leq \mathbb{P}(Y - \mathbb{E}Y \leq -\mathbb{E}Y) \leq \exp\left(-\frac{2(\mathbb{E}Y)^2}{\binom{n}{2}}\right). \quad (9.1)$$

It remains to show that $\mathbb{E}Y \geq n^{2-o(1)}$. Create an auxiliary graph \mathcal{H} whose vertices are the k -cliques in G , with a pair of k -cliques adjacent if they overlap in at least 2 vertices. Then $Y = \alpha(\mathcal{H})$. We would like to lower bound the independence number of this graph based on its average degree. Here are two ways to proceed:

1. Recall the Caro–Wei inequality (Corollary 2.3.5): for every graph H with average degree \bar{d} , we have

$$\alpha(H) \geq \sum_{v \in V(H)} \frac{1}{1 + d_v} \geq \frac{|V(H)|}{1 + \bar{d}} = \frac{|V(H)|^2}{|V(H)| + 2|E(H)|}.$$

2. Let H' be the induced subgraph obtained from H by keeping every vertex

independently with probability q . We have

$$\alpha(H) \geq \alpha(H') \geq |V(H')| - |E(H')|.$$

Taking expectations of both sides, and noting that $\mathbb{E}|V(H')| = q|V(H)|$ and $\mathbb{E}|E(H')| = q^2|E(H)|$ by linearity of expectations, we have

$$\alpha(H) \geq q\mathbb{E}|V(H)| - q^2|E(H)| \quad \text{for every } q \in [0, 1].$$

Provided that $|E(H)| \geq |V(H)|/2$, we can take $q = |V(H)|/(2|E(H)|) \in [0, 1]$ and obtain

$$\alpha(H) \geq \frac{|V(H)|^2}{4|E(H)|} \quad \text{if } |E(H)| \geq \frac{1}{2}|V(H)|.$$

(This method allows us to recover Turán's theorem up to a factor of 2, whereas the Caro–Wei inequality recovers Turán's theorem exactly. For the present application, we do not care about these constant factors.)

By a second moment argument (details again omitted; similar to Section 4.4 and ??), we have, with probability $1 - o(1)$, that the number of k -cliques in G is

$$|V(\mathcal{H})| \sim \mathbb{E}|V(\mathcal{H})| = \binom{n}{k} 2^{-\binom{k}{2}} = n^{3-o(1)}$$

and the number of unordered pairs of edge-overlapping k -cliques in G is

$$\mathbb{E}|E(\mathcal{H})| = n^{4-o(1)}.$$

Thus, with probability $1 - o(1)$, we can apply either of the above lower bounds on independent sets to obtain

$$\mathbb{E}Y \gtrsim \mathbb{E} \frac{|V(\mathcal{H})|^2}{|E(\mathcal{H})|} \gtrsim \mathbb{E} \frac{n^{6-o(1)}}{|E(\mathcal{H})|} \geq \frac{n^{6-o(1)}}{\mathbb{E}|E(\mathcal{H})|} = n^{2-o(1)}.$$

Together with (9.1), this completes the proof that $\mathbb{P}(\omega(G) < k) = e^{-n^{2-o(1)}}$. \square

Chromatic number of sparse random graphs

Let us show that $G(n, p)$ is concentrated on a constant size window if p is small enough.

Theorem 9.3.4 (Shamir and Spencer 1987)

Let $\alpha > 5/6$ be fixed. Then for $p < n^{-\alpha}$, $\chi(G(n, p))$ is concentrated on four values with probability $1 - o(1)$. That is, there exists $u = u(n, p)$ such that, as $n \rightarrow \infty$,

$$\mathbb{P}(u \leq \chi(G(n, p)) \leq u + 3) = 1 - o(1).$$

Proof. It suffices to show that for all $\varepsilon > 0$, there exists $u = u(n, p, \varepsilon)$ so that, provided $p < n^{-\alpha}$ and n is sufficiently large,

$$\mathbb{P}(u \leq \chi(G(n, p)) \leq u + 3) \geq 1 - 3\varepsilon.$$

Let u be the least integer so that

$$\mathbb{P}(\chi(G(n, p)) \leq u) > \varepsilon.$$

Now we make a clever choice of a random variable.

Let $G \sim G(n, p)$. Let $Y = Y(G)$ denote the minimum size of a subset $S \subseteq V(G)$ such that $G - S$ is u -colorable.

Note that Y changes by at most 1 if we change the edges around one vertex of G . Thus, by applying Theorem 9.1.1 with respect to vertex-exposure (Example 9.2.6), we have

$$\begin{aligned} \mathbb{P}(Y \leq \mathbb{E}Y - \lambda\sqrt{n}) &\leq e^{-2\lambda^2} \\ \text{and} \quad \mathbb{P}(Y \geq \mathbb{E}Y + \lambda\sqrt{n}) &\leq e^{-2\lambda^2}. \end{aligned}$$

We choose $\lambda = \lambda(\varepsilon) > 0$ so that $e^{-2\lambda^2} = \varepsilon$.

First, we use the lower tail bound to show that $\mathbb{E}Y$ must be small. We have

$$e^{-2\lambda^2} = \varepsilon < \mathbb{P}(\chi(G) \leq u) = \mathbb{P}(Y = 0) = \mathbb{P}(Y \leq \mathbb{E}Y - \mathbb{E}Y) \leq \exp\left(\frac{-2(\mathbb{E}Y)^2}{n}\right).$$

Thus

$$\mathbb{E}Y \leq \lambda\sqrt{n}.$$

Next, we apply the upper tail bound to show that Y is rarely large. We have

$$\mathbb{P}(Y \geq 2\lambda\sqrt{n}) \leq \mathbb{P}(Y \geq \mathbb{E}Y + \lambda\sqrt{n}) \leq e^{-2\lambda^2} = \varepsilon.$$

Each of the following three events occur with probability at least $1 - \varepsilon$, for large enough n ,

- By the above argument, there is some $S \subseteq V(G)$ with $|S| \leq 2\lambda\sqrt{n}$ and $G - S$

may be properly u -colored.

- By the next lemma, one can properly 3-color $G[S]$.
- $\chi(G) \geq u$ (by the minimality of u at the beginning of the proof).

Thus, with probability at least $1 - 3\varepsilon$, all three events occur, and so we have $u \leq \chi(G) \leq u + 3$. \square

Lemma 9.3.5

Fix $\alpha > 5/6$ and C . Let $p \leq n^{-\alpha}$. Then with probability $1 - o(1)$ every subset of at most $C\sqrt{n}$ vertices of $G(n, p)$ can be properly 3-colored.

Proof. Let $G \sim G(n, p)$. Assume that G is not 3-colorable. Choose minimum size $T \subseteq V(G)$ so that the induced subgraph $G[T]$ is not 3-colorable.

We see that $G[T]$ has minimum degree at least 3, since if $\deg_{G[T]}(x) < 3$, then $T - x$ cannot be 3-colorable either (if it were, then can extend coloring to x), contradicting the minimality of T .

Thus $G[T]$ has at least $3|T|/2$ edges. The probability that G has some induced subgraph on $t \leq C\sqrt{n}$ vertices and $\geq 3t/2$ edges is, by a union bound, (recall $\binom{n}{k} \leq (ne/k)^k$)

$$\begin{aligned} &\leq \sum_{t=4}^{C\sqrt{n}} \binom{n}{t} \binom{\binom{t}{2}}{3t/2} p^{3t/2} \leq \sum_{t=4}^{C\sqrt{n}} \left(\frac{ne}{t}\right)^t \left(\frac{te}{3}\right)^{3t/2} n^{-3t\alpha/2} \\ &\leq \sum_{t=4}^{C\sqrt{n}} \left(O(n^{1-3\alpha/2}\sqrt{t})\right)^t \leq \sum_{t=4}^{C\sqrt{n}} \left(O(n^{1-3\alpha/2+1/4})\right)^t. \end{aligned}$$

The sum is $o(1)$ provided that $\alpha > 5/6$. \square

Remark 9.3.6. Theorem 9.3.4 was subsequently improved (by a refinement of the above techniques) by [Łuczak \(1991\)](#) and [Alon and Krivelevich \(1997\)](#). We now know that the chromatic number of $G(n, n^{-\alpha})$ has two-point concentration for all $\alpha > 1/2$.

9.4 Isoperimetric inequalities: a geometric perspective

We shall explore the following connection, which are two sides of the same coin:

<i>Probability</i>	<i>Geometry</i>
Concentration of Lipschitz functions	Isoperimetric inequalities

Milman recognized the importance of the **concentration of measure phenomenon**, which he heavily promoted in the 1970's. The subject has been since then extensively developed. It plays a central role in probability theory, the analysis of Banach spaces, and it also has been influential in theoretical computer science.

Euclidean space

The classic isoperimetric theorem in \mathbb{R}^n says that among all subset of \mathbb{R}^n of given volume, the ball has the smallest surface volume. (The word “isoperimetric” refers to fixing the perimeter; equivalently we fix the surface area and ask to maximize volume.) This result (at least in two-dimensions) was known to the Greeks, but rigorous proofs were only found in towards the end of the nineteenth century.

Let (X, d_X) be a metric space. Let $A \subseteq X$. For any $x \in X$, write $d_X(x, A) := \inf_{a \in A} d_X(x, a)$ for the distance from x to A . Denote the set of all points within distance t from A by

$$A_t := \{x \in X : d_X(x, A) \leq t\} \quad (9.1)$$

This is also known as the **radius- t neighborhood of A** . One can visualize A_t by “expanding” A by distance t .

Theorem 9.4.1 (Isoperimetric inequality in Euclidean space)

Let $A \subseteq \mathbb{R}^n$ be a measurable set, and let $B \subseteq \mathbb{R}^n$ be a ball $\text{vol}(A) = \text{vol}(B)$. Then, for all $t \geq 0$,

$$\text{vol } A_t \geq \text{vol } B_t.$$

Remark 9.4.2. A clean way to prove the above inequality is via the Brunn–Minkowski theorem.

Classically, the isoperimetric inequality is stated as (here ∂A is the boundary of A)

$$\text{vol}_{n-1} \partial A \geq \text{vol}_{n-1} \partial B.$$

These two formulations are equivalent. Indeed, assuming Theorem 9.4.1, we have

$$\begin{aligned} \text{vol}_{n-1} \partial A &= \left. \frac{d}{dt} \right|_{t=0} \text{vol}_n A_t = \lim_{t \rightarrow 0} \frac{\text{vol } A_t - \text{vol } A}{t} \\ &\geq \lim_{t \rightarrow 0} \frac{\text{vol } B_t - \text{vol } B}{t} = \text{vol}_{n-1} \partial B. \end{aligned}$$

Conversely, we can obtain the neighborhood version from the boundary version by integrating (noting that B_t is always a ball).

The cube

We have an analogous result in the $\{0, 1\}^n$ with respect to Hamming distance. In Hamming cube, **Harper's theorem** gives the exact result. Below, for $A \subseteq \{0, 1\}^n$, we write A_t as in (9.1) for $X = \{0, 1\}^n$ and d_X being the Hamming distance.

Theorem 9.4.3 (Isoperimetric inequality in the Hamming cube; Harper 1966)

Let $A \subseteq \{0, 1\}^n$. Let $B \subseteq \{0, 1\}^n$ be a Hamming ball with $|A| \geq |B|$. Then for all $t \geq 0$,

$$|A_t| \geq |B_t|.$$

Remark 9.4.4. The above statement is tight when A has the same size as a Hamming ball, i.e., when $|A| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k}$ for some integer k . Actually, more is true. For any value of $|A|$ and t , the size of A_t is minimized by taking A to be an initial segment of $\{0, 1\}^n$ according to the *simplicial ordering*: first sort by Hamming weight, and for ties, sort by lexicographic order. For more on this topic, particularly extremal set theory, see the book *Combinatorics* by Bollobás (1986).

Combined with the isoperimetric inequality on the cube, we obtain the following surprising consequence. Suppose we start with just half of the cube, and then expand it by a bit (recall that the diameter of the cube is n , and we will be expanding it by $o(n)$), then resulting expansion occupies nearly all of the cube.

Theorem 9.4.5 (Rapid expansion from half to $1 - \varepsilon$)

Let $t > 0$. For every $A \subseteq \{0, 1\}^n$ with $|A| \geq 2^{n-1}$, we have

$$|A_t| > (1 - e^{-2t^2/n})2^n.$$

Proof. Let $B = \{x \in \{0, 1\}^n : \text{weight}(x) < n/2\}$, so that $|B| \leq 2^{n-1} \leq |A|$. Then by Harper's theorem (Theorem 9.4.3),

$$|A_t| \geq |B_t| = |\{x \in \{0, 1\}^n : \text{weight}(x) < n/2 + t\}| > (1 - e^{-2t^2/n})2^n$$

by the Chernoff bound. □

In fact, using the above, we can deduce that even if we start with a small fraction (e.g., 1%) of the cube, and expand it slightly, then we would cover most of the cube.

Theorem 9.4.6 (Rapid expansion from ε to $1 - \varepsilon$)

Let $\varepsilon > 0$ and $C = \sqrt{2 \log(1/(\varepsilon))}$. If $A \subseteq \{0, 1\}^n$ with $|A| \geq \varepsilon 2^n$, then

$$|A_{C\sqrt{n}}| \geq (1 - \varepsilon)2^n.$$

First proof via Harper's isoperimetric inequality. Let $t = \sqrt{\log(1/\varepsilon)n/2}$ so that $e^{-2t^2/n} = \varepsilon$. Applying Theorem 9.4.5 to $A' = \{0, 1\}^n \setminus A_t$, we see that $|A'| < 2^{n-1}$ (or else $|A'_t| > (1 - \varepsilon)2^n$, so A'_t would intersect A , which is impossible since the distance between A and A' is greater than t). Thus $|A_t| \geq 2^{n-1}$, and then applying Theorem 9.4.5 yields $|A_{2t}| \geq (1 - \varepsilon)2^n$. \square

Let us give another proof of Theorem 9.4.6 without using Harper's exact isoperimetric theorem in the Hamming cube, and instead use the bounded differences inequality that we proved earlier.

Second proof via the bounded differences inequality. Pick a uniform random $x \in \{0, 1\}^n$ and let $X = \text{dist}(x, A)$. Note that X changes by at most 1 if a single coordinate of x is changed. Applying the bounded differences inequality, Theorem 9.1.1, we have the lower tail

$$\mathbb{P}(X - \mathbb{E}X \leq -\lambda) \leq e^{-2\lambda^2/n} \quad \text{for all } \lambda \geq 0$$

We have $X = 0$ if and only if $x \in A$, so

$$\varepsilon \leq \mathbb{P}(x \in A) = \mathbb{P}(X = 0) = \mathbb{P}(X - \mathbb{E}X \leq -\mathbb{E}X) \leq e^{-2(\mathbb{E}X)^2/n}.$$

Thus

$$\mathbb{E}X \leq \sqrt{\frac{\log(1/\varepsilon)n}{2}} = \frac{C\sqrt{n}}{2}.$$

Now we apply the upper tail of the bounded differences inequality

$$\mathbb{P}(X - \mathbb{E}X \geq \lambda) \leq e^{-2\lambda^2/n} \quad \text{for all } \lambda \geq 0$$

to yield

$$\mathbb{P}(x \notin A_{C\sqrt{n}}) = \mathbb{P}(X > C\sqrt{n}) \leq \mathbb{P}\left(X \geq \mathbb{E}X + \frac{C\sqrt{n}}{2}\right) \leq \varepsilon. \quad \square$$

Isoperimetry versus concentration

The above two proofs illustrate the link between geometric isoperimetric inequalities and probabilistic concentration inequalities. Let now state a simple result that

formalizes this connection.

Definition 9.4.7 (Lipschitz functions)

Given two metric spaces (X, d_X) and (Y, d_Y) , we say that a function $f: X \rightarrow Y$ is ***C-Lipschitz*** if

$$d_Y(f(x), f(x')) \leq C d_X(x, x') \quad \text{for all } x, x' \in X.$$

So the bounded differences inequality applies to Lipschitz functions with respect to the Hamming distance. In particular, it tells us that if $f: \{0, 1\}^n \rightarrow \mathbb{R}$ is 1-Lipschitz (with respect to the Hamming distance on $\{0, 1\}^n$), it must be concentrated around its mean with respect to the uniform measure on $\{0, 1\}^n$:

$$\mathbb{P}(|f - \mathbb{E}f| \geq n\lambda) \leq 2e^{-2n\lambda^2}.$$

So f is *almost constant almost everywhere*. This is a counterintuitive high dimensional phenomenon.

Theorem 9.4.8 (Equivalence between notions of concentration of measure)

Let $t, \varepsilon \geq 0$. In a probability space (Ω, \mathbb{P}) equipped with a metric. The following are equivalent:

- (a) (Expansion/approximate isoperimetry) If $A \subseteq \Omega$ with $\mathbb{P}(A) \geq 1/2$, then

$$\mathbb{P}(A_t) \geq 1 - \varepsilon.$$

- (b) (Concentration of Lipschitz functions) If $f: \Omega \rightarrow \mathbb{R}$ is 1-Lipschitz and $m \in \mathbb{R}$ satisfies $\mathbb{P}(f \leq m) \geq 1/2$, then

$$\mathbb{P}(f > m + t) \leq \varepsilon.$$

Remark 9.4.9 (Median). In (b), we often take m to be a ***median*** of f , which is defined to be a value such that $\mathbb{P}(f \geq m) \geq 1/2$ and $\mathbb{P}(f \leq m) \geq 1/2$ (the median always exists but is not necessarily unique). For distributions with good concentration properties, the median and mean are usually close to each other. For example, we leave it as an exercise to check that if there is some m such that $\mathbb{P}(|f - m| \geq t) \leq 2e^{-t^2/2}$ for all $t \geq 0$, then the mean and the medians of f all lie within $O(1)$ of m .

Proof. (a) \implies (b): Let $A = \{x \in \Omega : f(x) \leq m\}$. So $\mathbb{P}(A) \geq 1/2$. Since f is 1-Lipschitz, we have $f(x) \leq m + t$ for all $x \in A_t$. Thus by (a)

$$\mathbb{P}(f > m + t) \leq \mathbb{P}(\overline{A_t}) \leq \varepsilon.$$

9 Concentration of measure

(b) \implies (a): Let $f(x) = \text{dist}(x, A)$ and $m = 0$. Then $\mathbb{P}(f \leq 0) = \mathbb{P}(A) \geq 1/2$. Also f is 1-Lipschitz. So by (b),

$$\mathbb{P}(\overline{A_t}) = \mathbb{P}(f > m + t) \leq \varepsilon. \quad \square$$

Informally, we say that a space (or rather, a sequence of spaces), has concentration of measure if ε decays rapidly as a function of t in the above theorem (the notion of “Lévy family” makes this precise). Earlier we saw that the Hamming cube exhibits has concentration of measure. Other notable spaces with concentration of measure include the sphere, Gauss space, orthogonal and unitary groups, positively-curved manifolds, and the symmetric group.

The sphere

We discuss analogs of the concentration of measure phenomenon in high dimensional geometry. This is rich and beautiful subject. An excellent introductory to this topic is the survey *An Elementary Introduction to Modern Convex Geometry* by [Ball \(1997\)](#).

Recall the isoperimetric inequality in \mathbb{R}^n says:

If $A \subseteq \mathbb{R}^n$ has the same measure as ball B , then $\text{vol}(A_t) \geq \text{vol}(B_t)$ for all $t \geq 0$.

Analogous exact isoperimetric inequalities are known in several other spaces. We already saw it for the boolean cube (Theorem 9.4.3). The case of sphere and Gaussian space are particularly noteworthy. The following theorem is due to Lévy (~1919).

Theorem 9.4.10 (Lévy’s isoperimetric inequality on the sphere)

On a sphere in \mathbb{R}^n , let A be a measurable subset and B a spherical cap with $\text{vol}_{n-1}(A) = \text{vol}_{n-1}(B)$. Then for all $t \geq 0$,

$$\text{vol}_{n-1}(A_t) \geq \text{vol}_{n-1}(B_t).$$

We have the following upper bound estimate on the size of spherical caps.

Theorem 9.4.11 (Upper bound on spherical cap size)

Let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ be a uniform random unit vector in \mathbb{R}^n . Then for any $\varepsilon \geq 0$,

$$\mathbb{P}(x_1 \geq \varepsilon) \leq e^{-n\varepsilon^2/2}.$$

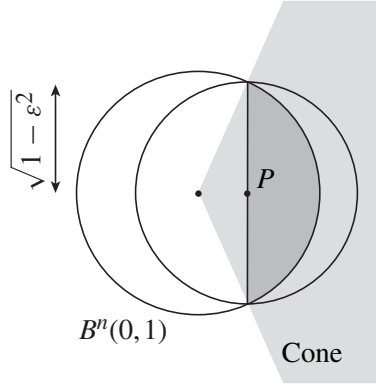
The following proof (including figures) is taken from [Tokz \(2012\)](#), building on the method by [Ball \(1997\)](#).

9.4 Isoperimetric inequalities: a geometric perspective

Proof. Let C denote the spherical cap consisting of unit vectors x with $x_1 \geq \varepsilon$. Write \tilde{C} for the convex hull of C with the origin, i.e., the conical sector spanned by C . The idea is to contain \tilde{C} in a ball of radius $r \leq e^{-\varepsilon^2/2}$. Writing $B(r)$ for a ball of radius r in \mathbb{R}^n so that, we have

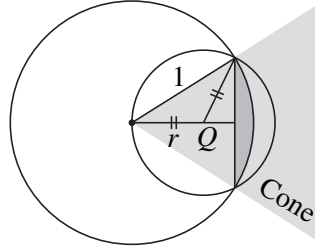
$$\frac{\text{vol}_{n-1} C}{\text{vol}_{n-1} S^{n-1}} = \frac{\text{vol}_n \tilde{C}}{\text{vol}_n B_n(1)} = \frac{\text{vol}_n B(r)}{\text{vol}_n B(1)} = r^n \leq e^{-\varepsilon^2 n/2}.$$

Case 1: $\varepsilon \in [0, 1/\sqrt{2}]$.



As shown above, \tilde{C} is contained in a ball of radius $r = \sqrt{1-\varepsilon^2} \leq e^{-\varepsilon^2/2}$.

Case 2: $\varepsilon \in [1/\sqrt{2}, 1]$.



Then \tilde{C} is contained in a ball of radius r as shown above. Using similar triangles, we find that $r/(1/2) = 1/\varepsilon$. So $r = 1/(2\varepsilon) \leq e^{-\varepsilon^2/2}$, where final inequality is equivalent to $e^{x^2/2} \leq 2x$ for all $[1/\sqrt{2}, 1]$, which, by convexity, only needs to be checked at the endpoints of the interval. \square

Combining the above two theorems, we deduce the following concentration of measure results.

Corollary 9.4.12 (Concentration of measure on the sphere)

Let A be a measurable subset of the unit sphere in \mathbb{R}^n , equipped with the metric inherited from \mathbb{R}^n . If $A \subseteq S^{n-1}$ has $\text{vol}_{n-1}(A)/\text{vol}_{n-1}(S^{n-1}) \geq 1/2$, then

$$\frac{\text{vol}_{n-1}(A_t)}{\text{vol}_{n-1}(S^{n-1})} \geq 1 - e^{-nt^2/4}.$$

Remark 9.4.13. See §14 in [Barvinok's notes](#) for a proof of the sharper estimate with $e^{-nt^2/4}$ replaced by $\sqrt{\pi/8}e^{-nt^2/2}$, where now we are using the geodesic distance on the sphere.

Corollary 9.4.14 (Concentration of measure on the sphere)

Let S^{n-1} denote the unit sphere in \mathbb{R}^n . If $f: S^{n-1} \rightarrow \mathbb{R}$ is a 1-Lipschitz measurable function, then there is some real m so that, for the uniform measure on the sphere,

$$\mathbb{P}(|f - m| > t) \leq 2e^{-nt^2/4}.$$

Informally: *every Lipschitz function on a high dimensional sphere is almost constant almost everywhere.*

This is a rather counterintuitive high-dimensional phenomenon.

Gauss space

Another related setting is the **Gauss space**, which is \mathbb{R}^n equipped with the probability measure γ_n induced by the Gaussian random vector whose coordinates are n iid standard normals, i.e., the normal random vector in \mathbb{R}^n with covariance matrix I_n . Its probability density function of γ_n at $x \in \mathbb{R}^n$ is $(2\pi)^{-n}e^{-|x|^2/2}$. The metric on \mathbb{R}^n is the usual Euclidean metric.

What would an isoperimetric inequality in Gauss space look like?

Although earlier examples of isoperimetric optimizers were all balls, for the Gauss space, the answer is actually a **half-spaces**, i.e., points on one side of some hyperplane.

The Gaussian isoperimetric inequality, below, was first shown independently by [Borell \(1975\)](#) and [Sudakov and Tsirel'son \(1974\)](#).

Theorem 9.4.15 (Gaussian isoperimetric inequality)

If $A, H \subseteq \mathbb{R}^n$, H a half-space, and $\gamma(A) = \gamma(H)$, then $\gamma(A_t) \geq \gamma(H_t)$ for all $t \geq 0$, where γ is the Gauss measure.

If $H = \{x_1 \leq 0\}$, then $H_t = \{x_1 \leq t\}$, which has Gaussian measure $\geq 1 - e^{-t^2/2}$. Thus:

Corollary 9.4.16 (Concentration of measure for Gaussian vectors)

If $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is 1-Lipschitz, and Z is a vector of i.i.d. standard normals, then $X = f(Z)$ satisfies, for some m ,

$$\mathbb{P}(|X - m| \geq t) \leq 2e^{-t^2/2}.$$

Here is a rather handwavy explanation why the half-space is a reasonable answer.

Consider $\{-1, 1\}^{mn}$, where both m and n are large. Let us group the coordinates of $\{-1, 1\}^{mn}$ into block of length m . The sum of entries in each block (after normalizing by \sqrt{m}) approximates normal random variable by the central limit theorem.

In the Hamming cube, Harper's theorem tells us Hamming balls are isoperimetric optimizers. Since a Hamming ball in $\{-1, 1\}^{mn}$ is given by all points whose sum of coordinates is below a certain threshold, we should look at the analogous subset in the Gauss space, which would then consist of all points whose sum of coordinates is below a certain threshold. The set of all points whose of coordinate sum is below a certain threshold is half-space. Note also that the Gaussian measure is radially symmetric.

The sphere as approximately a sum of independent Gaussians. The Gauss space is a nice space to work with because a standard normal vector simultaneously possesses two useful properties (and it is essentially the only such random vector to have both properties):

- (a) Rotational invariance
- (b) Independence of coordinates

The squared-length of a random Gaussian vector is $Z_1^2 + \dots + Z_n^2$ with iid $Z_1, \dots, Z_n \in N(0, 1)$. It has mean n and a $O(\sqrt{n})$ window of concentration (e.g., by a straightforward adaptation of the Chernoff bound proof). Since $\sqrt{n} + O(\sqrt{n}) = \sqrt{n} + O(1)$, the length of Gaussian vector is concentrated in a $O(1)$ window around \sqrt{n} (the concentration can also be deduced from the above corollary for $f(x) = |x|$). So most of the distribution in the Gauss space lies within a constant distance of a sphere of radius \sqrt{n} . Due to rotational invariance, we see that a Gaussian distribution approximates the uniform distribution on sphere of radius \sqrt{n} in high dimensions. In other words:

$$\text{random Gaussian vector} \approx \sqrt{n} \cdot \text{random unit vector}.$$

Random Gaussian vectors often yield easier calculations due to coordinate independence, and so they often give an accessible way to analyze random unit vectors.

Note that how a *half-space* in the Gauss space intersect the sphere in a *spherical cap*, with both italicized objects being isoperimetric optimizers in their respective spaces.

Sub-Gaussian distributions

We introduce some terminology that captures notions we have seen so far. It will also be convenient for later discussions.

Definition 9.4.17 (Sub-Gaussian distribution)

We say that a random variable X is ***K-subGaussian about its mean*** if

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-t^2/K^2} \quad \text{for all } t \geq 0.$$

Remark 9.4.18. This definition is not standard. Some places say σ^2 -subGaussian for what we mean by σ -subGaussian.

Usually we will not worry about constant factors. Thus, saying that a family of random variables X_n is $O(K_n)$ -subGaussian about its mean is the same as saying that there exist constant $C, c > 0$ such that

$$\mathbb{P}(|X_n - \mathbb{E}X_n| \geq t) \leq Ce^{-ct^2/K_n^2} \quad \text{for all } t \geq 0 \text{ and } n.$$

Also note that, up to changing the constants c, C , the definition does not change if we replace $\mathbb{E}X_n$ by a median of X_n above.

Example 9.4.19. The concentration inequalities so far can be rephrased in terms of subGaussian distributions. Below is summary of results of the form: if X is a random point drawn from the given space, and f is a 1-Lipschitz function, then $f(X)$ is K -subGaussian.

space	distance	-subGaussian	reference
$\{0, 1\}^n$	Hamming	$O(\sqrt{n})$	bounded diff. ineq. (Thm. 9.1.1)
S^{n-1}	Euclidean	$O(1/\sqrt{n})$	Lévy concentration (Cor. 9.4.14)
Gauss space \mathbb{R}^n	Euclidean	$O(1)$	Gaussian isoperimetric ineq. (Cor. 9.4.16)

The following lemma shows that for subGaussian random variables, it does not matter much if we define the tails around its median, mean, or root-mean-square.

Lemma 9.4.20 (Median vs. mean for subGaussian distributions)

There exists a constant $C > 0$ so that the following holds for any real random variable X satisfying, for some constants m and K ,

$$\mathbb{P}(|X - m| \geq t) \leq 2e^{-t^2/K^2} \quad \text{for all } t \geq 0.$$

(a) Every median $\mathbb{M}X$ of X satisfies

$$|\mathbb{M}X - m| \leq CK,$$

(b) The mean of X satisfies

$$|\mathbb{E}X - m| \leq CK,$$

(c) For any $p \geq 1$, writing $\|X\|_p := (\mathbb{E}|X|^p)^{1/p}$ for the L^p norm of X ,

$$|\|X\|_p - m| \leq CK\sqrt{p}.$$

(d) For every constant A there exists a constant $c > 0$ so that if $|m' - m| \leq AK$, then

$$\mathbb{P}(|X - m'| \geq t) \leq 2e^{-ct^2/K^2} \quad \text{for all } t \geq 0.$$

Proof. By considering X/K instead of X , we may assume that $K = 1$ for convenience.

(a) For any $t > \sqrt{2 \log 2}$, we have $\mathbb{P}(|X - m| \geq t) \leq 2e^{-t^2} < 1/2$. So every median of X lies within $\sqrt{2 \log 2}$ of m .

(b) We have

$$\begin{aligned} |\mathbb{E}X - m| &\leq \mathbb{E}|X - m| = \int_0^\infty \mathbb{P}(|X - m| \geq t) dt \\ &\leq \int_0^\infty 2e^{-t^2} dt = \sqrt{\pi}. \end{aligned}$$

(c) Using the triangle inequality on the L^p norm, we have

$$\begin{aligned} |\|X\|_p - m| &\leq \|X - m\|_p = (\mathbb{E}|X - m|^p)^{1/p} = \left(\int_0^\infty \mathbb{P}(|X - m|^p \geq t) dt \right)^{1/p} \\ &\leq \left(\int_0^\infty 2e^{-t^{2/p}} dt \right)^{1/p} = 2^{1/p} \Gamma\left(1 + \frac{p}{2}\right)^{1/p} = O(\sqrt{p}). \end{aligned}$$

(c) We can make c small enough so that $RHS = 2e^{-ct^2} \geq 1$ for $t \leq 2A$. For $t > 2A$,

9 Concentration of measure

we note that

$$\mathbb{P}(|X - m'| \geq t) \leq \mathbb{P}(|X - m| \geq t/2) \leq 2e^{-t^2/4}. \quad \square$$

Remark 9.4.21 (Equivalent characterization of subGaussian distributions). Given a real random variable X , if any of the below is true for some K_i , then the other conditions are true for some $K_j \leq CK_i$ for some absolute constant C .

(a) (Tails) $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/K_1^2}$ for all $t \geq 0$.

(b) (Moments) $\|X\|_{L^p} \leq K_2\sqrt{p}$ for all $p \geq 1$.

(c) (MGF of X^2) $\mathbb{E}e^{X^2/K_3^2} \leq 2$.

We leave the proofs as exercises. Also see §2.5.1 in the textbook *High-Dimensional Probability* by Vershynin (2018), which gives a superb introduction to the subject.

Johnson–Lindenstrauss Lemma

Given a set of N points in high-dimensional Euclidean space, the next result tells us that one can embed them in $O(\varepsilon^{-2} \log N)$ dimensions without sacrificing pairwise distances by more than $1 \pm \varepsilon$ factor. This is known as **dimension reduction**. It is an important tool in many areas, from functional analysis to algorithms.

Theorem 9.4.22 (Johnson and Lindenstrauss 1982)

There exists a constant $C > 0$ so that the following holds. Let $\varepsilon > 0$. Let X be a set of N points in \mathbb{R}^m . Then for any $d > C\varepsilon^{-2} \log N$, there exists $f: X \rightarrow \mathbb{R}^d$ so that

$$(1 - \varepsilon) |x - y| \leq |f(x) - f(y)| \leq (1 + \varepsilon) |x - y| \quad \text{for all } x, y \in X.$$

Remark 9.4.23. Here the requirement $d > C\varepsilon^{-2} \log N$ on the dimension is optimal up to a constant factor (Larsen and Nelson 2017).

We will take f to be $\sqrt{m/d}$ times an orthogonal projection onto a d -dimensional subspace chosen uniformly at random. The theorem then follows from the following lemma together with a union bound.

Lemma 9.4.24 (Random projection)

There exists a constant $C > 0$ so that the following holds. Let $m \geq d$ and let $P: \mathbb{R}^m \rightarrow \mathbb{R}^d$ denote the orthogonal projection onto the subspace spanned by the first d coordinates. Let z be a uniform random point on the unit sphere in \mathbb{R}^m . Let $y = Pz$ and $Y = |y|$. Then, for all $t \geq 0$,

$$\mathbb{P} \left(\left| Y - \sqrt{\frac{d}{m}} \right| \geq t \right) \leq 2e^{-Cmt^2}.$$

To prove the Theorem 9.4.22, for each pair of distinct points $x, x' \in X$, set

$$z = \frac{x - x'}{|x - x'|}, \quad \text{so that } \sqrt{\frac{m}{d}}Y = \frac{|f(x) - f(x')|}{|x - x'|}.$$

Then the length of the projection of z onto a uniform random d -dimensional subspace has the same distribution as Y in the lemma. So setting $t = \varepsilon\sqrt{d/m}$, we find that

$$\mathbb{P} \left(\left| \sqrt{\frac{m}{d}}Y - 1 \right| \geq \varepsilon \right) \leq 2e^{-C\varepsilon d} < 2N^{-cC}.$$

Provided that $C > 1/c$, we can take a union bound over all $\binom{N}{2} < N^2/2$ pairs of points of X to show that with some positive probability, the random f works.

Proof of the lemma. We have $z_1^2 + \cdots + z_n^2 = 1$ and each z_i has the same distribution, so $\mathbb{E}[z_i^2] = 1/m$ for each i . Thus

$$\mathbb{E}[Y^2] = \mathbb{E}[z_1^2 + \cdots + z_d^2] = \frac{d}{m}.$$

Note that P is 1-Lipschitz on the unit sphere. By Lévy's concentration measure theorem on the sphere, letting $\mathbb{M}Y$ denote the median of Y ,

$$\mathbb{P}(|Y - \mathbb{M}Y| \geq t) \leq 2e^{-mt^2/4}.$$

The result then follows by Lemma 9.4.20, using that $\|Y\|_2 = \sqrt{d/m}$. □

Here is a cute application of Johnson–Lindenstrauss (this is related to a homework problem on the Chernoff bound).

Corollary 9.4.25

There is a constant $c > 0$ so that for every positive integer d , there is a set of $e^{c\varepsilon^2 d}$ points in \mathbb{R}^d whose pairwise distances are in $[1 - \varepsilon, 1 + \varepsilon]$.

Proof. Applying Theorem 9.4.22 a regular simplex with unit edge lengths with N vertices in \mathbb{R}^{N-1} to yield N points in \mathbb{R}^d for $d = O(\varepsilon^{-2} \log N)$ and pairwise distances in $[1 - \varepsilon, 1 + \varepsilon]$. \square

9.5 Talagrand's inequality

Talagrand (1995) developed a powerful concentration inequality. It is applicable to many combinatorial optimization problems on independent random inputs. The most general form of Talagrand's inequality can be somewhat difficult to grasp. So we start by discussing a special case with an easier geometric statement. Though, to obtain the full power of Talagrand's inequality with combinatorial consequences, we will need the full statement to be given later.

We omit the proof of Talagrand's inequality (see the Alon–Spencer textbook or [Tao's blog post](#)) and instead focus on explaining the theorem and its applications.

Distance to a subspace

We start with a geometrically motivated question.

Problem 9.5.1

Let V be a *fixed* d -dimensional subspace. Let $x \sim \text{Unif}\{-1, 1\}^n$. How well is $\text{dist}(x, V)$ concentrated?

Let $P = (p_{ij}) \in \mathbb{R}^{n \times n}$ be the matrix giving the orthogonal projection onto V^\perp . We have $\text{tr } P = \dim V^\perp = n - d$. Then

$$\text{dist}(x, V)^2 = |x \cdot Px| = \sum_{i,j} x_i x_j p_{ij}.$$

So

$$\mathbb{E}[\text{dist}(x, V)^2] = \sum_i p_{ii} = \text{tr } P = n - d.$$

How well is $\text{dist}(x, V)$ concentrated around $\sqrt{n - d}$?

Some easier special cases (codimension-1):

- If V is a coordinate subspace, then $\text{dist}(x, V)$ is a constant not depending on x .
- If $V = (1, 1, \dots, 1)^\perp$, then $\text{dist}(x, V) = |x_1 + \dots + x_n|/\sqrt{n}$ which converge to $|Z|$ for $Z \sim N(0, 1)$. In particular, it is $O(1)$ -subGaussian.

- More generally, if for a hyperplane $V = \alpha^\perp$ for some unit vector $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$, one has $\text{dist}(x, V) = |\alpha \cdot x|$. Note that flipping x_i changes $|\alpha \cdot x|$ by at most $2|\alpha_i|$. So the bounded differences inequality Theorem 9.1.3, for every $t \geq 0$,

$$\mathbb{P}(|\text{dist}(x, V) - \mathbb{E} \text{dist}(x, V)| \geq t) \leq 2 \exp \left(\frac{-2t^2}{4(\alpha_1^2 + \dots + \alpha_n^2)} \right) \leq 2e^{-t^2/2}.$$

So again $\text{dist}(x, V)$ is $O(1)$ -subGaussian.

What about higher codimensional subspaces V ? Then

$$\text{dist}(x, V) = \sup_{\substack{\alpha \in V^\perp \\ |\alpha|=1}} |\alpha \cdot x|.$$

It is not clear how to apply the bounded difference inequality to all such α in the above supremum simultaneously.

The bounded difference inequality applied to the function $x \in \{-1, 1\}^n \mapsto \text{dist}(x, V)$, which is 2-Lipschitz (with respect to Hamming distance), gives

$$\mathbb{P}(|\text{dist}(x, V) - \mathbb{E} \text{dist}(x, V)| \geq t) \leq 2e^{-nt^2/2},$$

showing that $\text{dist}(x, V)$ is $O(\sqrt{n})$ -subGaussian—but this is a pretty bad result, as $|\text{dist}(x, V)| \leq \sqrt{n}$ (half the length of the longest diagonal of the cube).

Perhaps the reason why the above bound is so poor is that the bounded difference inequality is measuring distance in $\{-1, 1\}^n$ using the Hamming distance (ℓ_1) whereas we really care about the Euclidean distance (ℓ_2).

If, instead of sampling $x \in \{-1, 1\}^n$, we took x to be a uniformly random point on the radius \sqrt{n} sphere in \mathbb{R}^n (which contains $\{-1, 1\}^n$), then Lévy concentration on the sphere (Corollary 9.4.14) implies that $\text{dist}(x, V)$ is $O(1)$ -subGaussian. Perhaps a similar bound holds when x is chosen from $\{-1, 1\}^n$?

Here is a corollary of Talagrand's inequality, which we will state in its general form later.

Theorem 9.5.2

Let V be a fixed d -dimensional subspace in \mathbb{R}^n . For uniformly random $x \in \{-1, 1\}^n$, one has

$$\mathbb{P}(|\text{dist}(x, V) - \sqrt{n-d}| \geq t) \leq Ce^{-ct^2},$$

where $C, c > 0$ are some constants.

Convex Lipschitz functions of independent random variables

Let us now state Talagrand's inequality, first in a special case for convex functions, and then more generally. Below $\text{dist}(\cdot, \cdot)$ means Euclidean distance.

Theorem 9.5.3 (Talagrand)

Let $A \subseteq \mathbb{R}^n$ be convex. Let $x \sim \text{Unif}\{0, 1\}^n$. Then for any $t \geq 0$,

$$\mathbb{P}(x \in A)\mathbb{P}(\text{dist}(x, A) \geq t) \leq e^{-t^2/4}.$$

Remark 9.5.4. (1) Note that A is a convex body in \mathbb{R}^n and not simply a set of points in A .

(2) The bounded differences inequality gives us an upper bound of the form $e^{-ct^2/n}$, which is much worse than Talagrand's bound.

Example 9.5.5 (Talagrand's inequality fails for nonconvex sets). Let

$$A = \left\{x \in \{0, 1\}^n : \text{wt}(x) \leq \frac{n}{2} - \sqrt{n}\right\}$$

(here A is a discrete set of points and not their convex hull). Then for every $y \in \{0, 1\}^n$ with $\text{wt}(y) \geq n/2$, one has $\text{dist}(y, A) \geq n^{1/4}$ (note that this is Euclidean distance and not Hamming distance). Using the central limit theorem, we have, for some constant $c > 0$ and sufficiently large n , for $x \sim \text{Uniform}(\{-1, 1\}^n)$, $\mathbb{P}(x \in A) \geq c$ and $\mathbb{P}(\text{wt}(x) \geq n/2) \geq 1/2$, so the conclusion of Talagrand's inequality is false for $t = n^{1/4}$, in the case of this nonconvex A .

By an argument similar to our proof of Theorem 9.4.8 (the equivalence of notions of concentration of measure), one can deduce the following consequence.

Corollary 9.5.6 (Talagrand)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex and 1-Lipschitz function (with respect to Euclidean distance on \mathbb{R}^n). Let $x \sim \text{Unif}\{0, 1\}^n$. Then for any $r \in \mathbb{R}$ and $t \geq 0$,

$$\mathbb{P}(f(x) \leq r)\mathbb{P}(f(x) \geq r + t) \leq e^{-t^2/4}.$$

Remark 9.5.7. The proof below shows that the assumption that f is convex can be weakened to f being *quasiconvex*, i.e., $\{f \leq a\}$ is convex for every $a \in \mathbb{R}$.

Proof that Theorem 9.5.3 and Corollary 9.5.6 are equivalent. Theorem 9.5.3 implies Corollary 9.5.6: take $A = \{x : f(x) \leq r\}$. We have $f(x) \leq r + t$ whenever

$\text{dist}(a, A) \leq t$ since f is 1-Lipschitz. So $\mathbb{P}(f(x) \leq r) = \mathbb{P}(x \in A)$ and $\mathbb{P}(f(x) \geq r + t) \leq \mathbb{P}(\text{dist}(x, A) \geq t)$.

Corollary 9.5.6 implies Theorem 9.5.3: $r = 0$ and take $f(x) = \text{dist}(x, A)$, which is a convex function since A is convex. \square

Let us write $\mathbb{M}X$ to be a **median** for the random variable X , i.e., a non-random real so that $\mathbb{P}(X \geq \mathbb{M}X) \geq 1/2$ and $\mathbb{P}(X \leq \mathbb{M}X) \geq 1/2$.

Corollary 9.5.8 (Talagrand)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex and 1-Lipschitz function (with respect to Euclidean distance on \mathbb{R}^n). Let $x \sim \text{Unif}(\{0, 1\}^n)$. Then

$$\mathbb{P}(|f(x) - \mathbb{M}f(x)| \geq t) \leq 4e^{-t^2/4}.$$

Proof. Setting $r = \mathbb{M}f(x)$ in Corollary 9.5.6 yields

$$\mathbb{P}(f(x) \geq \mathbb{M}f(x) + t) \leq 2e^{-t^2/4}.$$

Setting $r = \mathbb{M}f(x) - t$ in Corollary 9.5.6 yields

$$\mathbb{P}(f(x) \leq \mathbb{M}f(x) - t) \leq 2e^{-t^2/4}.$$

\square

Combining the two tail bounds yields the corollary.

Theorem 9.5.2 then follows. Indeed, Corollary 9.5.8 shows that $\text{dist}(x, V)$ (which is a convex 1-Lipschitz function of $x \in \mathbb{R}^n$) is $O(1)$ -subGaussian, which immediately implies Theorem 9.5.2.

Example 9.5.9 (Operator norm of a random matrix). Let A be a random matrix whose entries are uniform iid from $\{-1, 1\}$. Viewing $A \mapsto \|A\|_{\text{op}}$ as a function $\mathbb{R}^{n^2} \rightarrow \mathbb{R}$, we see that it is convex (since the operator norm is a norm) and 1-Lipschitz (using that $\|\cdot\|_{\text{op}} \leq \|\cdot\|_{\text{HS}}$, where the latter is the Hilbert–Schmidt norm, also known as the Frobenius norm, i.e., the ℓ_2 -norm of the matrix entries). It follows by Talagrand's inequality (Corollary 9.5.8) that $\|A\|_{\text{op}}$ is $O(1)$ -subGaussian about its mean.

Convex distance

Talagrand's inequality has a much more general form, which has far-reaching combinatorial applications. We need to define a more subtle notion of distance.

We consider $\Omega = \Omega_1 \times \cdots \times \Omega_n$ with product probability measure (i.e., independent random variables).

9 Concentration of measure

Weighted hamming distance: given $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}_{\geq 0}^n$, $x, y \in \Omega$, we set

$$d_\alpha(x, y) := \sum_{i: x_i \neq y_i} \alpha_i$$

For $A \subseteq \Omega$,

$$d_\alpha(x, A) := \inf_{y \in A} d_\alpha(x, y).$$

Talagrand's **convex distance** between $x \in \Omega$ and $A \subseteq \Omega$ is defined by

$$d_T(x, A) := \sup_{\substack{\alpha \in \mathbb{R}_{\geq 0}^n \\ |\alpha|=1}} d_\alpha(x, A).$$

Here $|\alpha|$ denotes Euclidean length:

$$|\alpha|^2 := \alpha_1^2 + \dots + \alpha_n^2.$$

Example 9.5.10 (Euclidean distance to convex hull). If $A \subseteq \{0, 1\}^n$ and $x \in \{0, 1\}^n$, then $d_T(x, A)$ is the Euclidean distance from x to the convex hull of A .

Let us give another interpretation of convex distance. For $x, y \in \Omega$, let

$$\phi_x(y) = (1_{x_1 \neq y_1}, 1_{x_2 \neq y_2}, \dots, 1_{x_n \neq y_n}) \in \{0, 1\}^n$$

be the vector of coordinatewise disagreements between x and y . Write

$$\phi_x(A) = \{\phi_x(y) : y \in A\} \subseteq \{0, 1\}^n.$$

Then for any $\alpha \in \mathbb{R}_{\geq 0}^n$,

$$d_\alpha(x, A) = d_\alpha(\vec{0}, \phi_x(A)),$$

where the LHS is the weighted Hamming distance in Ω whereas the RHS takes place in $\{0, 1\}^n$. Taking the supremum over $\alpha \in \mathbb{R}_{\geq 0}^n$ with $|\alpha| = 1$, and using the Example 9.5.10, we deduce

$$d_T(x, A) = \text{dist}(\vec{0}, \text{ConvexHull } \phi_x(A)).$$

The general form of Talagrand's inequality says the following. Note that it reduces to the earlier special case Theorem 9.5.3 if $\Omega = \{0, 1\}^n$.

Theorem 9.5.11 (Talagrand's inequality: general form)

Let $A \subseteq \Omega = \Omega_1 \times \cdots \times \Omega_n$, with Ω equipped with a product probability measure. Let $x \in \Omega$ be chosen randomly with independent coordinates. Let $t \geq 0$. Then

$$\mathbb{P}(x \in A) \mathbb{P}(d_T(x, A) \geq t) \leq e^{-t^2/4}.$$

Let us see how Talagrand's inequality recovers a more general form of our geometric inequalities from earlier, extending from independent boolean random variables to independent bounded random variables.

Lemma 9.5.12 (Convex distance upper bounds Euclidean distance)

Let $A \subseteq [0, 1]^n$ and $x \in [0, 1]^n$. Then $\text{dist}(x, \text{ConvexHull } A) \leq d_T(x, A)$.

Proof. For any $\alpha \in \mathbb{R}^n$, and any $y \in [0, 1]^n$, we have

$$|(x - y) \cdot \alpha| \leq \sum_{i=1}^n |\alpha_i| |x_i - y_i| \leq \sum_{i: x_i \neq y_i}^n |\alpha_i|.$$

First taking the infimum over all $y \in A$, and then taking the supremum over unit vectors α , the LHS becomes $\text{dist}(x, \text{ConvexHull } A)$ and the RHS becomes $d_T(x, A)$. \square

Corollary 9.5.13 (Talagrand's inequality: convex sets and convex Lipschitz functions)

Let $x = (x_1, \dots, x_n) \in [0, 1]^n$ be independent random variables (not necessarily identical). Let $t \geq 0$. Let $A \subseteq [0, 1]^n$ be a convex set. Then

$$\mathbb{P}(x \in A) \mathbb{P}(\text{dist}(x, A) \geq t) \leq e^{-t^2/4}$$

where dist is Euclidean distance. Also, if $f: [0, 1]^n \rightarrow \mathbb{R}$ is a convex 1-Lipschitz function, then

$$\mathbb{P}(|f - \mathbb{M}f| \geq t) \leq 4e^{-t^2/4}.$$

Here is a form of Talagrand's inequality that is useful for combinatorial applications. Below, one should think of $f(x)$ as the value of some optimization problem on some random input x . There is a hypothesis on how much $f(x)$ can change if we alter x . An example that we will examine in the next section is the length of the shortest tour through n random points in the unit square (the Euclidean traveling salesman problem).

Theorem 9.5.14 (Talagrand's inequality — functions with weighted certificates)

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ equipped with the product measure. Let $f: \Omega \rightarrow \mathbb{R}$ be a function. Suppose for every $x \in \Omega$, there is some $\alpha(x) = (\alpha_1(x), \dots, \alpha_n(x)) \in \mathbb{R}_{\geq 0}^n$ such that

$$f(y) \geq f(x) - \sum_{i: x_i \neq y_i} \alpha_i(x) \quad \text{for all } y \in \Omega.$$

Then, for every $t \geq 0$, (recall $|\alpha|^2 = \sum_{i=1}^n \alpha_i(x)^2$)

$$\mathbb{P}(|f - \mathbb{M}f| \geq t) \leq 4e^{-t^2/K^2} \quad \text{where } K = 2 \sup_{x \in \Omega} |\alpha(x)|.$$

Remark 9.5.15. By considering $-f$ instead of f , we can change the hypothesis on f to

$$f(y) \leq f(x) + \sum_{i: x_i \neq y_i} \alpha_i(x) \quad \text{for all } y \in \Omega.$$

Note that x and y play asymmetric roles.

Remark 9.5.16. The vector $\alpha(x)$ measures the resilience of $f(x)$ under changing some coordinates of x . It is important that we can choose a different weight $\alpha(x)$ for each x . In fact, if we do not let $\alpha(x)$ change with x , then Theorem 9.5.14 recovers the bounded differences inequality Theorem 9.1.3 up to an unimportant constant factor in the exponent of the bound.

Proof. Let $r \in \mathbb{R}$. Let $A = \{y \in \Omega : f(y) \leq r - t\}$. Consider an $x \in \Omega$ with $f(x) \geq r$. By hypothesis, there is some $\alpha(x) \in \mathbb{R}_{\geq 0}^n$ such that

$$d_{\alpha(x)}(x, y) \geq f(x) - f(y) \geq t \quad \text{for all } y \in A.$$

Taking infimum over $y \in A$, we find

$$|\alpha(x)| d_T(x, A) \geq t.$$

So

$$d_T(x, A) \geq \frac{t}{|\alpha(x)|} \geq \frac{2t}{K}.$$

And hence by Talagrand's inequality Theorem 9.5.11,

$$\mathbb{P}(f \leq r - t) \mathbb{P}(f \geq r) \leq \mathbb{P}(x \in A) \mathbb{P}\left(d_T(x, A) \geq \frac{2t}{K}\right) \leq e^{-t^2/K^2}.$$

Taking $r = \mathbb{M}f + t$ yields

$$\mathbb{P}(f \geq \mathbb{M}f + t) \leq 2e^{-t^2/K^2},$$

and taking $r = \mathbb{M}f$ yields

$$\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2e^{-t^2/K^2}.$$

Putting them together yields the final result. \square

Largest eigenvalue of a random matrix

Theorem 9.5.17

Let $A = (a_{ij})$ be an $n \times n$ symmetric random matrix with independent entries in $[-1, 1]$. Let $\lambda_1(A)$ denote the largest eigenvalue of A . Then

$$\mathbb{P}(|\lambda_1(A) - \mathbb{M}\lambda_1(A)| \geq t) \leq 4e^{-t^2/32}.$$

Proof. We shall verify the hypotheses of Theorem 9.5.14. We would like to come up with a good choice of a weight vector $\alpha(A)$ for each matrix A so that for any other symmetric matrix B with $[-1, 1]$ entries,

$$\lambda_1(B) \geq \lambda_1(A) - \sum_{i \leq j: a_{ij} \neq b_{ij}} \alpha_{i,j}. \quad (9.1)$$

Note that in a random symmetric matrix we only have $n(n+1)/2$ independent random entries: the entries below the diagonal are obtained by reflecting the upper diagonal entries.

Let $v = v(A)$ be the unit eigenvector of A corresponding to the eigenvalue $\lambda_1(A)$. Then, by the Courant–Fischer characterization of eigenvalues,

$$v^\top A v = \lambda_1(A) \quad \text{and} \quad v^\top B v \leq \lambda_1(B).$$

Thus

$$\lambda_1(A) - \lambda_1(B) \leq v^\top (A - B) v \leq \sum_{i,j: a_{ij} \neq b_{ij}} |v_i| |v_j| |a_{ij} - b_{ij}| \leq \sum_{i,j: a_{ij} \neq b_{ij}} 2|v_i| |v_j|.$$

Thus (9.1) holds for the vector $\alpha(A) = (\alpha_{ij})_{i \leq j}$ defined by

$$\alpha_{ij} = \begin{cases} 4|v_i| |v_j| & \text{if } i < j \\ 2|v_i|^2 & \text{if } i = j. \end{cases}$$

9 Concentration of measure

We have

$$\sum_{i \leq j} \alpha_{ij}^2 \leq 8 \sum_{i,j} |v_i|^2 |v_j|^2 = 8 \left(\sum_i |v_i|^2 \right)^2 = 8.$$

So Theorem 9.5.14 yields the result. \square

Remark 9.5.18. If A has mean zero entries, then a moments computation shows that $\mathbb{E}\lambda_1(A) = O(\sqrt{n})$ (the constant can be computed as well). A much more advanced fact is that, say for uniform $\{-1, 1\}$ entries, the true scale of fluctuation is $n^{-1/6}$, and when normalized, the distribution converges to something known as the [Tracy–Widom](#) distribution. This limiting distribution is “universal” in the sense that it occurs in many naturally occurring problems, including the next example.

Certifiable functions and longest increasing subsequence

An **increasing subsequence** of a permutation $\sigma = (\sigma_1, \dots, \sigma_n)$ is defined to be some $(\sigma_{i_1}, \dots, \sigma_{i_\ell})$ for some $i_1 < \dots < i_\ell$.

Question 9.5.19

How well is the length X of the longest increasing subsequence of uniform random permutation concentrated?

While the entries of σ are not independent, we can generate a uniform random permutation by taking iid uniform $x_1, \dots, x_n \sim \text{Unif}[0, 1]$ and let σ record the ordering of the x_i 's. This trick converts the problem into one about independent random variables.

We leave it as an exercise to deduce that X is $\Theta(\sqrt{n})$ whp.

Changing one of the x_i 's changes LIS by at most 1, so the bounded differences inequality tells us that X is $O(\sqrt{n})$ -subGaussian. Can we do better?

The assertion that a permutation has an increasing permutation of length s can be checked by verifying s coordinates of the permutation. Talagrand's inequality tells us that in such situations the typical fluctuation should be on the order $O(\sqrt{MX})$, or $O(n^{1/4})$ in this case.

Definition 9.5.20

Let $\Omega = \Omega_1 \times \dots \times \Omega_n$. Let $A \subseteq \Omega$. We say that A is **s -certifiable** if for every $x \in A$, there exists a set $I(x) \subseteq [n]$ with $|I| \leq s$ such that for every $y \in \Omega$ with $x_i = y_i$ for all $i \in I(x)$, one has $y \in A$.

For example, for a random permutation as earlier, having an increasing subsequence of length $\geq s$ is s -certifiable (namely by the indices of the length s increasing subsequence).

Theorem 9.5.21 (Talagrand's inequality for certifiable functions)

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be equipped with a product measure. Let $f: \Omega \rightarrow \mathbb{R}$ be 1-Lipschitz with respect to Hamming distance on Ω . Suppose that $\{f \geq r\}$ is s -certifiable. Then, for every $t \geq 0$,

$$\mathbb{P}(f \leq r - t)\mathbb{P}(f \geq r) \leq e^{-t^2/(4s)}.$$

Proof. Let $A, B \subseteq \Omega$ be given by $A = \{x : f(x) \leq r - t\}$ and $B = \{y : f(y) \geq r\}$. For every $y \in B$, let $I(y) \subseteq [n]$ denote a set of $\leq s$ coordinates that certify $f \geq r$. Due to f being 1-Lipschitz, we see that every $x \in A$ disagrees with y on $\geq t$ coordinates of $I(y)$.

For every $y \in B$, let $\alpha(y)$ be the indicator vector for $I(y)$ normalized in length to a unit vector. Then for any $x \in A$,

$$d_\alpha(x, y) = \frac{|\{i \in I(y) : x_i \neq y_i\}|}{\sqrt{|I|}} \geq \frac{t}{\sqrt{s}}.$$

Thus $d_T(y, A) \geq t/\sqrt{s}$. Thus

$$\mathbb{P}(f \leq r - t)\mathbb{P}(f \geq r) \leq \mathbb{P}(A)\mathbb{P}(B) \leq \mathbb{P}(x \in A)\mathbb{P}(d_T(x, A) \geq t/\sqrt{s}) \leq e^{-t^2/(4s)}$$

by Talagrand's inequality (Theorem 9.5.11). □

Corollary 9.5.22 (Talagrand's inequality for certifiable functions)

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be equipped with a product measure. Let $f: \Omega \rightarrow \mathbb{R}$ be 1-Lipschitz with respect to Hamming distance on Ω . Suppose $\{f \geq r\}$ is r -certifiable for every r . Then for every $t \geq 0$,

$$\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2 \exp\left(\frac{-t^2}{4\mathbb{M}f}\right)$$

and

$$\mathbb{P}(f \geq \mathbb{M}f + t) \leq 2 \exp\left(\frac{-t^2}{4(\mathbb{M}f + t)}\right).$$

Proof. Applying the previous theorem, we have, for every $r \in \mathbb{R}$ and every $t \geq 0$,

$$\mathbb{P}(f \leq r - t)\mathbb{P}(f \geq r) \leq \exp\left(\frac{-t^2}{4r}\right).$$

9 Concentration of measure

Setting $r = \mathbb{M}f$, we obtain the lower tail.

$$\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2 \exp\left(\frac{-t^2}{4\mathbb{M}f}\right).$$

Setting $r = \mathbb{M}f + t$, we obtain the upper tail

$$\mathbb{P}(X \geq \mathbb{M}f + t) \leq 2 \exp\left(\frac{-t^2}{4(\mathbb{M}f + t)}\right). \quad \square$$

We can apply the above corollary to $[0, 1]^n$ with f being the length of the longest subsequence. Then $f \geq r$ is r -certifiable. It is also easy to deduce that $\mathbb{M}f = O(\sqrt{n})$. The above tail bounds give us a concentration window of width $O(n^{1/4})$.

Corollary 9.5.23 (Longest increasing subsequence)

Let X be the length of the longest increasing subsequence of a random permutation of $[n]$. Then for every $\varepsilon > 0$ there exists $C > 0$ so that

$$\mathbb{P}(|X - \mathbb{M}X| \leq Cn^{1/4}) \geq 1 - \varepsilon.$$

Remark 9.5.24. The distribution of the length X of longest increasing subsequence of a uniform random permutation is now well understood through some deep results.

Vershik and Kerov (1977) showed that $\mathbb{E}X \sim 2\sqrt{n}$.

Baik, Deift, and Johansson (1999) showed that the correct scaling factor is $n^{1/6}$, and furthermore, $n^{-1/6}(X - 2\sqrt{n})$ converges to the Tracy–Widom distribution, the same distribution for the top eigenvalue of a random matrix.

9.6 Euclidean traveling salesman problem

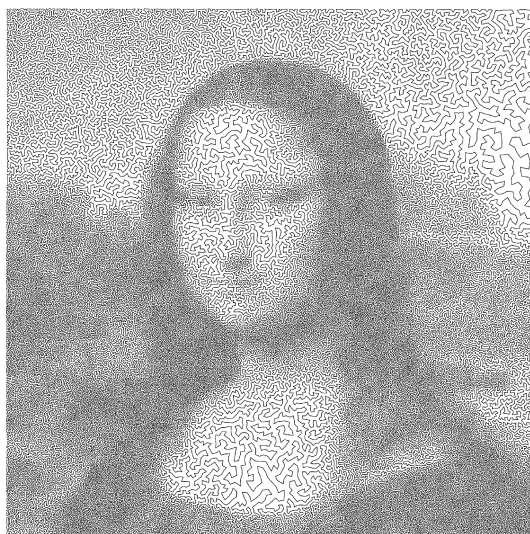
Given points $x_1, \dots, x_n \in [0, 1]^2$, let $L(x_1, \dots, x_n) = L(\{x_1, \dots, x_n\})$ denote the length of the shortest tour through all given points and returns to its starting point. Equivalently, $L(x_1, \dots, x_n)$ is the minimum of

$$|x_{\sigma(1)} - x_{\sigma(2)}| + |x_{\sigma(2)} - x_{\sigma(3)}| + \dots + |x_{\sigma(n)} - x_{\sigma(1)}|$$

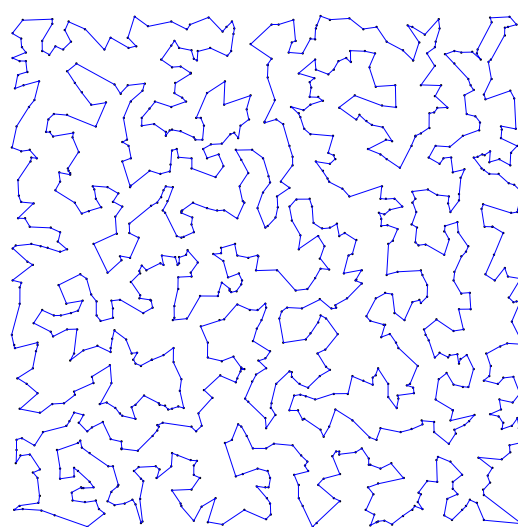
as σ ranges over all permutations of $[n]$. This Euclidean traveling salesman problem is NP-hard to solve exactly, although there is a $(1 + \varepsilon)$ -factor approximation algorithm with running polynomial time for any constant $\varepsilon > 0$ (Arora 1998).

Let

$$L_n = L(x_1, \dots, x_n) \quad \text{with i.i.d. } x_1, \dots, x_n \sim \text{Unif}([0, 1]^2)$$



The Mona Lisa TSP challenge.



A tour of 1000 random points.

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

Beardwood, Halton, and Hammersley (1959) showed that whp L_n/\sqrt{n} converges to some constant as $n \rightarrow \infty$ (the exact value of the constant is unknown).

We shall focus on the concentration of L_n .

We will present two methods that illustrate different techniques from this chapter.

Martingale methods

The following simple monotonicity property will be important for us: for any S and $x \in [0, 1]^2$,

$$L(S) \leq L(S \cup \{x\}) \leq L(S) + 2 \operatorname{dist}(x, S).$$

Here is the justification for the second inequality. Let y be the closest point in S to x . Consider a shortest tour through S of length $L(S)$. Let us modify this tour by first traversing through it, and when we hit y , we take a detour excursion from y to x and then back to y . The length of this tour, which contains $S \cup \{x\}$, is $L(S) + 2 \operatorname{dist}(x, S)$, and thus the shortest tour through $S \cup \{x\}$ has length at most $L(S) + 2 \operatorname{dist}(x, S)$.

If we simply apply the bounded difference inequality, we find that changing a single x_i might change $L(x_1, \dots, x_n)$ by $O(1)$ in the worst case, and so L_n is $O(\sqrt{n})$ -subGaussian about its mean. This is a trivial result since L_n is typically $\Theta(\sqrt{n})$.

To do better, we apply Azuma's inequality to the Doob martingale. The key observation is that the initially revealed points do not affect the conditional expectations by much even in the worst case.

Theorem 9.6.1 (Rhee and Talagrand 1987)

L_n is $O(\sqrt{\log n})$ -subGaussian about its mean. That is,

$$\mathbb{P}(|L_n - \mathbb{E}L_n| \geq t) \leq \exp\left(\frac{-ct^2}{\log n}\right) \quad \text{for all } t > 0,$$

where $c > 0$ is some constant.

We need the following estimate.

Lemma 9.6.2

Let S be a set of k random points chosen independently and uniformly in $[0, 1]^2$. For any (non-random) point $y \in [0, 1]^2$, one has

$$\mathbb{E} \text{dist}(y, S) \lesssim \frac{1}{\sqrt{k}}.$$

Proof. We have

$$\begin{aligned} \mathbb{E} \text{dist}(y, S) &= \int_0^{\sqrt{2}} \mathbb{P}(\text{dist}(y, S) \geq t) dt \\ &= \int_0^{\sqrt{2}} \left(1 - \text{area}\left(B(y, t) \cap [0, 1]^2\right)\right)^k dt \\ &\leq \int_0^{\sqrt{2}} \exp\left(-k \text{area}\left(B(y, t) \cap [0, 1]^2\right)\right) dt \\ &\leq \int_0^{\infty} \exp\left(-\Omega(kt^2)\right) dt \lesssim \frac{1}{\sqrt{k}}. \end{aligned} \quad \square$$

Proof of Theorem 9.6.1. Let

$$L_{n,i}(x_1, \dots, x_i) = \mathbb{E}[L_n(x_1, \dots, x_n) \mid x_1, \dots, x_i]$$

be the expectation of L_n conditional on the first i points (and averaging over the remaining $n - i$ points).

Claim. $L_{n,i}$ is $O\left(\frac{1}{\sqrt{n-i+1}}\right)$ -Lipschitz with respect to Hamming distance.

We have

$$\begin{aligned} L(x_1, \dots, x_i, \dots, x_n) &\leq L(x_1, \dots, x'_i, \dots, x_n) + 2 \text{dist}(x_i, \{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n\}) \\ &\leq L(x_1, \dots, x_i, \dots, x_n) + \begin{cases} 2 \text{dist}(x_i, \{x_{i+1}, \dots, x_n\}) & \text{if } i < n \\ O(1) & \text{if } i = n. \end{cases} \end{aligned}$$

Taking expectation over x_{i+1}, \dots, x_n , and applying the previous lemma, we find that

$$L_{n,i}(x_1, \dots, x_i) \leq L_{n,i}(x_1, \dots, x_{i-1}, x'_i) + O\left(\frac{1}{\sqrt{n-i+1}}\right).$$

This proves the claim. Thus the Doob martingale

$$Z_i = \mathbb{E}[L_n(x_1, \dots, x_n) \mid x_1, \dots, x_i] = L_{n,i}(x_1, \dots, x_i)$$

satisfies

$$|Z_i - Z_{i-1}| \lesssim \frac{1}{\sqrt{n-i+1}} \quad \text{for each } 1 \leq i \leq n.$$

Now we apply Azuma's inequality (Theorem 9.2.8). Since

$$\sum_{i=1}^n \left(\frac{1}{\sqrt{n-i+1}}\right)^2 = O(\log n),$$

we deduce that $Z_N = L_n$ is $O(\sqrt{\log n})$ -subGaussian about its mean. \square

Talagrand's inequality

Using Talagrand's inequality, we will prove the following stronger estimate.

Theorem 9.6.3 (Rhee and Talagrand 1989)

L_n is $O(1)$ -subGaussian about its mean. That is,

$$\mathbb{P}(|L_n - \mathbb{E}L_n| \geq t) \leq e^{-ct^2} \quad \text{for all } t > 0,$$

where $c > 0$ is some constant.

Remark 9.6.4. Rhee (1991) showed that this tail bound is sharp.

The proof below, following Steele (1997), applies the “space-filling curve heuristic.”

A **space-filling curve** is a continuous surjection from $[0, 1]$ to $[0, 1]^2$. Peano (1890) constructed the first space-filling curve. Hilbert (1891) constructed another space-filling curve known as the **Hilbert curve**. We will not give a precise description of the Hilbert curve here. Intuitively, the Hilbert curve is the pointwise limit of a sequence of piecewise linear curves illustrated in Figure 9.2. I recommend this [3Blue1Brown video](#) on YouTube for a beautiful animation of the Hilbert curve along with applications.

We will only need the following property of the Hilbert space filling curve.

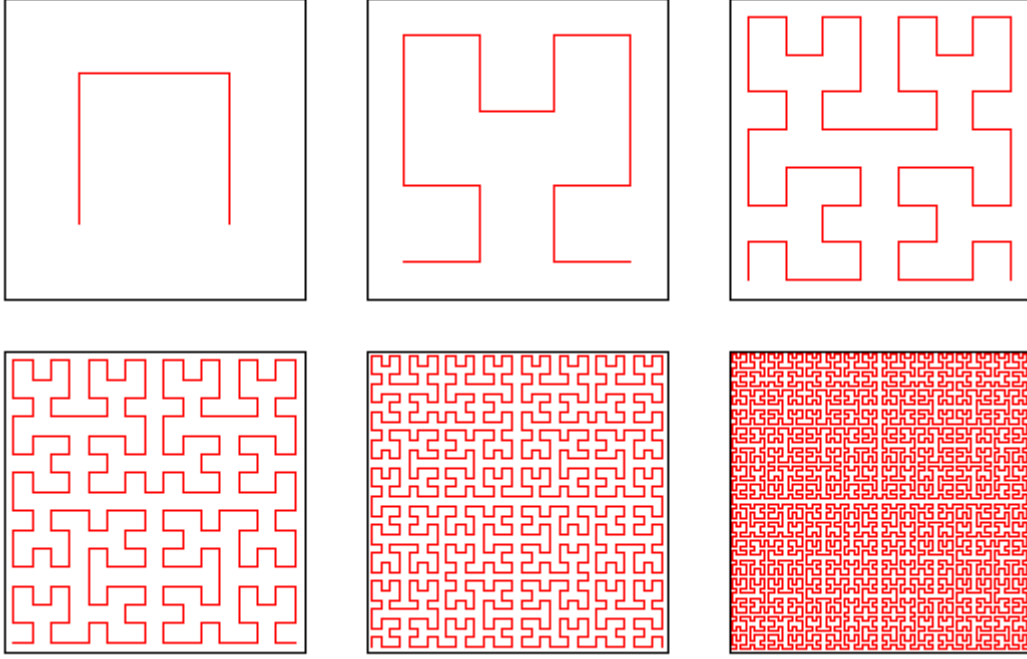


Figure 9.2: The Hilbert space-filling curve is the limit of discrete curves illustrated.

Definition 9.6.5 (Hölder continuity)

Given two metric spaces (X, d_X) and (Y, d_Y) , we say that a map $f: X \rightarrow Y$ is **Hölder continuous with exponent α** if there is some constant C (depending on f) so that

$$d_Y(f(x), f(x')) \leq C d_X(x, x')^\alpha \quad \text{for all } x, x' \in X.$$

Remark 9.6.6. Hölder continuity with exponent 1 is the same as Lipschitz continuity. Often X has bounded diameter, in which case if f is Hölder continuous with exponent α , then it is so with any exponent $\alpha' < \alpha$.

Theorem 9.6.7

The Hilbert curve $H: [0, 1] \rightarrow [0, 1]^2$ is Hölder continuous with exponent $1/2$.

Proof sketch. The Hilbert space-filling curve H sends every interval of the form $[(i-1)/4^n, i/4^n]$ to a square of the form $[(j-1)/2^n, j/2^n] \times [(k-1)/2^n, k/2^n]$. Indeed, for each fixed n , the discrete curves eventually all have this property.

Let $x, y \in [0, 1]$, and let n be the largest integer so that $x, y \in [(i-1)/4^n, (i+1)/4^n]$ for some integer i . Then $|x - y| = \Theta(4^{-n})$, and $|H(x) - H(y)| \lesssim 2^{-n}$. Thus $|H(x) - H(y)| \lesssim |x - y|^{1/2}$. \square

Remark 9.6.8. If a space filling space is Hölder continuous with exponent α , then $\alpha \leq 1/2$. Indeed, the images of the intervals $[(i-1)/k, i/k]$, $i = 1, \dots, k$, cover the unit square, and thus one intervals must have image diameter $\gtrsim 1/\sqrt{k}$.

Lemma 9.6.9 (Space-filling curve heuristic)

Let $x_1, \dots, x_n \in [0, 1]^2$. There is a permutation of σ of $[n]$ with (indices taken mod n)

$$\sum_{i=1}^n |x_{\sigma(i)} - x_{\sigma(i+1)}|^2 = O(1).$$

Proof. Order the points as they appear on the Hilbert space filling curve $H: [0, 1] \rightarrow [0, 1]^2$ (since H is not injective, there is more than one possible order). Then, there exist $0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq 1$ so that $H(t_i) = x_{\sigma(i)}$ for each i . Using that H is Hölder continuous with exponent $1/2$, we have

$$\sum_{i=1}^n |x_{\sigma(i)} - x_{\sigma(i+1)}|^2 = \sum_{i=1}^n |H(t_i) - H(t_{i+1})|^2 \lesssim \sum_{i=1}^n |t_i - t_{i+1}| \leq 2. \quad \square$$

Remark 9.6.10. We leave it as an exercise to find an elementary proof of the lemma without invoking the existence of a space-filling curve. Hint: consider a finite approximation of the Hilbert curve.

Using Talagrand's inequality in the form of Theorem 9.5.14, to prove Theorem 9.6.3 that L_n is $O(1)$ -subGaussian, it suffices to prove the following lemma.

Lemma 9.6.11

Let $\Omega = ([0, 1]^2)^n$ be the space of n -tuples of points in $[0, 1]^2$. There exists a map $\alpha: \Omega \rightarrow \mathbb{R}_{\geq 0}^n$ so that for all $x \in \Omega$, $\alpha(x) = (\alpha_1(x), \dots, \alpha_n(x)) \in \mathbb{R}_{\geq 0}^n$ satisfies

$$L(x) \leq L(y) + \sum_{i: x_i \neq y_i} \alpha_i(x) \quad \text{for all } x, y \in \Omega \quad (9.1)$$

and

$$\sup_{x \in \Omega} \sum_{i=1}^n \alpha_i(x)^2 = O(1). \quad (9.2)$$

Proof. Let $x = (x_1, \dots, x_n) \in \Omega$, and let σ be the permutation of $[n]$ given by Lemma 9.6.9, the space-filling curve heuristic. Then σ induces a tour of x_1, \dots, x_n . Let $\alpha_i(x)$ equal twice the sum of the lengths of the two edges incident to x_i in this tour

9 Concentration of measure

(indices taken mod n):

$$\alpha_i(x) = 2 \left(|x_i - x_{\sigma(\sigma^{-1}(i)+1)}| + |x_i - x_{\sigma(\sigma^{-1}(i)-1)}| \right).$$

Intuitively, this quantity captures “difficulty to serve” x_i .

Now we prove (9.1). First we take care of the first case when $x_i \neq y_i$ for all i : (9.1) follows from

$$L(x) \leq \sum_{i=1}^n |x_{\sigma(i)} - x_{\sigma(i+1)}| = \frac{1}{2} \sum_{i=1}^n \alpha_i(x).$$

Now suppose that $x_i = y_i$ for at least one i . Suppose we have a tour through y of length $L(y)$. Consider, for each i with $x_i \neq y_i$, the point x_i along with the two segments incident to x_i in the σ -induced tour through x (these are the “new edges”). Starting with an optimal tour through y , and by making various detours/excursions on the new edges, we can reach all the points of x , traversing each new edge at most twice. The length of the new tour is at most $L(y) + \sum_{i: x_i \neq y_i} \alpha_i(x)$. This proves (9.1).

Finally, it remains to prove (9.2). By Lemma 9.6.9,

$$\begin{aligned} \sum_{i=1}^n \alpha_i(x)^2 &\leq 4 \sum_{j=1}^n \left(|x_{\sigma(j)} - x_{\sigma(j+1)}| + |x_{\sigma(j)} - x_{\sigma(j-1)}| \right)^2 \\ &\lesssim \sum_{j=1}^n |x_{\sigma(j)} - x_{\sigma(j+1)}|^2 = O(1). \end{aligned} \quad \square$$

10 Entropy method

My greatest concern was what to call it. I thought of calling it “information,” but the word was overly used, so I decided to call it “uncertainty.” When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, “You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.”

Claude Shannon, 1971

For more information theory, see the textbook by [Cover and Thomas](#).

10.1 Basic properties

We define the (binary) entropy of a discrete random variable as follows.

Definition 10.1.1

Given a discrete random variable X taking values in S , with $p_s := \mathbb{P}(X = s)$, its *entropy* (or *binary entropy* to emphasize the base-2 logarithm) is defined to be

$$H(X) := \sum_{s \in S} -p_s \log_2 p_s$$

(by convention if $p_s = 0$ then the corresponding summand is set to zero).

Remark 10.1.2 (Base of the logarithm). It is also fine to use another base for the logarithm, e.g., the natural log, as long as we are consistent throughout. There is some combinatorial preference for base-2 due to its interpretation as counts bits. For certain results, such as Pinsker’s inequality (which we will unfortunately not cover here), the choice of the base does matter.

Remark 10.1.3 (Information theoretic interpretation). Intuitively, $H(X)$ measures the amount of “surprise” in the randomness of X . It can also be interpreted as the

10 Entropy method

amount of information learned by seeing the random variable X . A more rigorous interpretation of this intuition is given by the **Shannon source coding theorem**, which, informally, says that the minimum number of bits needed to encode n iid copies of X is $nH(X) + o(n)$.

Here are some basic properties. Throughout we only consider discrete random variables.

The proofs are all routine calculations. It will be useful to understand the information theoretic interpretations of these properties.

Lemma 10.1.4 (Uniform bound)

$$H(X) \leq \log_2 |\text{support}(X)|,$$

with equality if and only if X is uniformly distributed.

Proof. Let function $f(x) = -x \log_2 x$ is concave for $x \in [0, 1]$. Let $S = \text{support}(X)$. Then

$$H(X) = \sum_{s \in S} f(p_s) \leq |S| f\left(\frac{1}{|S|} \sum_{s \in S} p_s\right) = |S| f\left(\frac{1}{|S|}\right) = \log_2 |S|. \quad \square$$

We write $H(X, Y)$ for the entropy of the joint random variables (X, Y) . In other words, letting $Z = (X, Y)$,

$$H(X, Y) := H(Z) = \sum_{(x,y)} -\mathbb{P}(X = x, Y = y) \log_2 \mathbb{P}(X = x, Y = y).$$

We can similarly write $H(X_1, \dots, X_n)$ for joint entropy.

Lemma 10.1.5 (Independence)

If X and Y are independent random variables, then

$$H(X, Y) = H(X) + H(Y).$$

Proof.

$$\begin{aligned}
 H(X, Y) &= \sum_{(x,y)} -\mathbb{P}(X = x, Y = y) \log_2 \mathbb{P}(X = x, Y = y) \\
 &= \sum_{(x,y)} -p_x p_y \log_2 (p_x p_y) \\
 &= \sum_{(x,y)} -p_x p_y (\log_2 p_x + \log_2 p_y) \\
 &= \sum_x -p_x \log_2 p_x + \sum_y -p_y \log_2 p_y = H(X) + H(Y). \quad \square
 \end{aligned}$$

Definition 10.1.6 (Conditional entropy)

Given jointly distributed random variables X and Y , define

$$\begin{aligned}
 H(X|Y) &:= \mathbb{E}_y[H(X|Y = y)] \\
 &= \sum_y \mathbb{P}(Y = y) H(X|Y = y) \\
 &= \sum_y \mathbb{P}(Y = y) \sum_x -\mathbb{P}(X = x|Y = y) \log_2 \mathbb{P}(X = x|Y = y)
 \end{aligned}$$

(each line unpacks the previous line. In the summations, x and y range over the supports of X and Y respectively).

Intuitively, the conditional entropy $H(X|Y)$ measures the amount of additional information in X not contained in Y . This intuition is also captured by the next lemma.

Some important special cases:

- If $X = Y$, or $X = f(Y)$, then $H(X|Y) = 0$.
- If X and Y are independent, then $H(X|Y) = H(X)$
- If X and Y are conditionally independent on Z , then $H(X, Y|Z) = H(X|Z) + H(Y|Z)$ and $H(X|Y, Z) = H(X|Z)$.

Lemma 10.1.7 (Chain rule)

$$H(X, Y) = H(X) + H(Y|X)$$

Proof. Writing $p(x, y) = \mathbb{P}(X = x, Y = y)$, etc., we have by Bayes's rule

$$p(x|y)p(y) = p(x, y),$$

and so

$$\begin{aligned}
 H(X|Y) &:= \mathbb{E}_y[H(X|Y = y)] = \sum_y -p(y) \sum_x p(x|y) \log_2 p(x|y) \\
 &= \sum_{x,y} -p(x,y) \log_2 \frac{p(x,y)}{p(y)} \\
 &= \sum_{x,y} -p(x,y) \log_2 p(x,y) + \sum_y p(y) \log_2 p(y) \\
 &= H(X,Y) - H(Y).
 \end{aligned}$$

□

Lemma 10.1.8 (Subadditivity)

$H(X, Y) \leq H(X) + H(Y)$, and more generally,

$$H(X_1, \dots, X_n) \leq H(X_1) + \dots + H(X_n).$$

Proof. Let $f(t) = \log_2(1/t)$, which is convex. Then

$$\begin{aligned}
 H(X) + H(Y) - H(X, Y) &= \sum_{x,y} (-p(x,y) \log_2 p(x) - p(x,y) \log_2 p(y) + p(x,y) \log_2 p(x,y)) \\
 &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\
 &= \sum_{x,y} p(x,y) f\left(\frac{p(x)p(y)}{p(x,y)}\right) \\
 &\geq f\left(\sum_{x,y} p(x,y) \frac{p(x)p(y)}{p(x,y)}\right) = f(1) = 0
 \end{aligned}$$

More generally, by iterating the above inequality for two random variables, we have

$$\begin{aligned}
 H(X_1, \dots, X_n) &\leq H(X_1, \dots, X_{n-1}) + H(X_n) \\
 &\leq H(X_1, \dots, X_{n-2}) + H(X_{n-1}) + H(X_n) \\
 &\leq \dots \leq H(X_1) + \dots + H(X_n).
 \end{aligned}$$

□

Remark 10.1.9 (Mutual information). The nonnegative quantity

$$I(X; Y) := H(X) + H(Y) - H(X, Y)$$

is called **mutual information**. Intuitively, it measures the amount of common infor-

mation between X and Y .

Lemma 10.1.10 (Dropping conditioning)

$H(X|Y) \leq H(X)$ and more generally,

$$H(X|Y, Z) \leq H(X|Z).$$

Proof. By chain rule and subadditivity, we have

$$H(X|Y) = H(X, Y) - H(Y) \leq H(X).$$

The inequality conditioning on Z follows since the above implies that

$$H(X|Y, Z = z) \geq H(X|Z = z)$$

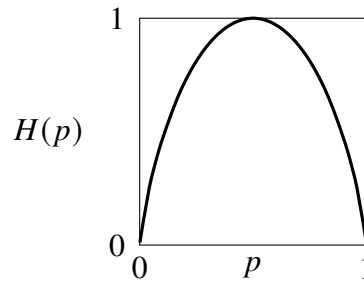
holds for every z , and taking expectation of z yields $H(X|Y, Z) \leq H(X|Z)$. \square

Remark 10.1.11. A related theorem is the **data processing inequality**: $H(X|f(Y)) \geq H(X|Y)$ for any function f . More generally, f can be random. In other words, if $X \rightarrow Y \rightarrow Z$ is a Markov chain, then $H(X|Z) \geq H(X|Y)$ (exercise: prove this).

Here are some simple applications of entropy to **tail bounds**.

Let us denote the entropy of a Bernoulli random variable by

$$H(p) := H(\text{Bernoulli}(p)) = -p \log_2 p - (1 - p) \log_2 (1 - p).$$



(This notation $H(\cdot)$ is standard but unfortunately ambiguous: $H(X)$ versus $H(p)$. It is usually clear from context which is meant.)

Theorem 10.1.12

If $0 < k \leq n/2$, then

$$\sum_{0 \leq i \leq k} \binom{n}{i} \leq 2^{H(k/n)n} = \left(\frac{n}{k}\right)^k \left(\frac{n}{n-k}\right)^{n-k}.$$

This bound can be established using our proof technique for Chernoff bound by applying Markov's inequality to the moment generating function:

$$\sum_{0 \leq i \leq k} \binom{n}{i} \leq \frac{(1+x)^n}{x^k} \quad \text{for all } x \in [0, 1].$$

The infimum of the RHS over $x \in [0, 1]$ is precisely $2^{H(k/n)n}$.

Now let us give a purely information theoretic proof to get some practice with entropy.

Proof. Let $(X_1, \dots, X_n) \in \{0, 1\}^n$ be chosen uniformly *conditioned* on $X_1 + \dots + X_n \leq k$. Then

$$\log_2 \sum_{0 \leq i \leq k} \binom{n}{i} = H(X_1, \dots, X_n) \leq H(X_1) + \dots + H(X_n).$$

Each X_i is a Bernoulli with probability $\mathbb{P}(X_i = 1)$. Note that conditioned on $X_1 + \dots + X_n = m$, one has $\mathbb{P}(X_i = 1) = m/n$. Varying over $m \leq k \leq n/2$, we find $\mathbb{P}(X_i = 1) \leq k/n$, so $H(X_i) \leq H(k/n)$. Hence

$$\log_2 \sum_{0 \leq i \leq k} \binom{n}{i} \leq H(k/n)n. \quad \square$$

Remark 10.1.13. One can extend the above proof to bound the tail of $\text{Binomial}(n, p)$ for any p . The result can be expressed in terms of the *relative entropy* (also known as the *Kullback–Leibler divergence* between two Bernoulli random variables). More concretely, for $X \sim \text{Binomial}(n, p)$, one has

$$\frac{\log \mathbb{P}(X \leq nq)}{n} \leq -q \log \frac{q}{p} - (1-q) \log \frac{1-q}{1-p} \quad \text{for all } 0 \leq q \leq p,$$

and

$$\frac{\log \mathbb{P}(X \geq nq)}{n} \leq -q \log \frac{q}{p} - (1-q) \log \frac{1-q}{1-p} \quad \text{for all } p \leq q \leq 1.$$

10.2 Permanent, perfect matchings, and Steiner triple systems

Permanent

We define the *permanent* of an $n \times n$ matrix A by

$$\text{per } A := \sum_{\sigma \in S_n} \prod_{i=1}^n a_{i, \sigma(i)}.$$

The formula for the permanent is simply that of the determinant without the sign factor:

$$\det A := \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n a_{i, \sigma(i)}.$$

We'll consider $\{0, 1\}$ -valued matrices. If A is the bipartite adjacency matrix of a bipartite graph, then

$\text{per } A =$ the number of perfect matchings.

The following theorem gives an upper bound on the number of perfect matchings of a bipartite graph with a given degree distribution. It was conjectured by [Minc \(1963\)](#) and proved by [Brégman \(1973\)](#).

Theorem 10.2.1 (Brégman–Minc inequality)

Let $A = (a_{ij}) \in \{0, 1\}^{n \times n}$, whose i -th row has sum d_i . Then

$$\text{per } A \leq \prod_{i=1}^n (d_i!)^{1/d_i}$$

Note that equality is attained when A consists diagonal blocks of 1's (corresponding to perfect matchings in a bipartite graph of the form $K_{d_1, d_1} \sqcup \dots \sqcup K_{d_t, d_t}$).

Let σ be a uniform random permutation of $[n]$ conditioned on $a_{i, \sigma(i)} = 1$ for all $i \in [n]$. Then

$$\log_2 \text{per } A = H(\sigma) = H(\sigma_1, \dots, \sigma_n) = H(\sigma_1) + H(\sigma_2 | \sigma_1) + \dots + H(\sigma_n | \sigma_1, \dots, \sigma_{n-1}).$$

We have

$$H(\sigma_i | \sigma_1, \dots, \sigma_{i-1}) \leq H(\sigma_i) \leq \log_2 |\text{support } \sigma_i| = \log_2 d_i,$$

but this step would be too lossy. In fact, what we just did amounts to a naive worst

case counting argument.

The key new idea is to **reveal the chosen entries in a uniform random order**.

Proof. (Radhakrishnan 1997) Let σ be as earlier. Consider a permutation of τ representing an ordering of the rows of the matrix. Say that i appears before j if $\tau_i < \tau_j$.

Let $N_i = N_i(\sigma, \tau)$ be the number of ones on row i that does not lie in the same column as some entry (j, σ_j) that comes before i . (Intuitively, N_i is the number of “greedily available” choices for σ_i before it is revealed.)

For any τ , the chain rule gives

$$H(\sigma) = \sum_{i=1}^n H(\sigma_i \mid \sigma_j : j \text{ comes before } i),$$

and the uniform bound gives

$$H(\sigma_i \mid \sigma_j : j \text{ comes before } i) \leq \mathbb{E}_\sigma \log_2 N_i.$$

Let τ vary uniformly over all permutations. Then,

$$H(\sigma) \leq \sum_{i=1}^n \mathbb{E}_{\sigma, \tau} \log_2 N_i.$$

For any fixed σ , as τ varies uniformly over all permutations of $[n]$, N_i varies uniformly over $[d_i]$. (Why?) Thus

$$\mathbb{E}_\tau \log_2 N_i = \frac{\log_2 1 + \cdots + \log_2 d_i}{d_i} = \frac{\log_2(d_i!)}{d_i}.$$

Taking expectation over σ and summing over i yields

$$\log_2 \text{per } A = H(\sigma) \leq \sum_{i=1}^n \mathbb{E}_{\sigma, \tau} \log_2 N_i \leq \sum_{i=1}^n \frac{\log_2(d_i!)}{d_i}. \quad \square$$

Corollary 10.2.2 (Kahn and Lovász)

Let G be a graph. Let d_v denote the degree of v . Then the number $\text{pm}(G)$ of perfect matchings of G satisfies

$$\text{pm}(G) \leq \prod_{v \in V(G)} (d_v!)^{1/(2d_v)} = \prod_{v \in V(G)} \text{pm}(K_{d_v, d_v})^{1/(2d_v)}.$$

Proof. (Alon and Friedland 2008) Brégman’s theorem implies the statement for bipar-

tite graphs G (by considering a bipartition on $G \sqcup G$). For the extension of non-bipartite G , one can proceed via a combinatorial argument that $\text{pm}(G \sqcup G) \leq \text{pm}(G \times K_2)$, which is left as an exercise. \square

The maximum number of Hamilton paths in a tournament

Question 10.2.3

What is the maximum possible number of directed Hamilton paths in an n -vertex tournament?

Earlier we saw that a uniformly random tournament has $n!/2^{n-1}$ Hamilton paths in expectation, and hence there is some tournament with at least this many Hamilton paths. This result, due to Szele, is the earliest application of the probabilistic method.

Using Brégman's theorem, Alon proved a nearly matching upper bound.

Theorem 10.2.4 (Alon 1990)

Every n -vertex tournament has at most $O(n^{3/2} \cdot n!/2^n)$ Hamilton paths.

Remark 10.2.5. The upper bound has been improved to $O(n^{3/2-\gamma} n!/2^n)$ for some small constant $\gamma > 0$ (Friedgut and Kahn 2005), while the lower bound $n!/2^{n-1}$ has been improved by a constant factor (Adler, Alon, and Ross 2001, Wormald 2004). It remains open to close this $n^{O(1)}$ factor gap.

We first prove an upper bound on the number of Hamilton cycles.

Theorem 10.2.6 (Alon 1990)

Every n -vertex tournament has at most $O(\sqrt{n} \cdot n!/2^n)$ Hamilton cycles.

Proof. Let A be an $n \times n$ matrix whose (i, j) entry is 1 if $i \rightarrow j$ is an edge of the tournament and 0 otherwise. Let d_i be the sum of the i -th row. Then $\text{per } A$ counts the number of 1-factors (spanning disjoint unions of directed cycles) of the tournament. So by Brégman's theorem, we have

$$\text{number of Hamilton cycles} \leq \text{per } A \leq \prod_{i=1}^n (d_i!)^{1/d_i}.$$

One can check (omitted) that the function $g(x) = (x!)^{1/x}$ is log-concave, i.e., $g(n)g(n+2) \geq g(n+1)^2$ for all $n \geq 0$. Thus, by a smoothing argument, among sequences (d_1, \dots, d_n) with sum $\binom{n}{2}$, the RHS above is maximized when all the d_i 's are within 1 of each other, which, by Stirling's formula, gives $O(\sqrt{n} \cdot n!/2^n)$. \square

Theorem 10.2.4 then follows by applying the above bound with the following lemma.

Lemma 10.2.7

Given an n -vertex tournament with P Hamilton paths, one can add a new vertex to obtain a $(n + 1)$ -vertex tournament with at least $P/4$ Hamilton cycles.

Proof. Add a new vertex and orient its incident edges uniformly at random. For every Hamilton path in the n -vertex tournament, there is probability $1/4$ that it can be closed up into a Hamilton cycle through the new vertex. The claim then follows by linearity of expectation. \square

Steiner triple systems

Definition 10.2.8 (Steiner triple system)

A **Steiner triple system (STS)** of order n is a 3-uniform hypergraph on n vertices where every pair of vertices is contained in exactly one triple.

Equivalently: an STS is a decomposition of a complete graph K_n into edge-disjoint triangles.

Example: the Fano plane is an STS of order 7.

It is a classic result that an STS of order n exists if and only if $n \equiv 1$ or $3 \pmod{6}$. It is not hard to see that this is necessary, since if an STS of order n exists, then $\binom{n}{2}$ should be divisible by 3, and $n - 1$ should be divisible by 2. [Keevash \(2014+\)](#) obtained a significant breakthrough proving the existence of more general designs.

Question 10.2.9

How many STS are there on n labeled vertices?

We shall prove the following result.

Theorem 10.2.10 (Upper bound on the number of STS — Linial and Luria 2013)

The number of Steiner triple systems on n labeled vertices is at most

$$\left(\frac{n}{e^2 + o(1)} \right)^{n^2}.$$

Remark 10.2.11. [Keevash \(2018\)](#) proved a matching lower bound when $n \equiv 1, 3 \pmod{6}$.

10.2 Permanent, perfect matchings, and Steiner triple systems

Proof. As in the earlier proof, the idea is to reveal the triples in a random order.

Let X denote a uniformly chosen STS on n vertices. We wish to upper bound $H(X)$.

We encode X as a tuple $(X_{ij})_{i < j} \in [n]^{\binom{n}{2}}$ where X_{ij} is the label of the unique vertex that forms a triple with i and j in the STS. Here whenever we write ij we mean the unordered pair $\{i, j\}$, i.e., an edge of K_n .

Let $y = (y_{ij})_{i < j} \in [0, 1]^{\binom{n}{2}}$, and we order the edges of K_n in decreasing y_{ij} :

$$kl < ij \quad \text{if} \quad y_{kl} > y_{ij}.$$

By the chain rule,

$$H(X) = \sum_{ij} H(X_{ij} \mid X_{kl} : kl < ij).$$

Let

$N_{ij} = N_{ij}(X, y)$ = the number of possibilities for X_{ij} after revealing X_{kl} for all $kl < ij$.

By the uniform bound, we have

$$H(X) \leq \sum_{ij} \mathbb{E}_X \log_2 N_{ij}.$$

Now let $y = (y_{ij})_{i < j} \in [0, 1]^{\binom{n}{2}}$ be chosen uniformly at random. We have

$$H(X) \leq \sum_{ij} \mathbb{E}_X \mathbb{E}_y \log_2 N_{ij}.$$

Write $y_{-ij} \in [0, 1]^{\binom{n}{2}-1}$ to mean y with the ij -coordinate removed. Let us bound $\mathbb{E}_{y_{-ij}} \log_2 N_{ij}$ as a function of y_{ij} .

We define ij shows up first in its triple to be the event that $ij < ik, jk$ where $k = X_{ij}$. We have, for any fixed X ,

$$\mathbb{P}_{y_{-ij}}(ij \text{ shows up first in its triple}) = \mathbb{P}_{y_{-ij}}(ij < ik, jk) = \mathbb{P}_{y_{-ij}}(y_{ij} > y_{ik}, y_{jk}) = y_{ij}^2.$$

If ij does not show up first in its triple, then X_{ij} has exactly one possibility (namely k) by the time it gets revealed, and so $N_{ij} = 1$ and $\log_2 N_{ij} = 0$. Thus

$$\begin{aligned} \mathbb{E}_{y_{-ij}} \log_2 N_{ij} &= y_{ij}^2 \mathbb{E}_{y_{-ij}} [\log_2 N_{ij} \mid ij \text{ shows up first in its triple}] \\ &\leq y_{ij}^2 \log_2 \mathbb{E}_{y_{-ij}} [N_{ij} \mid ij \text{ shows up first in its triple}]. \end{aligned}$$

Now we use linearity of expectations (over y_{-ij} with fixed X). For each $s \in [n] \setminus \{i, j, k\}$, if s is available as a possibility for X_{ij} by the time X_{ij} is revealed, then none

of the six edges of K_n consisting of the two triangle isX_{ij} and jsX_{js} may occur before X_{ij} ; the latter event occurs with probability y_{ij}^6 . So

$$\mathbb{E}_{y_{-ij}} [N_{ij} \mid ij \text{ shows up first in its triple}] \leq 1 + (n-3)y_{ij}^6.$$

Thus

$$\mathbb{E}_y \log_2 N_{ij} \leq \int_0^1 y_{ij}^2 \log_2(1 + (n-3)y_{ij}^3) dy_{ij} = \frac{1}{3} \int_0^1 \log_2(1 + (n-3)t^2) dt.$$

This integral actually has a closed-form antiderivative (e.g., check Mathematica/Wolfram Alpha), but it suffices for us to obtain the asymptotics. We have

$$\int_0^1 \log_2 \left(\frac{1}{n-3} + t^2 \right) dt \rightarrow \int_0^1 \log_2(t^2) dt = -2 \log_2 e$$

as $n \rightarrow \infty$ by the monotone convergence theorem. Thus

$$\mathbb{E}_y \log_2 N_{ij} \leq \frac{\log_2(n/e^2) + o(1)}{3}.$$

It follows therefore that the log-number of STS on n vertices is

$$H(X) \leq \sum_{ij} \mathbb{E}_X \mathbb{E}_y \log_2 N_{ij} \leq \binom{n}{2} \left(\frac{\log_2(n/e^2) + o(1)}{3} \right) = \frac{n^2}{6} \log_2 \left(\frac{n}{e^2 + o(1)} \right). \quad \square$$

Remark 10.2.12 (Guessing the formula). Here is perhaps how we might have guessed the formula for the number of STSs. Suppose we select $\frac{1}{3} \binom{n}{2}$ triangles in K_n independently at random. What is the probability that every edge is contained in exactly one triangle? Each edge is contained one triangle on expectation, and so by the Poisson approximation, the probability that a single fixed edge is contained in exactly one triangle is $1/e + o(1)$. Now let us pretend as if all the edges behave independently (!) — the probability that every edge is contained in exactly one triangle is $(1/e + o(1))^{\binom{n}{2}}$. This would then lead us to guessing that the number of STSs being

$$\frac{1}{\left(\frac{1}{3} \binom{n}{2}\right)!} \binom{n}{3}^{\frac{1}{3} \binom{n}{2}} \left(\frac{1}{e} + o(1) \right)^{\binom{n}{2}} = \left(\left(\frac{n^2}{6e} \right)^{-n^2/6} \left(\frac{n^3}{6} \right)^{n^2/6} \left(\frac{1}{e} \right)^{n^2/2} \right)^{1+o(1)} = \left(\frac{n}{e^2 + o(1)} \right)^{n^2/3}.$$

Here is another heuristic for getting the formula, and this time this method can actually be turned into a proof of matching lower bound on the number of STSs, though with a lot of work (Keevash 2018). Suppose we remove triangles from K_n one at a time. After k triangles have been removed, the number of edges remaining is $\binom{n}{2} - 3k$. Let us pretend that the remaining edges were randomly distributed. Then the number of

triangles should be about

$$\binom{n}{3} \left(1 - \frac{3k}{\binom{n}{2}}\right)^3 \sim \frac{36}{n^3} \left(\frac{1}{3} \binom{n}{2} - k\right)^3$$

If we multiply the above quantity over $0 \leq k < \frac{1}{3} \binom{n}{2}$, and then divide by $\left(\frac{1}{3} \binom{n}{2}\right)!$ to account for the ordering of the triangles, we get

$$\frac{\left(\frac{36}{n^3}\right)^{n^2/6} \left(\frac{1}{3} \binom{n}{2}\right)!^3}{\left(\frac{1}{3} \binom{n}{2}\right)!} \approx \left(\frac{n}{e^2 + o(1)}\right)^{n^2/3}.$$

10.3 Sidorenko's inequality

Given graphs F and G , a **graph homomorphism** from F to G is a map $\phi: V(F) \rightarrow V(G)$ of vertices that sends edges to edges, i.e., $\phi(u)\phi(v) \in E(G)$ for all $uv \in E(F)$.

Let

$\text{hom}(F, G)$ = the number of graph homomorphisms from F to G .

Define the **homomorphism density** (the **H-density in G**) by

$$\begin{aligned} t(F, G) &= \frac{\text{hom}(F, G)}{v(G)^{v(F)}} \\ &= \mathbb{P}(\text{a uniform random map } V(F) \rightarrow V(G) \text{ is a graph homomorphism } F \rightarrow G) \end{aligned}$$

In this section, we are interested in the regime of fixed F and large G , in which case almost all maps $V(F) \rightarrow V(G)$ are injective, so that there is not much difference between homomorphisms and subgraphs. More precisely,

$$\text{hom}(F, G) = \text{aut}(F)(\# \text{copies of } F \text{ in } G \text{ as a subgraph}) + O_F(v(G)^{v(F)-1}).$$

where $\text{aut}(F)$ is the number of automorphisms of F .

Inequalities between graph homomorphism densities is a central topic in extremal graph theory. For example, see Chapter 5 of my book *Graph Theory and Additive Combinatorics*. Much of the rest of this chapter is adapted from §5.5 of the book.

Question 10.3.1

Given a fixed graph F and constant $p \in [0, 1]$, what is the minimum possible F -density in a graph with edge density at least p ?

The F -density in the random graph $G(n, p)$ is $p^{e(F)} + o(1)$. Here p is fixed and $n \rightarrow \infty$.

Can one do better?

If F is non-bipartite, then the complete bipartite graph $K_{n/2, n/2}$ has F -density zero. (The problem of minimizing F -density is still interesting and not easy; it has been solved for cliques.)

[Sidorenko's conjecture \(1993\)](#) (also proposed by [Erdős and Simonovits \(1983\)](#)) says for any fixed bipartite F , the random graph asymptotically minimizes F -density. This is an important and well-known conjecture in extremal graph theory.

Conjecture 10.3.2 (Sidorenko)

For every bipartite graph F , and any graph G ,

$$t(F, G) \geq t(K_2, G)^{e(F)}.$$

The conjecture is known to hold for a large family of graphs F .

The entropy approach to Sidorenko's conjecture was first introduced by [Li and Szegedy \(2011\)](#) and later further developed in subsequent works. Here we illustrate the entropy approach to Sidorenko's conjecture with several examples.

We will construct a probability distribution μ on $\text{Hom}(F, G)$, the set of all graph homomorphisms $F \rightarrow G$. Unlike earlier applications of entropy, here we are trying to prove a lower bound on $\text{hom}(F, G)$ instead of an upper bound. So instead of taking μ to be a uniform distribution (which automatically has entropy $\log_2 \text{hom}(F, G)$), we actually take μ to be carefully constructed distribution, and apply the upper bound

$$H(\mu) \leq \log_2 |\text{support } \mu| = \log_2 \text{hom}(F, G).$$

We are trying to show that

$$\frac{\text{hom}(F, G)}{v(G)^{v(F)}} \geq \left(\frac{2e(G)}{v(G)^2} \right)^{e(F)}.$$

So we would like to find a probability distribution μ on $\text{Hom}(F, G)$ satisfying

$$H(\mu) \geq e(F) \log_2(2e(G)) - (2e(F) - v(F)) \log_2 v(G). \quad (10.1)$$

Theorem 10.3.3 (Blakey and Roy 1965)

Sidorenko's conjecture holds if F is a three-edge path.

Proof. We choose randomly a walk $XYZW$ in G as follows:

- XY is a uniform random edge of G (by this we mean first choosing an edge of G uniformly at random, and then let X be a uniformly chosen endpoint of this edge, and then Y the other endpoint);
- Z is a uniform random neighbor of Y ;
- W is a uniform random neighbor of Z .

Key observation: YZ is distributed as a uniform random edge of G , and likewise with ZW

Indeed, conditioned on the choice of Y , the vertices X and Z are both independent and uniform neighbors of Y , so XY and YZ are uniformly distributed.

Also, the conditional independence observation implies that

$$H(Z|X, Y) = H(Z|Y) \quad \text{and} \quad H(W|X, Y, Z) = H(W|Z)$$

and furthermore both quantities are equal to $H(Y|X)$ since XY, YZ, ZW are each distributed as a uniform random edge.

Thus

$$\begin{aligned} H(X, Y, Z, W) &= H(X) + H(Y|X) + H(Z|X, Y) + H(W|X, Y, Z) && \text{[chain rule]} \\ &= H(X) + H(Y|X) + H(Z|Y) + H(W|Z) && \text{[cond indep]} \\ &= H(X) + 3H(Y|X) \\ &= 3H(X, Y) - 2H(X) && \text{[chain rule]} \\ &\geq 3 \log_2(2e(G)) - 2 \log_2 v(G) \end{aligned}$$

In the final step we used $H(X, Y) = \log_2(2e(G))$ since XY is uniformly distributed among edges, and $H(X) \leq \log_2 |\text{support}(X)| = \log_2 v(G)$. This proves (10.1) and hence the theorem for a path of 4 vertices. (As long as the final expression has the “right form” and none of the steps are lossy, the proof should work out.) \square

Remark 10.3.4. See [this MathOverflow discussion](#) for the history as well as alternate proofs.

The above proof easily generalizes to all trees. We omit the details.

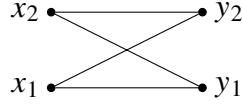
Theorem 10.3.5

Sidorenko's conjecture holds if F is a tree.

Theorem 10.3.6

Sidorenko's conjecture holds for all complete bipartite graphs.

Proof. Following the same framework as earlier, let us demonstrate the result for $F = K_{2,2}$. The same proof extends to all $K_{s,t}$.



We will pick a random tuple $(X_1, X_2, Y_1, Y_2) \in V(G)^4$ with $X_i Y_j \in E(G)$ for all i, j as follows.

- $X_1 Y_1$ is a uniform random edge;
- Y_2 is a uniform random neighbor of X_1 ;
- X_2 is a conditionally independent copy of X_1 given (Y_1, Y_2) .

The last point deserves more attention. Note that we are *not* simply uniformly randomly choosing a common neighbor of Y_1 and Y_2 as one might naively attempt. Instead, one can think of the first two steps as generating a distribution for (X_1, Y_1, Y_2) —according to this distribution, we first generate (Y_1, Y_2) according to its marginal, and then produce two conditionally independent copies of X_1 (the second copy is X_2).

As in the previous proof (applied to a 2-edge path), we see that

$$H(X_1, Y_1, Y_2) = 2H(X_1, Y_1) - H(X_1) \geq 2 \log_2(2e(G)) - \log_2 v(G).$$

So we have

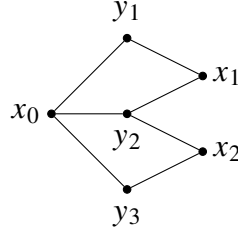
$$\begin{aligned}
 & H(X_1, X_2, Y_1, Y_2) \\
 &= H(Y_1, Y_2) + H(X_1, X_2 | Y_1, Y_2) && \text{[chain rule]} \\
 &= H(Y_1, Y_2) + 2H(X_1 | Y_1, Y_2) && \text{[conditional independence]} \\
 &= 2H(X_1, Y_1, Y_2) - H(Y_1, Y_2) && \text{[chain rule]} \\
 &\geq 2(2 \log_2(2e(G)) - \log_2 v(G)) - 2 \log_2 v(G). && \text{[prev. ineq. and uniform bound]} \\
 &= 4 \log_2(2e(G)) - 4 \log_2 v(G).
 \end{aligned}$$

So we have verified (10.1) for $K_{2,2}$. □

Theorem 10.3.7 (Conlon, Fox, Sudakov 2010)

Sidorenko's conjecture holds for a bipartite graph that has a vertex adjacent to all vertices in the other part.

Proof. Let us illustrate the proof for the following graph. The proof extends to the general case.



Let us choose a random tuple $(X_0, X_1, X_2, Y_1, Y_2, Y_3) \in V(G)^6$ as follows:

- X_0Y_1 is a uniform random edge;
- Y_2 and Y_3 are independent uniform random neighbors of X_0 ;
- X_1 is a conditionally independent copy of X_0 given (Y_1, Y_2) ;
- X_2 is a conditionally independent copy of X_0 given (Y_2, Y_3) .

(as well as other symmetric versions.) Some important properties of this distribution:

- X_0, X_1, X_2 are conditionally independent given (Y_1, Y_2, Y_3) ;
- X_1 and (X_0, Y_3, X_2) are conditionally independent given (Y_1, Y_2) ;
- The distribution of (X_0, Y_1, Y_2) is identical to the distribution of (X_1, Y_1, Y_2) .

We have

$$\begin{aligned}
 & H(X_0, X_1, X_2, Y_1, Y_2, Y_3) \\
 &= H(X_0, X_1, X_2 | Y_1, Y_2, Y_3) + H(Y_1, Y_2, Y_3) && \text{[chain rule]} \\
 &= H(X_0 | Y_1, Y_2, Y_3) + H(X_1 | Y_1, Y_2, Y_3) + H(X_2 | Y_1, Y_2, Y_3) + H(Y_1, Y_2, Y_3) && \text{[cond indep]} \\
 &= H(X_0 | Y_1, Y_2, Y_3) + H(X_1 | Y_1, Y_2) + H(X_2 | Y_2, Y_3) + H(Y_1, Y_2, Y_3) && \text{[cond indep]} \\
 &= H(X_0, Y_1, Y_2, Y_3) + H(X_1, Y_1, Y_2) + H(X_2, Y_2, Y_3) - H(Y_1, Y_2) - H(Y_2, Y_3). && \text{[chain rule]}
 \end{aligned}$$

The proof of Theorem 10.3.3 actually lower bounds the first three terms:

$$\begin{aligned}
 H(X_0, Y_1, Y_2, Y_3) &\geq 3 \log_2(2e(G)) - 2 \log_2 v(G) \\
 H(X_1, Y_1, Y_2) &\geq 2 \log_2(2e(G)) - \log_2 v(G) \\
 H(X_2, Y_2, Y_3) &\geq 2 \log_2(2e(G)) - \log_2 v(G).
 \end{aligned}$$

10 Entropy method

We can apply the uniform support bound on the remaining terms.

$$H(Y_1, Y_2) = H(Y_2, Y_3) \leq 2 \log_2 v(G).$$

Putting everything together, we have

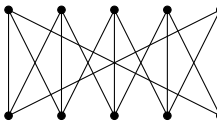
$$H(X_0, X_1, X_2, Y_1, Y_2, Y_3) \geq 7 \log_2(2e(G)) - 8 \log_2 v(G),$$

thereby verifying (10.1). \square

To check that you understand the above proof, where did we use the assumption that F has a vertex complete to the other part?

Many other graphs can be proved by extending this method.

Remark 10.3.8 (Möbius graph). An important open case (and the smallest in some sense) of Sidorenko conjecture is when F is the following graph, known as the **Möbius graph**. It is $K_{5,5}$ with a C_{10} removed. The name comes from it being the face-vertex incidence graph of the simplicial complex structure of the Möbius strip, built by gluing a strip of five triangles.

$$\text{Möbius graph} = K_{5,5} \setminus C_{10} =$$


10.4 Shearer's lemma

Shearer's entropy lemma extends the subadditivity property of entropy. Before stating it in full generality, let us first see the simplest instance of Shearer's lemma.

Theorem 10.4.1 (Shearer's lemma, special case)

$$2H(X, Y, Z) \leq H(X, Y) + H(X, Z) + H(Y, Z)$$

Proof. Using the chain rule and conditioning dropping, we have

$$\begin{aligned} H(X, Y) &= H(X) + H(Y|X) \\ H(X, Z) &= H(X) + H(Z|X) \geq H(X) + H(Z|X, Z) \\ H(Y, Z) &= H(Y) + H(Z|Y) \geq H(Y|X) + H(Z|X, Y) \end{aligned}$$

Applying conditioning dropping, we see that their sum is at least

$$2H(X) + 2H(Y|X) + 2H(Z|X, Y) = 2H(X, Y, Z). \quad \square$$

Question 10.4.2

What is the maximum volume of a body in \mathbb{R}^3 that has area at most 1 when projected to each of the three coordinate planes?

The cube $[0, 1]^3$ satisfies the above property and has area 1. It turns out that this is the maximum.

To prove this claim, first let us use Shearer's inequality to prove a discrete version.

Theorem 10.4.3

Let $S \subseteq \mathbb{R}^3$ be a finite set, and $\pi_{xy}(S)$ be its projection on the xy -plane, etc. Then

$$|S|^2 \leq |\pi_{xy}(S)| |\pi_{xz}(S)| |\pi_{yz}(S)|$$

Proof. Let (X, Y, Z) be a uniform random point of S . Then

$$\begin{aligned} 2 \log_2 |S| = 2H(X, Y, Z) &\leq H(X, Y) + H(X, Z) + H(Y, Z) \\ &\leq \log_2 |\pi_{xy}(S)| + \log_2 |\pi_{xz}(S)| + \log_2 |\pi_{yz}(S)|. \quad \square \end{aligned}$$

By approximating a body using cubes, we can deduce the following corollary.

Corollary 10.4.4

Let S be a body in \mathbb{R}^3 . Then

$$\text{vol}(S)^2 \leq \text{area}(\pi_{xy}(S)) \text{area}(\pi_{xz}(S)) \text{area}(\pi_{yz}(S)).$$

Let us now state the general form of Shearer's lemma. (Chung, Graham, Frankl, and Shearer 1986)

Theorem 10.4.5 (Shearer's lemma)

Let $A_1, \dots, A_s \subseteq [n]$ where each $i \in [n]$ appears in at least k sets A_j 's. Writing $X_A := (X_i)_{i \in A}$,

$$kH(X_1, \dots, X_n) \leq \sum_{j \in [s]} H(X_{A_j}).$$

The proof of the general form of Shearer's lemma is a straightforward adaptation of the proof of the special case earlier.

Like earlier, we can deduce an inequality about sizes of projections. (Loomis and Whitney 1949)

Corollary 10.4.6 (Loomis–Whitney inequality)

Writing π_i for the projection from \mathbb{R}^n onto the hyperplane $x_i = 0$, we have for every $S \subseteq \mathbb{R}^n$,

$$|S|^{n-1} \leq \prod_{i=1}^n |\pi_i(S)|$$

Corollary 10.4.7

Let $A_1, \dots, A_s \subseteq \Omega$ where each $i \in \Omega$ appears in at least k sets A_j . Then for every family \mathcal{F} of subsets of Ω ,

$$|\mathcal{F}|^k \leq \prod_{j \in [s]} |\mathcal{F}|_{A_j}|$$

where $\mathcal{F}|_A := \{F \cap A : F \in \mathcal{F}\}$.

Proof. Each subset of Ω corresponds to a vector $(X_1, \dots, X_n) \in \{0, 1\}^n$. Let (X_1, \dots, X_n) be a random vector corresponding to a uniform element of \mathcal{F} . Then

$$k \log_2 |\mathcal{F}| = kH(X_1, \dots, X_n) \leq \sum_{j \in [s]} H(X_{A_j}) = \log_2 |\mathcal{F}|_{A_j}|. \quad \square$$

Triangle-intersecting families

We say that a set \mathcal{G} of labeled graphs on the same vertex set is *triangle-intersecting* if $G \cap G'$ contains a triangle for every $G, G' \in \mathcal{G}$.

Question 10.4.8

What is the largest triangle-intersecting family of graphs on n labeled vertices?

The set of all graphs that contain a fixed triangle is triangle-intersecting, and they form a $1/8$ fraction of all graphs.

An easy upper bound: the edges form an intersecting family, so a triangle-intersecting family must be at most $1/2$ fraction of all graphs.

The next theorem improves this upper bound to $< 1/4$. It is also in this paper that Shearer's lemma was introduced.

Theorem 10.4.9 (Chung, Graham, Frankl, and Shearer 1986)

Every triangle-intersecting family of graphs on n labeled vertices has size $< 2^{\binom{n}{2}-2}$.

Proof. Let \mathcal{G} be a triangle-intersecting family of graphs on vertex set $[n]$ (viewed as a collection of subsets of edges of K_n)

For $S \subseteq [n]$ with $|S| = \lfloor n/2 \rfloor$, let $A_S = \binom{S}{2} \cup \binom{[n] \setminus S}{2}$ (i.e., A_S is the union of the clique on S and the clique on the complement of S). Let

$$r = |A_S| = \binom{\lfloor n/2 \rfloor}{2} + \binom{\lceil n/2 \rceil}{2} \leq \frac{1}{2} \binom{n}{2}.$$

For every S , every triangle has an edge in A_S , and thus \mathcal{G} restricted to A_S must be an intersecting family. Hence

$$|\mathcal{G}|_{A_S} \leq 2^{|A_S|-1} = 2^{r-1}.$$

Each edge of K_n appears in at least

$$k = \frac{r}{\binom{n}{2}} \binom{n}{\lfloor n/2 \rfloor}$$

different A_S with $|S| = \lfloor n/2 \rfloor$ (by symmetry and averaging). Applying Corollary 10.4.7, we find that

$$|\mathcal{G}|^k \leq \left(2^{r-1}\right)^{\binom{n}{\lfloor n/2 \rfloor}}.$$

Therefore

$$|\mathcal{G}| \leq 2^{\binom{n}{2} - \frac{\binom{n}{2}}{r}} < 2^{\binom{n}{2}-2}. \quad \square$$

Remark 10.4.10. A tight upper bound of $2^{\binom{n}{2}-3}$ (matching the construction of taking all graphs containing a fixed triangle) was conjectured by Simonovits and Sós (1976) and proved by Ellis, Filmus, and Friedgut (2012) using Fourier analytic methods. Berger and Zhao (2021+) gave a tight solution for K_4 -intersecting families. The general conjecture for K_r -intersecting families is open.

The number of independent sets in a regular bipartite graph

Question 10.4.11

Fix d . Which d -regular graph on a given number of vertices has the most number of independent sets? Alternatively, which graph G maximizes $i(G)^{1/v(G)}$?

(Note that the number of independent sets is multiplicative: $i(G_1 \sqcup G_2) = i(G_1)i(G_2)$.)

Alon and Kahn conjectured that for graphs on n vertices, when n is a multiple of $2d$, a disjoint union of $K_{d,d}$'s maximizes the number of independent sets.

Alon (1991) proved an approximate version of this conjecture. Kahn (2001) proved it assuming the graph is bipartite. Zhao (2010) proved it in general.

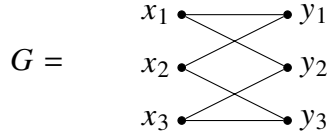
Theorem 10.4.12 (Kahn, Zhao)

Let G be an n -vertex d -regular graph. Then

$$i(G) \leq i(K_{d,d})^{n/(2d)} = (2^{d+1} - 1)^{n/(2d)}$$

where $i(G)$ is the number of independent sets of G .

Proof assuming G is bipartite. (Kahn) Let us first illustrate the proof for



Among all independent sets of G , choose one uniformly at random, and let $(X_1, X_2, X_3, Y_1, Y_2, Y_3) \in \{0, 1\}^6$ be its indicator vector. Then

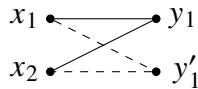
$$\begin{aligned} 2 \log_2 i(G) &= 2H(X_1, X_2, X_3, Y_1, Y_2, Y_3) \\ &= 2H(X_1, X_2, X_3) + 2H(Y_1, Y_2, Y_3 | X_1, X_2, X_3) && \text{[chain rule]} \\ &\leq H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3) \\ &\quad + 2H(Y_1 | X_1, X_2, X_3) + 2H(Y_2 | X_1, X_2, X_3) + 2H(Y_3 | X_1, X_2, X_3) && \text{[Shearer]} \\ &= H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3) \\ &\quad + 2H(Y_1 | X_1, X_2) + 2H(Y_2 | X_1, X_3) + 2H(Y_3 | X_2, X_3) && \text{[cond indep]} \end{aligned}$$

Here we are using that (a) Y_1, Y_2, Y_3 are conditionally independent given (X_1, X_2, X_3) and (b) Y_1 and (X_3, Y_2, Y_3) are conditionally independent given (X_1, X_2) . A more general statement is that if $S \subseteq V(G)$, then the restrictions to the different connected components of $G - S$ are conditionally independent given X_S .

It remains to prove that

$$H(X_1, X_2) + 2H(Y_1 | X_1, X_2) \leq \log_2 i(K_{2,2})$$

and two other analogous inequalities. Let Y'_1 be conditionally independent copy of Y_1 given (X_1, X_2) . Then (X_1, X_2, Y_1, Y'_1) is the indicator vector of an independent set of $K_{2,2}$ (though not necessarily chosen uniformly).



Thus we have

$$\begin{aligned}
H(X_1, X_2) + 2H(Y_1|X_1, X_2) &= H(X_1, X_2) + H(Y_1|X_1, X_2) + H(Y'_1|X_1, X_2) \\
&= H(X_1, X_2, Y_1, Y'_1) && \text{[chain rule]} \\
&\leq \log_2 i(G) && \text{[uniform bound]}
\end{aligned}$$

This concludes the proof for $G = K_{2,2}$, which works for all bipartite G . Here are the details.

Let $V = A \cup B$ be the vertex bipartition of G . Let $X = (X_v)_{v \in V}$ be the indicator function of an independent set chosen uniformly at random. Write $X_S := (X_v)_{v \in S}$. We have

$$\begin{aligned}
d \log_2 i(G) &= dH(X) = dH(X_A) + dH(X_B|X_A) && \text{[chain rule]} \\
&\leq \sum_{b \in B} H(X_{N(b)}) + d \sum_{b \in B} H(X_b|X_A) && \text{[Shearer]} \\
&\leq \sum_{b \in B} H(X_{N(b)}) + d \sum_{b \in B} H(X_b|X_{N(b)}) && \text{[drop conditioning]}
\end{aligned}$$

For each $b \in B$, we have

$$\begin{aligned}
H(X_{N(b)}) + dH(X_b|X_{N(b)}) &= H(X_{N(b)}) + H(X_b^{(1)}, \dots, X_b^{(d)}|X_{N(b)}) \\
&= H(X_b^{(1)}, \dots, X_b^{(d)}, X_{N(b)}) \\
&\leq \log_2 i(K_{d,d})
\end{aligned}$$

where $X_b^{(1)}, \dots, X_b^{(d)}$ are conditionally independent copies of X_b given $X_{N(b)}$. Summing over all b yields the result. \square

Now we give the argument from [Zhao \(2010\)](#) that removes the bipartite hypothesis. The following combinatorial argument reduces the problem for non-bipartite G to that of bipartite G .

Starting from a graph G , we construct its **bipartite double cover** $G \times K_2$ (see Figure 10.1), which has vertex set $V(G) \times \{0, 1\}$. The vertices of $G \times K_2$ are labeled v_i for $v \in V(G)$ and $i \in \{0, 1\}$. Its edges are u_0v_1 for all $uv \in E(G)$. Note that $G \times K_2$ is always a bipartite graph.

Lemma 10.4.13

Let G be any graph (not necessarily regular). Then

$$i(G)^2 \leq i(G \times K_2).$$

Once we have the lemma, Theorem 10.4.12 then reduces to the bipartite case, which we already proved. Indeed, for a d -regular G , since $G \times K_2$ is bipartite, the bipartite

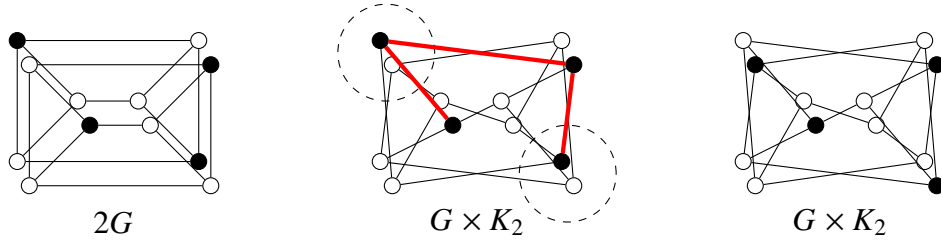


Figure 10.1: The bipartite swapping trick in the proof of Lemma 10.4.13: swapping the circled pairs of vertices (denoted A in the proof) fixes the bad edges (red and bolded), transforming an independent set of $2G$ into an independent set of $G \times K_2$.

case of the theorem gives

$$i(G)^2 \leq i(G \times K_2) \leq i(K_{d,d})^{n/d},$$

Proof of Lemma 10.4.13. Let $2G$ denote a disjoint union of two copies of G . Label its vertices by v_i with $v \in V$ and $i \in \{0, 1\}$ so that its edges are $u_i v_i$ with $uv \in E(G)$ and $i \in \{0, 1\}$. We will give an injection $\phi: I(2G) \rightarrow I(G \times K_2)$. Recall that $I(G)$ is the set of independent sets of G . The injection would imply $i(G)^2 = i(2G) \leq i(G \times K_2)$ as desired.

Fix an arbitrary order on all subsets of $V(G)$. Let S be an independent set of $2G$. Let

$$E_{\text{bad}}(S) := \{uv \in E(G) : u_0, v_1 \in S\}.$$

Note that $E_{\text{bad}}(S)$ is a bipartite subgraph of G , since each edge of E_{bad} has exactly one endpoint in $\{v \in V(G) : v_0 \in S\}$ but not both (or else S would not be independent). Let A denote the first subset (in the previously fixed ordering) of $V(G)$ such that all edges in $E_{\text{bad}}(S)$ have one vertex in A and the other outside A . Define $\phi(S)$ to be the subset of $V(G) \times \{0, 1\}$ obtained by “swapping” the pairs in A , i.e., for all $v \in A$, $v_i \in \phi(S)$ if and only if $v_{1-i} \in S$ for each $i \in \{0, 1\}$, and for all $v \notin A$, $v_i \in \phi(S)$ if and only if $v_i \in S$ for each $i \in \{0, 1\}$. It is not hard to verify that $\phi(S)$ is an independent set in $G \times K_2$. The swapping procedure fixes the “bad” edges.

It remains to verify that ϕ is an injection. For every $S \in I(2G)$, once we know $T = \phi(S)$, we can recover S by first setting

$$E'_{\text{bad}}(T) = \{uv \in E(G) : u_i, v_i \in T \text{ for some } i \in \{0, 1\}\},$$

so that $E_{\text{bad}}(S) = E'_{\text{bad}}(T)$, and then finding A as earlier and swapping the pairs of A back. (Remark: it follows that $T \in I(G \times K_2)$ lies in the image of ϕ if and only if $E'_{\text{bad}}(T)$ is bipartite.) \square

The entropy proof of the bipartite case of Theorem 10.4.12 extends to graph homomorphisms, yielding the following result.

Theorem 10.4.14 (Galvin and Tetali 2004)

Let G be an n -vertex d -regular bipartite graph. Let H be any graph allowing loops. Then

$$\text{hom}(G, H) \leq \text{hom}(K_{d,d}, H)^{n/(2d)}$$

Some important special cases:

- $\text{hom}(G, \text{⬤} \text{---} \text{⬤}) = i(G)$, the number of independent sets of G ;
- $\text{hom}(G, K_q) =$ the number of proper q -colorings of G .

The bipartite hypothesis in Theorem 10.4.14 cannot be always be removed. For example, if $H = \text{⬤} \text{---} \text{⬤}$, then $\log_2 \text{hom}(G, H)$ is the number of connected components of G , so that the maximizers of $\log_2 \text{hom}(G, H)/v(G)$ are disjoint unions of K_{d+1} 's.

For $H = K_q$, corresponding to the proper q -colorings, the bipartite hypothesis was recently removed.

Theorem 10.4.15 (Sah, Sawhney, Stoner, and Zhao 2020)

Let G be an n -vertex d -regular graph. Then

$$c_q(G) \leq c_q(K_{d,d})^{n/(2d)}$$

where $c_q(G)$ is the number of q -colorings of G .

Furthermore, it was also shown in the same paper that in Theorem 10.4.14, the bipartite hypothesis on G can be weakened to triangle-free. Furthermore triangle-free is the weakest possible hypothesis on G so that the claim is true for all H .

For more discussion and open problems on this topic, see the survey by Zhao (2017).

11 The container method

Many problems in combinatorics can be phrased in terms of independent sets in hypergraphs.

For example, here is a model question:

Question 11.0.1

How many triangle-free graphs are there on n vertices?

By taking all subgraphs of $K_{n/2, n/2}$, we obtain $2^{n^2/4}$ such graphs. It turns out this gives the correct exponential asymptotic.

Theorem 11.0.2 (Erdős, Kleitman, and Rothschild 1973)

The number of triangle-free graphs on n vertices is $2^{n^2/4+o(n^2)}$.

Remark 11.0.3. It does not matter here whether we consider vertices to be labeled, it affects the answer up to a factor of at most $n! = e^{O(n \log n)}$.

Remark 11.0.4. Actually the original Erdős–Kleitman–Rothschild paper showed an even stronger result: $1 - o(1)$ fraction of all n -vertex triangle-free graphs are bipartite. The above asymptotic can be then easily deduced by counting subgraphs of complete bipartite graphs. The container methods discussed in this section are not strong enough to prove this finer claim.

We can convert this asymptotic enumeration problem into a problem about independent sets in a 3-uniform hypergraph H :

- $V(H) = \binom{[n]}{2}$
- The edges of H are triples of the form $\{xy, xz, yx\}$, i.e., triangles

We then have the correspondence:

- A subset of $V(H)$ = a graph on vertex set $[n]$
- An independent set of $V(H)$ = a triangle-free graph

(Here an *independent set* in a hypergraph is a subset of vertices containing no edges.)

Naively applying first moment/union bound does not work—there are too many events to union bound over.

For example, Mantel’s theorem tell us the maximum number of edges in an n -vertex triangle-free graph is $\lfloor n^2/4 \rfloor$, obtained by $K_{\lfloor n/2 \rfloor, \lceil n/2 \rceil}$. With a fixed triangle-free graph G , the number of subgraphs of G is $2^{e(G)}$, and each of them is triangle-free. Perhaps we could union bound over all maximal triangle-free graphs? It turns out that there are $2^{n^2/8+o(n^2)}$ such maximal triangle-free graphs, so a union bound would be too wasteful.

In many applications, independent sets are clustered into relatively few highly correlated sets. In the case of triangle-free graphs, each maximal triangle-free graph is very “close” to many other maximal triangle-free graphs.

Is there a more efficient union bound that takes account of the clustering of independent sets?

The container method does exactly that. Given some hypergraph with controlled degrees, one can find a collection of *containers* satisfying the following properties:

- Each container is a subset of vertices of the hypergraph.
- Every independent set of the hypergraph is a subset of some container.
- The total number of containers in the collection is relatively small.
- Each container is not too large (in fact, not too much larger than the maximum size of an independent set)

We can then union bound over all such containers. If the number of containers is not too small, then the union bound is not too lossy.

Here are some of the most typical and important applications of the container method:

- Asymptotic enumerations:
 - Counting H -free graphs on n vertices
 - Counting H -free graphs on n vertices and m edges
 - Counting k -AP-free subsets of $[n]$ of size m
- Extremal and Ramsey results in random structures:
 - The maximum number of edges in an H -free subgraph of $G(n, p)$
 - Szemerédi’s theorem in a p -random subset of $[n]$
- List coloring in graphs/hypergraphs

The method of hypergraph containers is one of the most exciting developments in this past decade. Some references and further reading:

- The graph container method was developed by [Kleitman and Winston \(1982\)](#) (for counting C_4 -free graphs) and [Sapozhenko \(2001\)](#) (for bounding the number of independent sets in a regular graph, giving an earlier version of Theorem 10.4.12)
- The hypergraph container theorem was proved independently by [Balogh, Morris, and Samotij \(2015\)](#), and [Saxton and Thomason \(2015\)](#).
- See the [2018 ICM survey of Balogh, Morris, and Samotij](#) for an introduction to the topic along with many applications
- See [Samotij's survey article \(2015\)](#) for an introduction to the graph container method
- See [Morris' lecture notes \(2016\)](#) for a gentle introduction to the proof and applications of hypergraph containers.

11.1 Containers for triangle-free graphs

The number of triangle-free graphs

We will prove Theorem 11.0.2 that the number of triangle-free graphs on n vertices is $2^{n^2/4+o(n^2)}$.

Theorem 11.1.1 (Containers for triangle-free graphs)

For every $\varepsilon > 0$, there exists $C > 0$ such that the following holds.

For every n , there is a collection C of graphs on n vertices, with

$$|C| \leq n^{Cn^{3/2}}$$

such that

- (a) every $G \in C$ has at most $(\frac{1}{4} + \varepsilon)n^2$ edges, and
- (b) every triangle-free graph is contained in some $G \in C$.

Proof of upper bound of Theorem 11.0.2. We want to show that the number of n -vertex triangle-free graphs is at most $2^{n^2/4+o(n^2)}$. Let $\varepsilon > 0$ be any real number (arbitrarily small). Let C be produced by Theorem 11.1.1.

Then every $G \in C$ has at most $(\frac{1}{4} + \varepsilon)n^2$ edges, and every triangle-free graph is

11 The container method

contained in some $G \in \mathcal{C}$. Hence the number of triangle-free graphs is

$$|\mathcal{C}| 2^{\left(\frac{1}{4}+\delta\right)n^2} \leq 2^{\left(\frac{1}{4}+\varepsilon\right)n^2+O_\varepsilon(n^{3/2}\log n)}.$$

Since $\varepsilon > 0$ can be made arbitrarily small, the number triangle-free graphs on n vertices is $2^{(\frac{1}{4}+o(1))n^2}$. \square

The same proof technique, with an appropriate container theorem, can be used to count H -free graphs.

We write $\text{ex}(n, H)$ for the maximum number of edges in an n -vertex graph without H as a subgraph.

Theorem 11.1.2 (Erdős–Stone–Simonovits)

Fix a non-bipartite graph H . Then

$$\text{ex}(n, H) = \left(1 - \frac{1}{\chi(H) - 1} + o(1)\right) \binom{n}{2}.$$

Note that for bipartite graphs H , the above theorem just says $o(n^2)$, though more precise estimates are available. Although we do not know the asymptotic for $\text{ex}(n, H)$ for all H , e.g., it is still open for $H = K_{4,4}$ and $H = C_8$.

Theorem 11.1.3

Fix a non-bipartite graph H . Then the number of H -free graphs on n vertices is $2^{(1+o(1))\text{ex}(n,H)}$.

The analogous statement for bipartite graphs is false. The following conjecture remains of great interest, and it is known for certain graphs, e.g., $H = C_4$.

Conjecture 11.1.4

Fix a bipartite graph H with a cycle. The number of H -free graphs on n vertices is $2^{O(\text{ex}(n,H))}$.

Mantel's theorem in random graphs

Theorem 11.1.5

If $p \gg 1/\sqrt{n}$, then with probability $1 - o(1)$, every triangle-free subgraph of $G(n, p)$ has at most $(\frac{1}{4} + o(1))pn^2$ edges.

Remark 11.1.6. In fact, a much stronger result is true: the triangle-free subgraph of $G(n, p)$ with the maximum number of edges is whp bipartite (DeMarco and Kahn 2015).

Remark 11.1.7. The statement is false for $p \ll 1/\sqrt{n}$. Indeed, in this case, then the expected number of triangles is $O(n^3 p^3)$, whereas there are whp $n^2 p/2$ edges, and $n^3 p^3 \ll n^2 p$, so we can remove $o(n^2 p)$ edges to make the graph triangle-free.

Proof. We prove a slightly weaker result, namely that the result is true if $p \gg n^{-1/2} \log n$. The version with $p \gg n^{-1/2}$ can be proved via a stronger formulation of the container lemma (using “fingerprints” as discussed later).

Let $\varepsilon > 0$ be arbitrarily small. Let C be a set of containers for n -vertex triangle-free graphs in Theorem 11.1.1. For every $G \in C$, $e(G) \leq \left(\frac{1}{4} + \varepsilon\right) n^2$, so by an application of the Chernoff bound,

$$\mathbb{P}\left(e(G \cap G(n, p)) > \left(\frac{1}{4} + 2\varepsilon\right) n^2 p\right) \leq e^{-\Omega_\varepsilon(n^2 p)}$$

Since every triangle-free graph is contained in some $G \in C$, by taking a union bound over C , we see that

$$\begin{aligned} & \mathbb{P}\left(G(n, p) \text{ has a triangle-free subgraph with } > \left(\frac{1}{4} + 2\varepsilon\right) n^2 p \text{ edges}\right) \\ & \leq \sum_{G \in C} \mathbb{P}\left(e(G \cap G(n, p)) > \left(\frac{1}{4} + 2\varepsilon\right) n^2 p\right) \\ & \leq |C| e^{-\Omega_\varepsilon(n^2 p)} \\ & \leq e^{O_\varepsilon(n^{3/2} \log n) - \Omega_\varepsilon(n^2 p)} \\ & = o(1) \end{aligned}$$

provided that $p \gg n^{-1/2} \log n$. □

11.2 Graph containers

Theorem 11.2.1 (Container theorem for independent sets in graphs)

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 . There exists a collection \mathcal{C} of subsets of V , with

$$|\mathcal{C}| \leq \binom{|V|}{\leq 2\delta |V|/d}$$

such that

1. Every independent set I of G is contained in some $C \in \mathcal{C}$.
2. $|C| \leq (1 - \delta) |V|$ for every $C \in \mathcal{C}$.

Each $C \in \mathcal{C}$ is called a “container.” Every independent set of G is contained in some container.

Remark 11.2.2. The requirement $|C| \leq (1 - \delta) |V|$ looks quite a bit weaker than in Theorem 11.1.1, where each container is only slightly larger than the maximum independent set. In a typical application of the container method, one iteratively applies the (hyper)graph container theorem (e.g., Theorem 11.2.1 and later Theorem 11.3.1) to the subgraphs induced by the slightly smaller containers in the previous iteration. One iterates until the containers are close to their minimum possible size.

For this iterative application of container theorem to work, one usually needs a **super-saturation** result, which, roughly speaking, says that every subset of vertices that is slightly larger than the independence number necessarily induces a lot of edges. This property is common to all standard applications of the container method.

The container theorem is proved using

The graph container algorithm (for a fixed given graph G)

Input: a maximal independent set $I \subseteq V$.

Output: a “fingerprint” $S \subseteq I$ of size $\leq 2\delta |V|/d$, and a container $C \supseteq I$ which depends only on S .

Throughout the algorithm, we will maintain a partition $V = A \cup S \cup X$, where

- A , the “available” vertices, initially $A = V$
- S , the current fingerprint, initially $S = \emptyset$
- X , the “excluded” vertices, initially $X = \emptyset$.

The *max-degree order* of $G[A]$ is an ordering of A in by the degree of the vertices in $G[A]$, with the largest first, and breaking ties according to some arbitrary predetermined ordering of V .

While $|X| < \delta |V|$:

1. Let v be the first vertex of $I \cap A$ in the max-degree order on $G[A]$.
2. Add v to S .
3. Add the neighbors of v to X .
4. Add vertices preceding v in the max-degree order on $G[A]$ to X .
5. Remove from A all the new vertices added to $S \cup X$.

Claim: when the algorithm terminates, we obtain a partition $V = A \cup S \cup X$ such that $|X| \geq \delta |V|$ and $|S| \leq 2\delta |V| / d$.

Proof idea: due to the degree hypotheses, in every iteration, at least $\geq d/2$ new vertices are added to X (provided that $d \leq 2\delta |V|$). See [Morris' lecture notes](#) for details.

Key facts:

- Two different maximal independent sets $I, I' \subseteq V$ that produce the same fingerprint S in the algorithm necessarily produces the same partition $V = A \cup S \cup X$
- The final set $S \cup A$ contains I (since only vertices not in I are ever moved to I)

Therefore, the total number possibilities for containers $S \cup A$ is at most the number of sets $S \subseteq V$. Since $|S| \leq 2\delta |V| / d$ and $|A \cup S| \leq (1 - \delta) |V|$, this concludes the proof of the graph container lemma.

The fingerprint obtained by the proof actually gives us a stronger consequence that will be important for some applications.

Theorem 11.2.3 (Graph container theorem, with fingerprints)

For every $c > 0$, there exists $\delta > 0$ such that the following holds.

Let $G = (V, E)$ a graph with average degree d and maximum degree at most cd .

Writing \mathcal{I} for the collection of independent sets of G , there exist functions

$$S: \mathcal{I} \rightarrow 2^V \quad \text{and} \quad A: 2^V \rightarrow 2^V$$

(one only needs to define $A(\cdot)$ on sets in the image of S)

such that, for every $I \in \mathcal{I}$,

- $S(I) \subseteq I \subseteq S(I) \cup A(S(I))$
- $|S(I)| \leq 2\delta |V| / d$
- $|S(I) \cup A(S(I))| \leq (1 - \delta) |V|$

11.3 Hypergraph container theorem

An independent set in a hypergraph is a subset of vertices containing no edges.

Given an r -uniform hypergraph H and $1 \leq \ell < r$, we write

$$\Delta_\ell(H) = \max_{A \subseteq V(H): |A|=\ell} \text{the number of edges containing } A$$

Theorem 11.3.1 (Container theorem for 3-uniform hypergraph)

For every $c > 0$ there exists $\delta > 0$ such that the following holds.

Let H be a 3-uniform hypergraph with average degree $d \geq \delta^{-1}$ and

$$\Delta_1(H) \leq cd \quad \text{and} \quad \Delta_2(H) \leq c\sqrt{d}.$$

Then there exists a collection \mathcal{C} of subsets of $V(H)$ with

$$|\mathcal{C}| \leq \binom{v(H)}{\leq v(H)/\sqrt{d}}$$

such that

- Every independent set of H is contained in some $C \in \mathcal{C}$, and
- $|C| \leq (1 - \delta)v(H)$ for every $C \in \mathcal{C}$.

Like the graph container theorem, the hypergraph container theorem is proved by designing an algorithm to produce, from an independent set $I \subseteq V(H)$, a fingerprint $S \subseteq I$ and a container $C \supset I$.

The hypergraph container algorithm is more involved compared to the graph container algorithm. In fact, the 3-uniform hypergraph container algorithm calls the graph container algorithm.

Container algorithm for 3-uniform hypergraphs (a very rough sketch):

Throughout the algorithm, we will maintain

- A fingerprint S , initially $S = \emptyset$
- A 3-uniform hypergraph A , initially $A = H$
- A graph G of “forbidden” pairs on $V(H)$, initially $G = \emptyset$

While $|S| \leq v(H)/\sqrt{d} - 1$:

- Let u be the first vertex in I in the max-degree order on A
- Add u to S
- Add xy to $E(G)$ whenever $uxy \in E(H)$
- Remove from $V(A)$ the vertex u as well as all vertices proceeding u in the max-degree order on A
- Remove from $V(A)$ every vertex whose degree in G is larger than $c\sqrt{d}$.
- Remove from $E(A)$ every edge that contains an edge of G .

Finally, it is will be the case that either

- We have removed many vertices from $V(A)$
- Or the final graph G has at least $\Omega(\sqrt{d}n)$ edges and has maximum degree $O(\sqrt{d})$, so that we can apply the graph container lemma to G .

In either case, the algorithm produces a container with the desired properties. Again see [Morris’ lecture notes](#) for details.