Lecture notes (MIT 18.226, Fall 2020)

Probabilistic Methods in Combinatorics

Yufei Zhao

Massachusetts Institute of Technology

yufeiz@mit.edu

http://yufeizhao.com/pm/

Contents

1 Introduction **7**

1.1 Lower bounds to Ramsey numbers 7

1.1.1 Erdős' original proof 8

1.1.2 Alteration method 9

1.1.3 Lovász local lemma 10

1.2 Set systems 11

1.2.1 Sperner's theorem 11

1.2.2 Bollobás two families theorem 12

1.2.3 Erdős–Ko–Rado theorem on intersecting families 13

1.3 2-colorable hypergraphs 13

1.4 List chromatic number of $K_{n,n}$ 15

2 Linearity of expectations **17**

2.1 Hamiltonian paths in tournaments 17

2.2 Sum-free set 18

2.3 Turán's theorem and independent sets 18

2.4 Crossing number inequality 20

2.4.1 Application to incidence geometry 22

2.5 Dense packing of spheres in high dimensions 23

2.6 Unbalancing lights 26

3 Alterations **28**

3.1 Ramsey numbers 28

3.2 Dominating set in graphs 28

3.3 Heilbronn triangle problem 29

3.4 Markov's inequality 31

3.5 High girth and high chromatic number 31

3.6 Greedy random coloring 32

4 Second moment method **34**

4.1 Threshold functions for small subgraphs in random graphs 36

4.2 Existence of thresholds 42

4.3 Clique number of a random graph 47

4.4 Hardy–Ramanujan theorem on the number of prime divisors 49

4.5 Distinct sums 52

4.6 Weierstrass approximation theorem 54

5 Chernoff bound **56**

5.1 Discrepancy 58

5.2 Hajós conjecture counterexample 60

6 Lovász local lemma **63**

6.1 Statement and proof 63

6.2 Algorithmic local lemma 66

6.3 Coloring hypergraphs 68

6.3.1 Compactness argument 69

6.4 Decomposing coverings 70

6.5 Large independent sets 72

6.6 Directed cycles of length divisible by k 73

6.7 Lopsided local lemma 74

6.7.1 Random permutations and positive dependencies 76

6.7.2 Latin square transversals 77

7 Correlation inequalities **79**

7.1 Harris–FKG inequality 79

7.2 Applications to random graphs 81

7.2.1 Triangle-free probability 81

7.2.2 Maximum degree 82

8 Janson inequalities **85**

8.1 Probability of non-existence . . . 85
8.2 Lower tails . . . 91
8.3 Clique and chromatic number of $G(n, 1/2)$ . . . 94

9 Concentration of measure **99**
9.1 Martingales concentration inequalities . . . 101
9.2 Chromatic number of random graphs . . . 105
9.2.1 Concentration of chromatic number . . . 105
9.2.2 Clique number, again . . . 106
9.2.3 Chromatic number of sparse random graphs . . . 108
9.3 Isoperimetric inequalities: a geometric perspective . . . 110
9.3.1 The sphere and Gauss space . . . 114
9.3.2 Johnson–Lindenstrauss Lemma . . . 117
9.4 Talagrand inequality . . . 119
9.4.1 Convex Lipschitz functions of independent random variables . . . 119
9.4.2 Convex distance . . . 123
9.4.3 How to apply Talagrand's inequality . . . 125
9.4.4 Largest eigenvalue of a random matrix . . . 126
9.4.5 Certifiable functions and longest increasing subsequence . . . 127

10 Entropy method **130**
10.1 Basic properties . . . 130
10.2 Upper bound on the permanent and the number of perfect matchings . . . 135
10.2.1 The maximum number of Hamilton paths in a tournament . . . 136
10.3 Sidorenko's inequality . . . 137
10.4 Shearer's lemma . . . 142
10.4.1 Triangle-intersecting families . . . 144
10.4.2 The number of independent sets in a regular bipartite graph . . . 145

11 The container method **150**

11.1 Containers for triangle-free graphs 152
11.1.1 The number of triangle-free graphs 152
11.1.2 Mantel's theorem in random graphs 153
11.2 Graph containers 154
11.3 Hypergraph container theorem 156

These notes were created primarily for my own lecture preparation. The writing style is far below that of formal writing and publications (in terms of complete sentences, abbreviations, citations, etc.). The notes are not meant to be a replacement of the lectures.

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 or by emailing me at yufeiz@mit.edu.

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 $C > 0$)
- $f/g \to 0$, $f \ll g$, $f = o(g)$ (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)

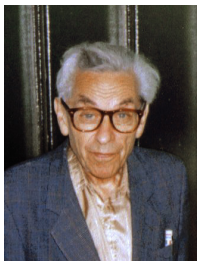


Figure 1: Paul Erdős (1913–1996) is considered the father of the probabilistic method. He published around 1,500 papers during his lifetime, and had more than 500 collaborators. To learn more about Erdős, see his biography *The man who loved only numbers* by Hoffman and the documentary *N is a number*.

1 Introduction

Probabilistic method: to prove that an object exists, show that a random construction works with positive probability

Tackle combinatorics problems by introducing randomness

Theorem 1.0.1. Every graph $G = (V, E)$ contains a bipartite subgraph with at least $|E|/2$ edges.

Proof. Randomly color every vertex of G with black or white, iid uniform

Let $E' =$ edges with one end black and one end white

Then (V, E') is a bipartite subgraph of G

Every edge belongs to E' with probability $\frac{1}{2}$, so by linearity of expectation, $\mathbb{E}[|E'|] = \frac{1}{2}|E|$.

Thus there is some coloring with $|E'| \geq \frac{1}{2}|E|$, giving the desired bipartite subgraph. □

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.

e.g., $R(3, 3) = 6$

Ramsey (1929) proved that $R(k, \ell)$ exists and is finite

Figure 2: 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.

1.1.1 Erdős' original proof

The probabilistic method started with:
P. Erdős, Some remarks on the theory of graphs, BAMS, 1947

Remark 1.1.1 (Hungarian names). Typing "Erdős" in LaTeX: `Erd\H{o}s` and *not* `Erd\"os`

Hungarian pronunciations: s = /sh/ and sz = /s/, e.g., Erdős, Szekeres, Lovász

Theorem 1.1.2 (Erdős 1947)**.** If $\binom{n}{k}2^{1-\binom{k}{2}} < 1$, then $R(k,k) > n$. In other words, there exist a red-blue edge-coloring of K_n without a monochromatic K_k.

Proof. Color edges uniformly at random

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

$$\mathbb{P}(\text{there exists a monochromatic } K_k) = \mathbb{P}\left(\bigcup_{R\in\binom{[n]}{k}} A_R\right) \le \sum_{R\in\binom{[n]}{k}} \mathbb{P}(A_R) = \binom{n}{k}2^{1-\binom{k}{2}} < 1.$$

Thus, with positive probability, the random coloring gives no monochromatic K_k. □

Remark 1.1.3. By optimizing n (using Stirling's formula) above, we obtain

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

Can be alternatively phrased as counting: of all $2^{\binom{n}{2}}$ possible colorings, not all are bad (this was how the argument was phrased in the original Erdős 1947 paper.

In this course, we almost always only consider finite probability spaces. While in principle the finite probability arguments can be rephrased as counting, but some of the later more

involved arguments are impractical without a probabilistic perspective.

Constructive lower bounds? Algorithmic? Open! "Finding hay in a haystack"

Remark 1.1.4 (Ramsey number upper bounds). Erdős–Szekeres (1935):

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

Recent improvements by Conlon (2009), and most recently Sah (2020+):

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

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

1.1.2 Alteration method

Two steps: (1) randomly color (2) get rid of bad parts

Theorem 1.1.5. For any k, n, we have $R(k,k) > n - \binom{n}{k}2^{1-\binom{k}{2}}$.

Proof. Construct in two steps:
(1) Randomly 2-color the edges of K_n
(2) Delete a vertex from every monochromatic K_k
Final graph has no monochromatic K_k

After step (1), every fixed K_k is monochromatic with probability $2^{1-\binom{k}{2}}$, let X be the number of monochromatic K_k's. $\mathbb{E}X = \binom{n}{k}2^{1-\binom{k}{2}}$.

We delete at most $|X|$ vertices in step (2). Thus final graph has size $\ge n - |X|$, which has expectation $n - \binom{n}{k}2^{1-\binom{k}{2}}$.

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

Remark 1.1.6. By optimizing the choice of n in the theorem, we obtain

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

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

1.1.3 Lovász local lemma

We give one more improvement to the lower bound, using the Lovász local lemma, which we will prove later in the course

Consider "bad events" $E_1, \ldots, E_n$. We want to avoid all.

If all $\mathbb{P}(E_i)$ small, say $\sum_i \mathbb{P}(E_i) < 1$, then can avoid all bad events.

Or, if they are all 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 there are some weak dependencies?

Theorem 1.1.7 (Lovász local lemma)**.** Let $E_1, \ldots, E_n$ be events, with $\mathbb{P}[E_i] \leq p$ for all i. Suppose that each E_i is independent of all other E_j except for at most d of them. If

$$ep(d+1) < 1,$$

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

Remark 1.1.8. The meaning of "independent of ..." is actually somewhat subtle (and easily mistaken). We will come back to this issue later on when we discuss the local lemma in more detail.

Theorem 1.1.9 (Spencer 1977)**.** If $e\left(\binom{k}{2}\binom{n}{k-2} + 1\right) 2^{1-\binom{k}{2}} < 1$, then $R(k,k) > n$.

Proof. Random 2-color edges of K_n

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

E_R is independent of all E_S other than those such that $|R \cap S| \geq 2$

For each R, there are at most $\binom{k}{2}\binom{n}{k-2}$ choices S with $|S| = k$ and $|R \cap S| \geq 2$.

Apply Lovász local lemma to the events $\left\{E_R : R \in \binom{V}{k}\right\}$ and $p = 2^{1-\binom{k}{2}}$ and $d = \binom{k}{2}\binom{n}{k-2}$, we get that with positive probability none of the events E_R occur, which gives a coloring with no monochromatic K_k's. □

Remark 1.1.10. 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

1.2.1 Sperner's theorem

Let $\mathcal{F}$ a collection of subsets of $\{1,2,\ldots,n\}$. We say that $\mathcal{F}$ is an **antichain** if no set in $\mathcal{F}$ is contained in another set in $\mathcal{F}$.

Question 1.2.1. What is the maximum number of sets in an antichain?

Example: $\mathcal{F} = \binom{[n]}{k}$ has size $\binom{n}{k}$. Maximized when $k = \lfloor \frac{n}{2} \rfloor$ or $\lceil \frac{n}{2} \rceil$. The next result shows that we cannot do better.

Theorem 1.2.2 (Sperner 1928)**.** If $\mathcal{F}$ is an antichain of subsets of $\{1,2,\ldots,n\}$, 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

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

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

Proof. Consider a random permutation σ of $\{1,2,\ldots,n\}$, and its associated chain of subsets

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

where the last set is always equal to $\{1,2,\ldots,n\}$. For each $A \subset \{1,2,\ldots,n\}$, let E_A denote the event that A is found in this chain. Then

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

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 event, and thus their probabilities sum to at most 1. $\square$

1.2.2 Bollobás two families theorem

Sperner's theorem is generalized by the following celebrated result of Bollobás, which has many more generalizations that we will not discuss here.

Theorem 1.2.4 (Bollobás (1965) "two families theorem"). 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 $m \leq \binom{r+s}{r}$.

Remark 1.2.5. The 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 $\tau(\mathcal{F} \setminus \{S\}) < \tau(\mathcal{F})$ for all $S \in \mathcal{F}$.

Question 1.2.6. 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-critcal r-uniform with $\tau(\mathcal{F}) = s+1$. (why?)

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

Theorem 1.2.7. 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. Consider a uniform random ordering of all elements.

Let X_i be the event that all elements of A_i come before B_i.

Then $\mathbb{P}[X_i] = \binom{|A_i|+|B_i|}{|A_i|}^{-1}$ (all permutations of $A_i \cup B_i$ are equally likely to occur).

Note that the events X_i are disjoint (X_i and X_j both occuring would contradict the hypothesis for A_i, B_i, A_j, B_j). Thus $\sum_i \mathbb{P}[X_i] \leq 1$. □

1.2.3 Erdős–Ko–Rado theorem on intersecting families

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

Question 1.2.8. 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 $1, 2, \ldots, n$ (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 contingous 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 $|\mathcal{F}| \leq \frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1}$. □

1.3 2-colorable hypergraphs

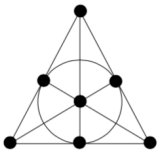
An **k-uniform hypergraph** (or **k-graph**) is a pair $H = (V, E)$, where V (vertices) is a finite set and E (edges) is a set of k-element subsets of E, i.e., $E \subseteq \binom{V}{k}$ (so hypergraphs are really the same concept as set families).

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)

$m(2) = 3$

$m(3) = 7$. Example: Fano plane (below) is not 2-colorable (the other direction is by exhaustive search)



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

Exact value of $m(k)$ is unknown for all $k \geq 5$

The probabilistic method gives a short proof of a lower bound (random coloring):

Theorem 1.3.1 (Erdős 1964)**.** For any $k \geq 2$, $m(k) \geq 2^{k-1}$, i.e., 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 $\leq 2^{-k+1}m < 1$. □

Remark 1.3.2. Later on we will prove an 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)**.** $m(k) = O(k^2 2^k)$, i.e., there exists a k-uniform hypergraph with $O(k^2 2^k)$ edges that is not 2-colorable.

Proof. Fix $|V| = n$ to be decided. Let H be the k-uniform hypergraph obtained by choosing m random edges (with replacement) $S_1, \ldots, S_m$.

Given a coloring $\chi\colon V \to [2]$, let A_χ denote the event that χ is a proper coloring (i.e., no monochromatic edges). It suffices to check that $\sum_\chi \mathbb{P}[A_\chi] < 1$.

If χ colors a vertices with one color and b vertices with the other color, then the probability that (random) S_1 is monochromatic under (fixed) χ is

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

Setting $n=k^2$, we see that the above quantity is at least $c2^{-k}$ for some constant $c>0$.

Thus, the probability that χ is a proper coloring (i.e., no monochromatic edges) is at most $(1-c2^{-k})^m \leq e^{-c2^{-k}m}$ (using $1+x \leq e^x$ for all real x).

Thus, $\sum_\chi \mathbb{P}[A_\chi] \leq 2^n e^{-c2^{-k}m} < 1$ for some $m = O(k^2 2^k)$ (recall $n=k^2$). □

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

It should be clear that $\chi(G) \leq \mathrm{ch}(G)$, but the inequality may be strict.

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!):

$$\begin{matrix}\{2,3\} & \{2,3\}\\ \{1,3\} & \{1,3\}\\ \{1,2\} & \{1,2\}\end{matrix}$$

Easy to check then that $\mathrm{ch}(K_{3,3}) = 3$.

Question 1.4.1. What is the asymptotic behavior of $\mathrm{ch}(K_{n,n})$?

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

Theorem 1.4.2. If $n < 2^{k-1}$, then $K_{n,n}$ is k-choosable.

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

Proof. For each color, mark it either "L" or "R" iid uniformly.

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 $2n2^{-k} < 1$. So with positive probability, every vertex has some color remaining. Assign the colors arbitrarily for a valid coloring. □

The lower bound on $\mathrm{ch}(K_{n,n})$ 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 k-uniform hypergraph $|E| = n$ edges. Label the vertex of $K_{n,n}$ by v_e and w_e as e ranges over E. View V as colors and assign to both v_e and w_e a list of colors given by the k-element set e.

If this $K_{n,n}$ has a proper list coloring with the assigned colors. Let C be the colors used among the n vertices. Then we get a proper 2-coloring of H by setting C black and $V \setminus C$ white. So if H is not 2-colorable, then this $K_{n,n}$ is not k-choosable. □

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

Putting these bounds together:

Corollary 1.4.4. $\mathrm{ch}(K_{n,n}) = (1 + o(1)) \log_2 n$

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

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

Let $X = c_1X_1 + \cdots + c_nX_n$ where $X_1, \ldots, X_n$ are random variables, and $c_1, \ldots, c_n$ constants. Then

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

Note: 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 question with a simple solution (there are also much more involved solutions via enumerations, but linearity of expectations nearly trivializes the problem).

Question 2.0.1. What is the average number of fixed points of a random permutation of $[n]$ chosen uniformly at random?

Let X_i be the event that i is fixed. Then $\mathbb{E}[X_i] = 1/n$. So the expected number of fixed points is $\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = 1$

2.1 Hamiltonian paths in tournaments

Important observation for proving existence: With positive probability, $X \geq \mathbb{E}[X]$ (likewise for $X \leq \mathbb{E}[X]$)

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

Theorem 2.1.1 (Szele 1943)**.** There is a tournament on n vertices with at least $n!2^{-(n-1)}$ Hamiltonian paths

Proof. Let X be the number of Hamiltonian paths in a random tournament.

For every permutation σ of $[n]$, one has the directed path $\sigma(1) \to \sigma(2) \to \cdots \to \sigma(n)$ with probability 2^{-n+1}.

Let X be the number of σ satisfying the above. $\mathbb{E}X = n!2^{-n+1}$. □

This was considered the first use of the probabilistic method. Szele conjectured that the maximum number of Hamiltonian 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 the course when discussing the entropy method).

2.2 Sum-free set

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 (Erdős 1965)**.** Every set of n nonzero integers contains a sum-free subset of size $\geq n/3$.

Proof. Let $A \subset \mathbb{Z} \setminus \{0\}$ 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, $\mathbb{E}|A_\theta| = n/3$. □

Remark 2.2.2. Alon and Kleitman (1990) noted that one can improve the bound to $\geq (n+1)/3$ by noting that $|A_\theta| = 0$ for $\theta \approx 0$.

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

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

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

2.3 Turán's theorem and independent sets

Question 2.3.1. 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. □

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

Remark 2.3.4 (Derandomization). Here is an alternative "greedy algorithm" proof of the Caro–Wei inequality.

Permute the vertices in non-increasing order of their degree.

And then greedily construct an independent set: at each step, take the first available vertex (in this order) and then discarding 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.

Taking the complement

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).

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 1941) Every n-vertex K_{r+1}-free graph has $\leq \left(1 - \frac{1}{r}\right) \frac{n^2}{2}$ edges.

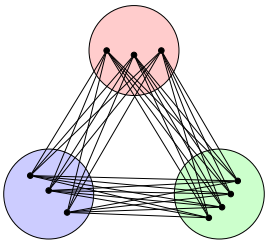


Figure 3: The Turán graph $T_{10,3}$.

Proof. Since G is K_{r+1}-free, by Corollary 2.3.5, letting $\overline{d}$ be average degree and $m = n\overline{d}/2$ be the number of edges, we see that the size $\omega(G)$ of the largest clique of G satisfies

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

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.

2.4 Crossing number inequality

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

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

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

$K_{3,3}$ and K_5 are non-planar; furthermore, the following famous 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

(Also related: Wagner's theorem (1937) says that 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)

Question 2.4.1. 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 2.4.2 (Crossing number inequality)**.** In a graph $G = (V, E)$, if $|E| \geq 4|V|$, then

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

Corollary 2.4.3. In a graph $G = (V, E)$, if $|E| \gtrsim |V|^2$, then $\operatorname{cr}(G) \gtrsim |V|^4$.

Proof. Recall **Euler's formula:** $v - e + f = 2$ for every connected planar graph

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, $|E| \leq 3|V| - 6$.

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

By deleting one edge for each crossing, we get a planar graph, so $|E| - \operatorname{cr}(G) > 3|V|$, i.e.,

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

This is a "cheap bound" that we will boost using the probabilistic method.

For graphs with $|E| = \Theta(n^2)$, this gives $\operatorname{cr}(G) \gtrsim n^2$. This 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

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

So

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

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

$$p^4\,\mathrm{cr}(G) \ge p^2|E| - 3p|V|.$$

Thus

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

Setting $p \in [0,1]$ so that $4p^{-3}|V| = p^{-2}|E|$, we obtain $\mathrm{cr}(G) \gtrsim |E|^3 / |V|^2$. □

2.4.1 Application to incidence geometry

Question 2.4.4. What is the maximum number of incidences between n distinct points and n distinct lines on a plane?

Let $\mathcal{P}$ be a set of points and $\mathcal{L}$ a set of lines. Denote the number of incidences by

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

Example: n points and n lines:

$$\mathcal{P} = [k] \times [2k^2] \quad \text{and} \quad \mathcal{L} = \{y = mx + b : m \in [k], b \in [k^2]\}$$

Every line contains k points from $\mathcal{P}$. Taking $3k^3 \approx n$ gives $k^4 = \Theta(n^{4/3})$ incidences.

Can we do better?

No. The following foundational theorem in incidence geometry implies that one has $O(n^{4/3})$ incidences between n points and n lines.

Theorem 2.4.5 (Szemerédi–Trotter 1983)**.** Given a set $\mathcal{P}$ of points and $\mathcal{L}$ of lines in $\mathbb{R}^2$,

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

We will show how to prove the Szemerédi–Trotter theorem using the crossing number inequality. This proof is due to Székely (1997).

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

Using that every pair of points determine at most one line, and counting triples $(p, p', \ell) \in \mathcal{P} \times \mathcal{P} \times \mathcal{L}$ with $p \ne p'$ and $p, p' \in \ell$, this is $\le |\mathcal{P}|^2$ and

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

Combining we get

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

By point-line duality, also

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

This gives $n^{3/2}$ for n points and n lines. Can we do better? Note that this is tight for planes over finite fields. Need to use topology of Euclidean space.

Proof of Szemerédi–Trotter theorem. Assume that there are no lines with < 2 incidences (otherwise remove such lines repeatedly until this is the same; we remove $\leq |\mathcal{L}|$ incidences this way).

Draw a graph based on incidences. Vertices are point in $\mathcal{P}$ and edges join consecutive points of $\mathcal{P}$ on a given line of $\mathcal{L}$.

A line with k incidences gives $k-1 \geq k/2$ edges, so the total number of edges is $\leq |I(\mathcal{P},\mathcal{L})|/2$.

There are at most $|\mathcal{L}|^2$ crossings. So by crossing number inequality

$$|\mathcal{L}|^2 \geq \mathrm{cr}(G) \gtrsim \frac{|E|^3}{|V|^2} \gtrsim \frac{|I(\mathcal{P},\mathcal{L})|^3}{|\mathcal{P}|^2} \quad \text{if } |I(\mathcal{P},\mathcal{L})| \geq 8|\mathcal{P}|.$$

So $I(\mathcal{P},\mathcal{L}) \lesssim |\mathcal{P}|^{2/3}|\mathcal{L}|^{2/3} + |\mathcal{P}|$. Remember to add $|\mathcal{L}|$ to the bound from the first step of the proof (removing lines with < 2 incidences). □

2.5 Dense packing of spheres in high dimensions

Question 2.5.1. 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 $n \to \infty$)

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

Exact maximum only solved in dimension $1, 2, 3, 8, 24$. Dimensions 8 and 24 were only solved recently (see this Quanta magazine story). 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 $\mathrm{vol}(B(1/2)) = \frac{\pi^{n/2}}{(n/2)!2^n} < n^{-cn}$.

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

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.

How does one pick a random lattice?

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

So every matrix in $\mathrm{SL}_n(\mathbb{R})$ 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 $A, A' \in \mathrm{SL}_n(\mathbb{R})$ represent the same lattice iff $A' = AU$ for some $U \in \mathrm{SL}_n(\mathbb{Z})$.

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

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 2.5.2 (Siegel mean value theorem)**.** Let L be the random lattice in $\mathbb{R}^n$ as above and $S \subset \mathbb{R}^n$. 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 choosen 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 2.5.3 (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 $1-\epsilon$ for arbitrarily small $\epsilon > 0$ if you like) centered at the origin. By the Siegel mean value theorem, the random lattice is 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, a bunch of 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.

Skipping details, one can extend the earlier idea to choose a random unimodular lattice in 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 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 2.5.4 (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 $\gamma = 0.577\ldots$ is Euler's constant.

Open problem 2.5.5. Do there exist lattices (or sphere packings) in $\mathbb{R}^n$ with density $\geq (c + o(1))^n$ for some constant $c > 1/2$?

2.6 Unbalancing lights

Theorem 2.6.1. Let $a_{ij} = \pm 1$ for all $i, j \in [n]$. There exists $x_i, y_j \in \{-1, 1\}$ for all $i, j \in [n]$ such that

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

Interpretation: $n \times n$ array of lights. Can flip rows and columns. Want to turn on as many lights as possible.

Proof. Choose $y_1, \ldots, y_n$ randomly. And then choose x_i to make the i-th row sum nonnnegative. Let

$$R_i = \sum_{j=1}^{n} a_{ij} y_j \qquad \text{and} \qquad R = \sum_{i=1}^{n} |R_i|.$$

How is R_i distributed? Same distribution as $S_n = \epsilon_1 + \cdots + \epsilon_n$, a sum of n i.i.d. uniform $\{-1, 1\}$. And so for every i

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

e.g., by central limit theorem

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

(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

$$\mathbb{E}[R] = \left(\sqrt{\frac{2}{\pi}} + o(1)\right) n^{3/2}.$$

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

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 2.6.2. Let $V = V_1 \cup \cdots \cup V_k$, where $V_1, \ldots, V_k$ 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 $V_1, \ldots, V_k$ is colored blue. Then there exists $S \subset V$ 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. Let's do this proof for $k = 3$. Proof easily generalizes to other 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

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

equals to some polynomial

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

(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 $|a_{i,j,k}| \le n^3$. We are done after the following lemma

Lemma 2.6.3. Let P_k denote the set of polynomials $g(p_1, \ldots, p_k)$ 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 $p_1, \ldots, p_k \in [0, 1]$ with $|g(p_1, \ldots, p_k)| \ge c$.

Proof of Lemma. Set $M(g) = \sup_{p_1, \ldots, p_k \in [0,1]} |g(p_1, \ldots, p_k)|$ (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 $M \colon P_k \to \mathbb{R}$ is continuous, M attains a minimum value $c = M(g) > 0$ for some $g \in P_k$. ■ □

3 Alterations

3.1 Ramsey numbers

Recall from Section 1.1:

$R(s,t)$ = smallest n such that every red/blue edge coloring of K_n contains a red K_s or a blue K_t

Using the basic method (union bounds), we deduce

Theorem 3.1.1. If there exists $p \in [0,1]$ with

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

then $R(s,t) > n$.

Proof sketch. Color edge red with prob p and blue with prob $1-p$. LHS upper bounds the probability of a red K_s or a blue K_t. □

Using the alteration method, we deduce

Theorem 3.1.2. For all $p \in [0,1]$ and n,

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

Proof sketch. Color edge red with prob p and blue with prob $1-p$ remove one vertex from each red K_s or blue K_t. RHS lower bounds the expected number remaining vertices. □

3.2 Dominating set in graphs

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

Theorem 3.2.1. Every graph on n vertices with minimum degree $\delta > 1$ has a dominating set of size at most $\left(\frac{\log(\delta+1)+1}{\delta+1}\right) n$.

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

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

3.3 Heilbronn triangle problem

Question 3.3.1. How can one place n points in the unit square so that no three points forms a triangle with small area?

Let

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

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 consectutive 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 3.3.2. The currently best upper bound known is $\Delta(n) \leq n^{-8/7+o(1)}$ (Komlós, Pintz, and Szemerédi 1981)

Theorem 3.3.3. 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

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

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

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

Thus, with p, q, r at random

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

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

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

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

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

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

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 $[p]^2 \subset \mathbb{Z}^2$. 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 $2n$.)

3.4 Markov's inequality

We note an important tool that will be used next.

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. $\mathbb{E}[X] \geq \mathbb{E}[X 1_{X \geq a}] \geq \mathbb{E}[a 1_{X \geq a}] = a\mathbb{P}(X \geq a)$ □

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

3.5 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 3.5.1 (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 3.5.2. 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 $\epsilon = \epsilon_k$ such that for all sufficiently large n there exists an n-vertex graph with chromatic number $> k$ but every subgraph on $\lfloor \epsilon n \rfloor$ vertices is 3-colorable. (In fact, one can take $G \sim G(n, C/n)$; see "Probabilistic Lens: Local coloring" in Alon–Spencer)

3.6 Greedy random coloring

Recall $m(k)$ is the minimum number of edges in a k-uniform hypergraph that is not 2-colorable.

Earlier we proved that $m(k) \geq 2^{k-1}$. Indeed, given a k-graph with $< 2^{k-1}$ edges, by randomly coloring the vertices, the expected number of monochromatic numbers is < 1.

We also proved an upper bound $m(k) = O(k^2 2^k)$ by taking a random k-uniform hypergraph on k^2 vertices.

Here is the currently best known lower bound.

Theorem 3.6.1 (Radhakrishnan and Srinivasan (2000)). $m(k) \gtrsim \sqrt{\frac{k}{\log k}} 2^k$

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. Suppose H is a k-graph with m edges.

Map $V(H) \to [0,1]$ uniformly at random.

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 been colored blue).

The resulting coloring has no all-blue edges. What is the probability of seeing a red edge?

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**.

Want to bound probability of seeing a conflicting pair in a random $V(H) \to [0,1]$.

Here is an attempt (an earlier weaker result due to Pluhaár (2009)). Each pair of edges with exactly one vertex in common conflicts with probability $\frac{(k-1)!^2}{(2k-1)!} = \frac{1}{2k-1}\binom{2k-2}{k-1}^{-1} \asymp k^{-1/2}2^{-2k}$; union bounding over $< m^2$ pairs of edges, the probability of of getting a conflicting edge is $\lesssim m^2 k^{-1/2}2^{-2k}$, which is < 1 for some $m \asymp k^{1/4}2^k$.

We'd like to do better by more carefully analyzing conflicting edges. Continuing ...

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

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

The probability that a given edge lands entirely in L is $(\frac{1-p}{2})^k$, and likewise with R

So probability that some edge of H is entirely contained in L or contained in R is $\leq 2m(\frac{1-p}{2})^k$.

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 $x(1-x) \leq 1/4$)

$$\int_{(1-p)/2}^{(1+p)/2} x^{k-1}(1-x)^{k-1}\,dx \leq p4^{-k+1}.$$

Thus the probability of seeing any conflicting pair is

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

Set $p = \log(2^{-k+2}k/m)/k$, we find that the above probability is < 1 for $m = c2^k\sqrt{k/\log k}$, with $c > 0$ being a sufficiently small constant. □

4 Second moment method

Previously, we used $\mathbb{E}X \geq a$ to deduce $\mathbb{P}(X \geq a) > 0$. We also saw from Markov's inequality that for $X \geq 0$, if $\mathbb{E}X$ is very small, then X is small with high probability.

Does $\mathbb{E}X$ being (very) large imply that X is large with high probability?

No! X could be almost always small but $\mathbb{E}X$ could still be large due to outliers (rare large values of X).

Often we want 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. We begin with the simplest method. It is the easiest to execute, requires the least hypotheses, but only produces weak (though often useful) concentration bounds.

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

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

Notation convention: mean μ, variance σ^2, standard deviation σ.

Theorem 4.0.1 (Chebyshev's inequality)**.** Let X be a random variable with mean μ and standard deviation σ. For any $\lambda > 0$

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

Proof. By Markov's inequality,

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

Remark 4.0.2. Concentration bounds that show small probability of deviating from the mean are called **tail bounds** (also: upper tail bounds for bounding $\mathbb{P}(X \geq \mu + a)$ and lower tail bounds for bounding $\mathbb{P}(X \leq \mu - a)$). Chebyshev's inequality gives tail bounds with polynomial decay. Later on we will see tools that give much better decay (usually exponential) provided additional assumptions on the random variable (e.g., independence).

We can rewrite Chebyshev's inequality as

$$\mathbb{P}(|X - \mathbb{E}X| \geq \epsilon\mathbb{E}X) \leq \frac{\mathrm{Var}\, X}{\epsilon^2(\mathbb{E}X)^2}.$$

Corollary 4.0.3. If $\mathrm{Var}[X] = o(\mathbb{E}X)^2$ then $X \sim \mathbb{E}X$ whp.

Remark 4.0.4. We are invoking asymptotics here (so we are actually considering a sequence X_n of random variables instead of a single one). The conclusion is equivalent to that for every $\epsilon > 0$, one has $|X - \mathbb{E}X| \leq \epsilon \mathbb{E}X$ with probability $1 - o(1)$ as $n \to \infty$.

Variance can be calculated from pairwise covariances. Recall the **covariance**

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

So $\mathrm{Var}[X] = \mathrm{Cov}[X, X]$. Covariance is bilinear in X and Y, i.e., for constants $a_1, \ldots$ and $b_1, \ldots,$ one has

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

Thus, given $X = X_1 + \cdots + X_n$ (no assumptions on dependencies between the X_i's), we have

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

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

Example 4.0.5 (Sum of independent Bernoulli)**.** Suppose $X = X_1 + \cdots + X_n$ with X_i iid $X_i \sim \mathrm{Bernoulli}(p)$, i.e., $X = 1$ with prob p and $X = 0$ with prob $1 - p$.

Then $\mu = np$ and $\sigma^2 = np(1 - p)$. If $np \gg 1$ then $\sigma \ll \mu$ 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.

Example 4.0.6 (The number of triangles in a random graph)**.** Let

$$X = \text{the number of triangles in the random graph } G(n, p).$$

For vertices $i, j, k \in [n]$, denote the edge indicator variables by $X_{ij} = 1_{ij \text{ is an edge}}$. Let the triangle indicator variables be $X_{ijk} = 1_{ijk \text{ is a triangle}} = X_{ij} X_{ik} X_{jk}$. Then

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

Its expectation is easy to compute, since $\mathbb{E}[X_{ij}X_{ik}X_{jk}] = \mathbb{E}[X_{ij}]\mathbb{E}[X_{ik}]\mathbb{E}[X_{jk}] = p^3$ by independence. So

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

Now we compute $\operatorname{Var} X$. Unlike in the earlier example, the summands of X are not all independent. Nonetheless, it is easy to compute the variance.

Given two triples T_1, T_2 of vertices

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

Thus

$$\operatorname{Var} X = \sum_{T_1,T_2} \operatorname{Cov}[X_{T_1}, X_{T_2}] = \binom{n}{3}(p^3 - p^6) + \binom{n}{2}n(n-1)(p^5 - p^6) \lesssim n^3p^3 + n^4p^5$$

When do we have $\sigma \ll \mu$? It is equivalent to satisfying both $n^{3/2}p^{3/2} \ll n^3p^3$ (which gives $p \gg 1/n$) and $n^2p^{5/2} \ll n^3p^3$ (which gives $p \gg n^{-2}$). So $\sigma \ll \mu$ if and only if $p \gg 1/n$, and as we saw earlier, in this case $X \sim \mathbb{E}X$ with high probability.

Remark 4.0.7. Later on we will use more powerful tools (including martingale methods/Azuma-Hoeffding inequalities, and also Janson inequalities) to prove better tail bounds on triangle (and other subgraph) counts.

Remark 4.0.8. Actually the number X of triangles in $G(n,p)$ satisfies an asymptotic central limit theorem, i.e., $(X-\mu)/\sigma \to N(0,1)$ in distribution (Rucinski 1988), initially proved via moment of moments (by showing that higher moments of $(X-\mu)/\sigma$ match those of the normal distribution). Later a different proof was found using the "method of projections."

On the other hand, for much sparser random graphs, when $p \lesssim 1/n$, X is asymptotically Poisson.

4.1 Threshold functions for small subgraphs in random graphs

Question 4.1.1. For which $p = p_n$ is $K_4 \subset G(n,p)$ true with high probability (i.e., with probability $1 - o(1)$)?

There are two statements that one wants to show:

- (0-statement) if $p = p_n$ is small, then $\mathbb{P}(K_4 \subset G(n,p)) \to 0$ as $n \to \infty$.
- (1-statement) if $p = p_n$ is large, then $\mathbb{P}(K_4 \subset G(n,p)) \to 1$ as $n \to \infty$.

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

- To show the 0-statement, it suffices to have $\mathbb{E}X \to 0$, in which case Markov's inequality implies that $\mathbb{P}(X \geq 1) \leq \mathbb{E}X \to 0$ (here we are only using the first moment method).
- To show the 1-statement, it suffices to show $\operatorname{Var} X = o((\mathbb{E}X)^2)$, by the lemma below (second moment method).

For simple applications, e.g., $K_4 \subset G(n,p)$, these two methods turn out to be sufficient. Other applications may require stronger techniques (though sometimes "only" second moment, but much more difficult applications).

Lemma 4.1.2. For any random variable X,

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

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

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

Corollary 4.1.3. If $\operatorname{Var} X = o((\mathbb{E}X)^2)$, then $X > 0$ with probability $1 - o(1)$.

Remark 4.1.4. Here is a slightly stronger inequality in the case of nonnegative random variables. It is a special case of the Paley–Zygmund inequality. I am showing it here because it is neat. It makes no difference for our applications whether we use the next lemma or the previous one.

Lemma 4.1.5. For any random variable $X \geq 0$,

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

Proof. We have $\mathbb{P}(X > 0) = \mathbb{E}[1_{X>0}]$. By the Cauchy–Schwarz inequality

$$\mathbb{E}[1_{X>0}]\ \mathbb{E}[X^2] \geq (\mathbb{E}[1_{X>0}X])^2 = (\mathbb{E}X)^2. \qquad \square$$

Definition 4.1.6 (Graph properties)**.** A **graph property** $\mathcal{P}$ is a subset of all graphs. We say that $\mathcal{P}$ is **monotone (increasing)** if whenever $G \in \mathcal{P}$, then any graph obtained by adding edges to G also satisfies $\mathcal{P}$. We say that $\mathcal{P}$ is **non-trivial** if for all sufficiently large n, there exists an n-vertex graph in $\mathcal{P}$ and an n-vertex graph not in $\mathcal{P}$.

Example 4.1.7. Examples of graph properties

- Contains K_4; i.e., $\mathcal{P} = \{G : K_4 \subset G\}$
- Connected
- Hamiltonian
- 3-colorable (a monotone decreasing property)
- Planar (monotone decreasing)
- Contains a vertex of degree 1 (not monotone increasing or decreasing)

Definition 4.1.8 (Threshold function)**.** We say that r_n is a **threshold function** for some graph property $\mathcal{P}$ if

$$\mathbb{P}(G(n, p_n) \text{ satisfies } \mathcal{P}) \to \begin{cases} 0 & \text{if } p_n/r_n \to 0, \\ 1 & \text{if } p_n/r_n \to \infty. \end{cases}$$

Remark 4.1.9. The above definition is most suitable for monotone increasing properties. For other types of properties one may need to adjust the definition appropriately.

Remark 4.1.10. From the definition, we see that if r_n and r'_n are both threshold functions, then they must be within a constant factor of each other. So it is fine to say "*the* threshold" of some property, with the understanding that we do not care about constant factors. Later on we will see that every monotone property *has* a threshold function.

Theorem 4.1.11. A threshold function for containing a K_3 is $1/n$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}(K_3 \subset G(n, p_n)) = \begin{cases} 0 & \text{if } p_n n \to 0 \\ 1 & \text{if } p_n n \to \infty \end{cases}$$

Proof. Let X be the number of triangles in $G(n, p)$. Then $\mu := \mathbb{E}X = \binom{n}{3} p^3 \sim n^3 p^3/6$. Let $\sigma^2 = \operatorname{Var} X$.

If $p \ll 1/n$, then $\mu = o(1)$, so $\mathbb{P}(X \geq 1) = o(1)$ by Markov, and hence $X = 0$ w.h.p.

If $p \gg 1/n$, then $\mu \to \infty$, and we saw earlier that $\sigma \ll \mu$, so whp $X \sim \mu$ and thus $X > 0$ whp. □

Question 4.1.12. What is the threshold for containing a fixed H as a subgraph?

The next calculation is similar in spirit to what we did earlier for triangles, but we would like to be more organized as there may be more interacting terms in the variance calculation.

General setup. 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. (For variance calculation, we are only considering pairwise dependence. Warning: later on when we study the Lovász Local Lemma, we will need a strong notion of a dependency graph.)

If $i \neq j$ and $i \not\sim j$ then $\mathrm{Cov}[X_i, X_j] = 0$. Otherwise,

$$\mathrm{Cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] \leq \mathbb{E}[X_i X_j] = \mathbb{P}[A_i \wedge A_j].$$

Thus

$$\mathrm{Var}\, X = \sum_{i,j} \mathrm{Cov}[X_i, X_j] \leq \mathbb{E}X + \Delta$$

where

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

The earlier second moment results (Corollary 4.0.3) imply that

$$\text{If } \mathbb{E}X \to \infty \text{ and } \Delta = o(\mathbb{E}X)^2 \text{ then and } X \sim \mathbb{E}X \text{ and } X > 0 \text{ whp.}$$

We have

$$\sum_{(i,j):i\sim j} \mathbb{P}(A_i \wedge A_j) = \sum_i \mathbb{P}(A_i) \sum_{j:j\sim i} \mathbb{P}\left(A_j \mid A_i\right)$$

In many symmetric situations (e.g. our examples), the following quantity does not depend on i:

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

(or take Δ^* to be the maximum such value ranging over all i). Then

$$\Delta = \sum_i \mathbb{P}[A_i]\Delta^* = \Delta^* \mathbb{E}X$$

Thus we have

Lemma 4.1.13. If $\mathbb{E}X \to \infty$ and $\Delta^* = o(\mathbb{E}X)$, then $X \sim \mathbb{E}X$ and $X > 0$ whp.

Theorem 4.1.14. A threshold function for containing K_4 is $n^{-2/3}$.

Proof. Let X denote the number of copies of K_4 in $G(n,p)$. Then $\mathbb{E}X = \binom{n}{4}p^6 \sim n^4p^6/24$.

If $p \ll n^{-2/3}$ then $\mathbb{E}X = o(1)$ so $X = 0$ whp

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

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 $6\binom{n}{2} = O(n^2)$, and for each such S' one has $\mathbb{P}(A_{S'}|A_S) = p^5$ (as there are 5 additional edges, no in the S-clique, that needs 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 Then

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

Thus $X > 0$ whp by Lemma 4.1.13. □

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

Example 4.1.15 (First moment is not enough)**.** Let $H =$  . We have $\mathbb{E}X_H \asymp n^5p^7$. If $\mathbb{E}X = o(1)$ then $X = 0$ whp. But what if $\mathbb{E}X \to \infty$, 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 $K_4 \subset H$.

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 $H =$  is actually $n^{-2/3}$, and not $n^{-5/7}$ as one might have naively predicted from the first moment alone.

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

Definition 4.1.16. Define the **edge-vertex ratio** of a graph H by $\rho(H) = e_H/v_H$. Define the **maximum edge-vertex ratio of a subgraph** of H:

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

Example 4.1.17. Let $H =$  . 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.

Theorem 4.1.18 (Bollobás 1981)**.** Fix a graph H with v_H vertices and e_H edges. Then $p = n^{-1/m(H)}$ is a threshold function for containing H has a subgraph. Furthermore, if $p \gg n^{-1/m(H)}$, then the number X_H of copies of H in $G(n,p)$ satisfies, with probability $1 - o(1)$,

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

Proof. Let H' be a subgraph of H achieving the maximum edge-vertex ratio, i.e., $\rho(H') = m(H)$.

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,

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

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

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

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 consider 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 4.1.13 yields the result. □

4.2 Existence of thresholds

Question 4.2.1. Does every monotone graph property $\mathcal{P}$ have a threshold function?

E.g., could it be the case that $\mathbb{P}(G(n, n^{-1/3}) \in \mathcal{P}), \mathbb{P}(G(n, n^{-1/4}) \in \mathcal{P}) \in [0.1, 0.9]$ for all sufficiently large n?

First, an even simpler question, why is it that if $\mathcal{P}$ is a nontrivial monotone property, then $\mathbb{P}(G(n,p) \in \mathcal{P})$ is an increasing function of p? This is intuitively obvious, but how to prove it?

Let us give two (related) proofs of this basic fact. Both are quite instructive.

More abstractly, this is not really about graphs, but rather about random subsets (for random graphs, we are taking random subgraphs of edges).

Given a collection $\mathcal{F}$ of subsets of $[n]$, we say that $\mathcal{F}$ is an **upward closed set** (or **up-set**) if whenever $A \subset B$ and $A \in \mathcal{F}$ then $B \in \mathcal{F}$. We say that an up-set $\mathcal{F}$ is nontrivial if $\emptyset \notin \mathcal{F}$ and $[n] \in \mathcal{F}$.

Let $[n]_p$ denote the random subset of $[n]$ obtained by including every element independently with probability p.

Theorem 4.2.2. Let $\mathcal{F}$ a nontrivial up-set of $[n]$. Then $p \mapsto \mathbb{P}([n]_p \in \mathcal{F})$ is a strictly increasing function.

The first proof is by **coupling**. Coupling is powerful probabilistic idea. Given two random variables X and Y with individually prescibed distributions, we "couple" them together by considering a single probabilistic process that generates both X and Y in a way that clarifies their relationship. More formally, we construct a joint distribution (X, Y) whose marginals agree with those of X and Y.

Proof 1. (By coupling) Let $0 \leq p < q \leq 1$. Consider the following process to generate two random subsets of $[n]$: pick a uniform random vector $(x_1, \ldots, x_n) \in [0,1]^n$. Let $A = \{i : x_i \leq p\}$ and $B = \{i : x_i \leq q\}$. Then A has the same distribution as $[n]_p$ and B has the same distribution as $[n]_q$. Furthermore, we see that $A \in \mathcal{F}$ implies $B \in \mathcal{F}$. Thus

$$\mathbb{P}([n]_p \in \mathcal{F}) = \mathbb{P}(A \in \mathcal{F}) \leq \mathbb{P}(B \in \mathcal{F}) = \mathbb{P}([n]_q \in \mathcal{F}).$$

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., $A = \emptyset$ and $B = [n]$). □

The second proof is also uses coupling, but viewed somewhat differently. The idea is that we can obtain $[n]_p$ as the union of several independent $[n]_{p'}$ for some smaller values of p'.

In other words, we are exposing the random subset in several rounds.

Proof 2. (By two-round exposure) Let $0 \le p < q \le 1$. Note that $B = [n]_q$ has the same distribution as the union of two independent $A = [n]_p$ and $A' = [n]_{p'}$, where p' is chosen to satisfy $1 - q = (1 - p)(1 - p')$. Thus

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

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

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).

Theorem 4.2.3 (Bollobás and Thomason 1987)**.** Every nontrivial monotone graph property has a threshold function.

Proof. Note that $G(n, 1-(1-p)^k)$ has the same distribution has the union of k independent copies $G^1, \dots, G^k$ of $G(n, p)$. Furthermore, by the monotonicity of the property, if $G^1 \cup \dots \cup G^k \notin \mathcal{P}$, then $G^1, \dots, G^k \notin \mathcal{P}$. By independence,

$$\mathbb{P}(G(n, 1 - (1 - p)^k) \notin \mathcal{P}) = \mathbb{P}(G^1 \cup \dots \cup G^k \notin \mathcal{P}) \le \mathbb{P}(G^1 \notin \mathcal{P}) \cdots \mathbb{P}(G^k \notin \mathcal{P})$$

To simplify notation, let us write

$$f_p = f_p(n) = \mathbb{P}(G(n, p) \in \mathcal{P}).$$

Since $1 - (1 - p)^k \le kp$ (check by convexity), we have that for any monotone graph property $\mathcal{P}$, any positive integer $k \le 1/p$,

$$1 - f_{kp} \le 1 - f_{1-(1-p)^k} \le (1 - f_p)^k. \tag{4.1}$$

Fix any large enough n (so that set of n-vertex graphs satisfying the property $\mathcal{P}$ is a nontrival up-set). Since $p \mapsto f_p(n)$ 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 for each fixed n), there is some "critical" $p_c = p_c(n)$ with $f_{p_c}(n) = 1/2$.

We claim that p_c is a threshold function. Indeed, (4.1) implies, if $p = p(n) \gg p_c(n)$, then, letting $k = k(n) = \lfloor p/p_c \rfloor \to \infty$,

$$1 - f_p \le (1 - f_{p_c})^k = 2^{-k} \to 0$$

so $f_p \to 1$. Likewise, if $p \ll p_c$, then, letting $k = \lfloor p_c/p \rfloor \to \infty$, we have

$$\frac{1}{2} = 1 - f_{p_c} \leq (1 - f_p)^k,$$

and thus $f_p \to 0$ as $n \to \infty$. Thus $p_c(n)$ is a threshold function for $\mathcal{P}$. □

Remark 4.2.4. Note that, by definition, if $p_1(n)$ and $p_2(n)$ are both threshold functions for the same property, then $cp_1(n) \leq p_2(n) \leq Cp_2(n)$ for some constants $0 < c < C$.

Last section we identified the threshold for the property of containing a fixed subgraph. Let us state the result (at least in the case of triangles, but similar results are known for every subgraph) a bit more precisely, where we use the fact that for a constant $c > 0$, the number of triangles in $G(n, c/n)$ converges to a Poisson distribution with mean $c^3/6$ (this can be proved using the "method of moments" but we will not do it here). So

$$\mathbb{P}\left(G\left(n, \frac{c_n}{n}\right) \text{ contains a triangle}\right) \to \begin{cases} 0 & \text{if } c_n \to 0 \\ 1 - e^{-c^3/6} & \text{if } c_n \to c \text{ constant} \\ 1 & \text{if } c_n \to \infty \end{cases}$$

What about other graph properties? It turns out that we can sometimes identity the transition very precisely.

Example 4.2.5. Here are some more examples of threshold functions. The first two statements are in the original Erdős–Rényi (1959) paper on random graphs. The first is an easy (and instructive) exercise in the second moment method.

- With $p = \dfrac{\log n + c_n}{n}$

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

- With $p = \dfrac{\log n + c_n}{n}$

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

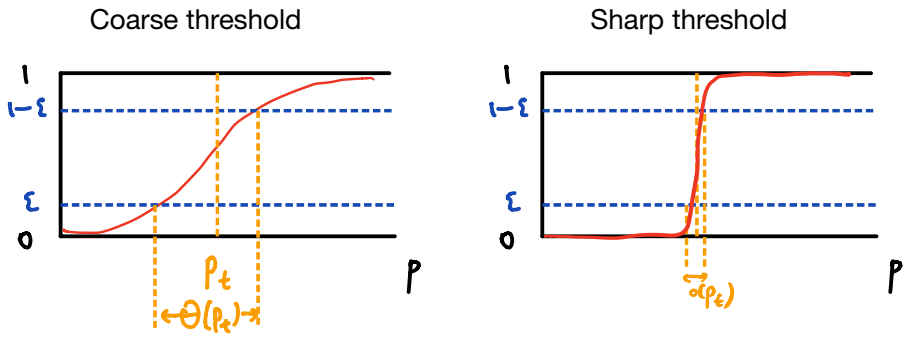


Figure 4: Examples of coarse and sharp thresholds. The vertical axis is the probability that $G(n,p)$ satisfies the property.

In fact, a much stronger statement is true, connecting the above two examples: consider a process where one adds an random edges one at a time, then with probability $1-o(1)$, the graph becomes connected as soon as there are no more isolated vertices.

- With $p = \dfrac{\log n + \log\log n + c_n}{n}$

$$\mathbb{P}(G(n,p) \text{ has a Hamiltonian cycle}) \to \begin{cases} 0 & \text{if } c_n \to -\infty \\ e^{-e^{-c}} & \text{if } c_n \to c \\ 1 & \text{if } c_n \to \infty \end{cases}$$

Like earlier, it is true that with high probability, a random graph becomes Hamiltonian as soon as its minimum degree reaches 2.

In the above 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 function, 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**.

Definition 4.2.6 (Sharp thresholds)**.** We say that r_n is a **sharp threshold** for some graph property $\mathcal{P}$ if, for every $\delta > 0$,

$$\mathbb{P}(G(n,p_n) \text{ satisfies } \mathcal{P}) \to \begin{cases} 0 & \text{if } p_n \le (1-\delta)r_n, \\ 1 & \text{if } p_n \ge (1-\delta)r_n. \end{cases}$$

Equivalently, a graph property $\mathcal{P}$ exhibits a sharp threshold at r_n if, for every $\epsilon > 0$,

for a given large n, as p increases from 0 to 1, the probability $\mathbb{P}(G(n,p) \in \mathcal{P})$ increases from ϵ to $1-\epsilon$ over a short window of width $o(r_n)$ around r_n. On the other hand, if this transition window has width $\Omega(r_n)$ for some $\epsilon > 0$, then we say that it is a **coarse threshold**. See Figure 4.

We saw coarse thresholds for the "local" property of containing some given subgraph, whereas we saw 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, roughly 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 $G_1, \ldots, G_m$ such that $\mathcal{P}$ is "close to" the property of containing one of $G_1, \ldots, G_m$ as a subgraph.

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 $\mathbb{N} = \mathbb{N}_1 \cup \cdots \cup \mathbb{N}_k$ 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 4.2.7 (of Friedgut's sharp threshold theorem)**.** Suppose $r(n)$ is a coarse threshold function of some graph property. Then there is a partition of $\mathbb{N} = \mathbb{N}_1 \cup \cdots \cup \mathbb{N}_k$ and rationals $\alpha_1, \ldots, \alpha_k > 0$ such that $r(n) \asymp n^{-\alpha_j}$ for every $n \in \mathbb{N}_j$.

In particular, if $(\log n)/n$ is a threshold function 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 4.2.8 (k-colorability threshold)**.** For every $k \geq 3$ there is some real constant $d_k > 0$ such that for any constant $d > 0$,

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

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

Theorem 4.2.9 (Achlioptas and Friedgut 2000)**.** For every $k \geq 3$, there exists a function $d_k(n)$ such that for every $\epsilon > 0$, and sequence $d(n) > 0$,

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

On the other hand, it is not known whether $\lim_{n\to\infty} d_k(n)$ exists, which would imply Conjecture 4.2.8. 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).

4.3 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 4.3.1. What is the clique number of $G(n, 1/2)$?

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

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

Theorem 4.3.2. Let $k = k(n)$ satisfy $f(k) \to \infty$. Then $\omega(G(n, 1/2)) \geq k$ whp.

Proof. 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. Let $X = \sum_{S \in \binom{[n]}{k}} X_S$ denote the number of k-cliques.

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

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

It then follows from Lemma 4.1.13 that $X > 0$ (i.e., $\omega(G) \geq k$) whp. □

Theorem 4.3.3 (Bollobás–Erdős 1976 and Matula 1976)**.** There exists a $k = k(n) \sim 2 \log_2 n$ such that $\omega(G(n, 1/2)) \in \{k, k+1\}$ whp.

Proof. (Sketch) For $k \sim 2 \log_2 n$,

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

So the value of $f(k)$ drops rapidly for $k \sim 2 \log_2 n$. Let $k_0 = k_0(n)$ be the value with $f(k_0) \geq 1 > f(k_0 + 1)$. If n is such that $f(k_0) \to \infty$ while $f(k_0 + 1) \to 0$ (it turns out that this is true for most integers n), and thus $\omega(G) = k_0$ whp. When $f(k_0) = O(1)$, we have $f(k_0 - 1) \to \infty$ and $f(k_0 + 1) \to 0$ so one has $\omega(G(n, 1/2)) \in \{k_0 - 1, k_0\}$ whp. □

Remark 4.3.4. The result also implies the same about size of largest independent set in $G(n, 1/2)$ (take complement). Also extends to constant p: $\omega(G(n, p)) \sim 2 \log_{1/(1-p)} n$ whp.

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

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

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 know show that $\chi(G(n, p))$ is tightly concentrated around its mean without a priori needing to know where the mean is located.

4.4 Hardy–Ramanujan theorem on the number of prime divisors

Let $\nu(n)$ denote the number of primes p dividing n (do not count multiplicities).

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

Theorem 4.4.1 (Hardy and Ramanujan 1917)**.** For every $\epsilon > 0$, there exists C such that all but ϵ-fraction of $x \in [n]$ satisfy

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

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

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

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

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

We have $\nu(x) - 10 \le X(x) \le \nu(x)$ 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 $\le n$ are divisible by p, we have

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

So

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

Here we are applying Merten's theorem from analytic number theory: $\sum_{p \le n} 1/p = \log\log n + O(1)$ (the $O(1)$ error term converges to the Meissel–Mertens constant).

Next we compute the variance. The intuition is that distinct primes should be have independently. Indeed, if pq divides n, then X_p and X_q are independent. Then pq does not divide n, but n is large enough, then there is some small covariance contribution. (Contrast to the earlier calculations in random graphs, where there are very few nonzero

covariance terms, but each can be more significant.)

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

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

Thus

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

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

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

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

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

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

Theorem 4.4.2 (Erdős and Kac 1940)**.** With $x \in [n]$ uniformly chosen at random, $\nu(x)$ is asymptotically normal, i.e., for every $\lambda \in \mathbb{R}$,

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

The intuition is that the number of prime divisors $X = \sum_p X_p$ (from the previous proof) behaves like a sum of independent random variables, the central limit theorem should imply an asymptotic normal distribution.

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).

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

Theorem 4.4.3 (Method of moments)**.** Let X_n be a sequence of real valued random variables such that for every positive integer k, $\lim_{n\to\infty} \mathbb{E}[X_n^k]$ equals to the k-th moment of the standard normal distribution. Then X_n converges in distribution to the standard normal, i.e., $\lim_{n\to\infty} \mathbb{P}(X_n \le a) = \mathbb{P}(Z \le a)$ for every $a \in \mathbb{R}$, where Z is a standard normal.

Remark 4.4.4. The same conclusion holds for any probability distribution (other than normal) 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.

Proof sketch of Erdős–Kacs Theorem 4.4.2. 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 $s(n) \to \infty$ sufficiently slowly. As earlier, $\nu(x) - s(n) \le \nu(x) \le v(x)$.

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

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

and

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

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

By the central limit theorem (e.g., the Lindeberg CLT), $\widetilde{Y} \to N(0,1)$ in distribution. In particular, $\mathbb{E}[\widetilde{Y}^k] \sim \mathbb{E}[Z^k]$ (asymptotics as $n \to \infty$) where Z is a standard normal.

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

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

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

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

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

So the moments of $\widetilde{X}$ approach those of $N(0,1)$. The method of moments theorem from probability then implies that $\widetilde{X}$ is asymptotically normally distributed. □

4.5 Distinct sums

Question 4.5.1. Let S be a k-element subset of positive integers such that all 2^k subset sums of S. What is the minimum possible $\max S$?

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

We begin with an easy pigeonhole argument. On the other hand, since all 2^k sums are distinct and are at most $k \max S$, we have $2^k \leq k \max S$, so $\max S \geq 2^k/k$.

Erdős offered \$300 for a proof or disproof that $\max S \gtrsim 2^k$. This remains an interesting open problem.

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 4.5.2. Let S be a k-element subset of positive integers such that all 2^k subset sums of S. Then $\max S \gtrsim 2^k/\sqrt{k}$.

Proof. Let $S = \{x_1, \ldots, x_k\}$ and $n = \max S$. Set

$$X = \epsilon_1 x_1 + \cdots + \epsilon_k x_k$$

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

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

and

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

By Chebyshev,

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

Since X takes distinct values for every $(\epsilon_1, \dots, \epsilon_k) \in \{0,1\}^k$, we have $\mathbb{P}(X = x) \le 2^{-k}$ for all x, so we have the lower bound

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

Putting them together, we get

$$2^{-k}(2n\sqrt{k} + 1) \le \frac{3}{4}.$$

So $n \gtrsim 2^k/\sqrt{k}$. □

Recently, this July, Dubroff–Fox–Xu gave another short proof of this result (with an improved error term $O(1)$) 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).

Here for the "n-dimensional boolean cube" we consider the graph on the vertex set $\{0,1\}^n$ with an edge between every pair of n-tuples that differ in exactly one coordinate. Given $A \subseteq \{0,1\}^n$, let δA be the set of all vertices outside A that is adjacent to some vertex of A.

Theorem 4.5.3 (Harper 1966)**.** Every $A \subset \{0,1\}^k$ with $|A| = 2^{k-1}$ has $|\partial A| \ge \binom{k}{\lfloor k/2 \rfloor}$.

Remark 4.5.4. Harper's theorem, more generally, gives the precise value of $\min_{A \subset \{0,1\}^n : |A| = m} |\partial A|$ for every (n, m). Basically, the minimum is achieved when A is a 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 4.5.5 (Dubroff–Fox–Xu)**.** If S is a set of k positive integers with distinct subset sums, then

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

Remark 4.5.6. The above bound has the currently best known leading constant factor.

Proof. Let $S = \{x_1, \dots, x_k\}$. Let

$$A = \left\{ (\epsilon_1, \dots, \epsilon_k) \in \{0,1\}^k : \epsilon_1 x_1 + \cdots + \epsilon_1 x_k < \frac{x_1 + \cdots + x_k}{2} \right\}.$$

Note that due to the distinct sum hypothesis, one can never have $x_1 s_1 + \cdots + x_n s_n = (s_1 + \cdots + s_n)/2$. It thus follows by symmetry that $|A| = 2^{k-1}$.

Note that every element of ∂A corresponds to some subset sum in the open interval

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

Since all subset sums are distinct, we must have $\max S \geq |\partial A| \geq \binom{k}{\lfloor k/2 \rfloor}$ by Harper's theorem (Theorem 4.5.3). □

4.6 Weierstrass approximation theorem

We finish off the chapter with an application to analysis.

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

Theorem 4.6.1 (Weierstrass approximation theorem 1885)**.** Let $f\colon [0,1] \to \mathbb{R}$ be a continuous function. Let $\epsilon > 0$. Then there is a polynomial $p(x)$ such that $|p(x) - f(x)| \leq \epsilon$ for all $x \in [0,1]$.

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

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

where $E_j(x)$ chosen as some polynomials peaks at $x = i/n$ and then decaysaway from $x = i/n$. To this end, set set

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

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

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

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

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

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

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

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

Note that in contrast, $\operatorname{Var} S_n = n$, so Chebyshev's inequality would only give a tail bound $\leq 1/\lambda^2$

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

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

We have (by comparing Taylor series coefficients $\frac{1}{(2n)!} \leq \frac{1}{n!2^n}$), for all $t \geq 0$,

$$\frac{e^{-t} + e^t}{2} \leq e^{t^2/2}.$$

By Markov's inequality,

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

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

Remark 5.0.2. The technique of considering the moment generating function can 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 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. $\mathbb{P}(|S_n| \geq \lambda\sqrt{n}) \leq 2e^{-\lambda^2/2}$

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. Let each X_i be an independent random variable taking values in $[-1, 1]$ and $\mathbb{E}X_i = 0$. Then $S_n = X_1 + \cdots + X_n$ 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 (not necessarily the same probability). Let $\mu = \mathbb{E}X$ and $\lambda > 0$. Then 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

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

Set $t = \lambda$ and get

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

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

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

The same proof method allows you to prove bounds for other sums of random variables, suitable for whatever application you have in mind. 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 $\epsilon > 0$,

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

and

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

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 $X \sim \mathrm{Bin}(n, c/n)$, which, for a constant $c > 0$, converges as $n \to \infty$ to a Poisson distribution with mean c, whose value at k is $c^k e^k/k! = e^{-\Theta(k \log k)}$ and not $e^{-\Omega(k^2)}$ 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, $\epsilon \in [0, 1]$.

5.1 Discrepancy

Theorem 5.1.1. Let $\mathcal{F}$ be a collection of m subsets of $[n]$. Then there exists some assignment $[n] \to \{-1, 1\}$ so that the sum on every set in $\mathcal{F}$ is at most $2\sqrt{n \log m}$ 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 $2\sqrt{n \log m}$. □

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 semi-random assignment algorithm.

Theorem 5.1.3 (Spencer (1985))**.** Let $\mathcal{F}$ be a collection of n subsets of $[n]$. Then there exists some assignment $[n] \to \{-1, 1\}$ 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 $11\sqrt{n \log(2m/n)}$.

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. Another very nice algorithm with another beautiful proof of the algorithmic result was given by Lovett and Meka (2015).

Here is a famous conjecture on discrepancy.

Conjecture 5.1.5 (Komlós)**.** There exists some absolute constant K so that for every set of vectors $v_1,\dots,v_m$ in the unit ball in R^n, there exists signs $\epsilon_1,\dots,\epsilon_m \in \{-1,1\}$ such that

$$\epsilon_1 v_1 + \cdots + \epsilon_m v_m \in [-K,K]^n.$$

Banaszczyk (1998) proved the bound $K = O(\sqrt{\log n})$ in a beautiful paper using deep ideas from convex geometry.

Spencer's theorem's implies the Komlós conjecture if 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 $m \le n$. 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 $|\{i : a_i \notin \{-1,1\}\}| \le n$ such that

$$a_1 v_1 + \cdots + a_m v_m = 0$$

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 to move $(a_i)_{i\in I}$'s to new values, while preserving $a_1 v_1 + \cdots + a_m v_m = 0$, and end up with at one additional a_i taking $\{-1,1\}$-value. □

Letting $a_1,\dots,a_m$ and $I = \{i : a_i \notin \{-1,1\}\}$ as in the Lemma, we then take $\epsilon_i = a_i$ for all $i \notin I$, and apply a corollary of Spencer's theorem to find $\epsilon_i \in \{-1,1\}^n$, $i \in I$ with

$$\sum_{i\in I} (\epsilon_i - a_i) v_i \in [-K,K]^n,$$

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 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** if 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 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.

Hadwidger posed the following conjecture, which is one of the biggest open conjectures in graph theory.

Conjecture 5.2.1 (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 found 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!)

To be continued

Theorem 5.2.2. With probability $1 - o(1)$, $G(n, 1/2)$ has no K_t-subdivision with $t = \lceil 10\sqrt{n} \rceil$.

From Theorem 4.3.3 we show 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 $S \subset V$, $|S| = t$, then at most $n - t \leq n$ of the edges in the subdivision can be paths with at least two edges (since they must use distinct vertices outside S). So S must induce at least $\binom{t}{2} - n \geq \frac{3}{4}\binom{t}{2}$ edges in G.

By Chernoff bound, for fixed t-vertex subset S

$$\mathbb{P}\left(e(S) \geq \frac{3}{4}\binom{t}{2}\right) \leq e^{-t^2/10}.$$

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 5.2.3. One can ask the following quantitative question regarding Hadwidger's conjecture:

Can we show that 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.

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. The best current bound was proved this past summer. Postle (2020+) showed that if every graph with no K_t-minor is $O(t(\log \log t)^6)$-colorable.

For more on Hadwiger's conjecture, see Seymour's survey (2016).

6 Lovász local lemma

The Lovász local lemma (LLL), introduced in the paper of Erdős and Lovász (1975) is a powerful tool in the probabilistic method. It is some form of interpolation between the following two extreme (easy) scenerios

- Complete independence: if we have an arbitrary number of independent bad events, each occurring with probability < 1, then it is possible to avoid all of them (although with tiny probability)
- Union bound: if we have a collection of bad events whose total probability is < 1 (but usually much smaller), then it is possible to avoid all of them (often with high probability)

The local lemma deals with the case when each bad event is independent with most other bad events, but possibly dependent with a small number of other events.

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

Here is the **setup** for the local lemma:

- We have "bad events" $A_1, A_2, \ldots, A_n$
- For each i there is some subset $N(i) \subseteq [n]$ such that A_i is independent from $\{A_j : j \notin N(i) \cup \{i\}\}$.

Here we say that event A_0 is **independent** from $\{A_1, \ldots, A_m\}$ if A_0 is independent of every event of the form $B_1 \wedge \cdots \wedge B_m$ where each B_i is either A_i or $\overline{A}_i$, i.e.,

$$\mathbb{P}(A_0 B_1 \cdots B_m) = \mathbb{P}(A_0)\mathbb{P}(B_1 \cdots B_m),$$

or, equivalently, using Bayes's rule: $\mathbb{P}(A_0 | B_1 \cdots B_m) = \mathbb{P}(A_0)$. (Here $\wedge$ = 'and' and $\vee$ = 'or', and we may omit $\wedge$ symbols, similar to multiplication)

We can represent the above relations by a **dependency (di)graph** whose vertices are indexed by the events (or equivalently $V = [n]$), and the (out-)neighbors of i are $N(i)$. (Mostly we'll just work with undirected dependency graphs for simplicity, but in general it may be helpful to think of them as directed—hence digraphs.)

Remark 6.1.1 (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., $\mathbb{P}(A_iA_j) \neq \mathbb{P}(A_i)\mathbb{P}(A_j)$).

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).

A related note: there could be more than one choices for dependency graphs. So we speak of "a dependency graph" instead of "the dependency graph."

Remark 6.1.2 (**Random variable model** / **hypergraph coloring**). Many common applications of the local lemma can be phrased in the following form:

- A collection of independent random variables $x_1, \ldots, x_N$
- Each event A_i only depends on $\{x_j : j \in S_i\}$ for some subset $S_i \subseteq [N]$

In this case, valid dependency graph can be formed by placing an edge $i \sim j$ whenever $S_i \cap S_j \neq \emptyset$.

We can also view the above as coloring a hypergraph with vertices labeled by $[N]$, using independent random colors $x_1, \ldots, x_N$ for each vertex, so that various constraints on edges $S_1, S_2, \cdots \subseteq [N]$ are satisfied.

An example of such a problem is the **satisfiability problem (SAT)**: given a **CNF formula** (conjunctive normal norm = *and-of-or*'s), e.g.,

$$(x_1 \vee x_2 \vee x_3) \wedge (\overline{x_1} \vee x_2 \vee x_4) \wedge (\overline{x_2} \vee x_4 \vee x_5) \wedge \cdots$$

the problem is to find a satisfying assignment with boolean variables $x_1, x_2, \ldots$. Many problems in computer science can be modeled using this way.

The following formulation of the local lemma is easiest to apply and is the most commonly used.

Theorem 6.1.3 (Lovász local lemma; symmetric form)**.** Let $A_1, \ldots, A_n$ 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 6.1.4. The constant e is best possible (Shearer 1985).

Theorem 6.1.5 (Lovász local lemma; general form)**.** Let $A_1, \ldots, A_n$ 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 $x_1, \ldots, x_n \in [0, 1)$ satisfy

$$\mathbb{P}(A_i) \leq x_i \prod_{j \in N(i)} (1 - x_j) \quad \forall i \in [n],$$

then with probability $\geq \prod_{i=1}^{n}(1 - x_i)$, none of the events A_i occur.

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. □

Here is another corollary of the general form. It says that the local lemma works if the total probability of any neighborhood in a dependency graph is small.

Corollary 6.1.6. In the setup of Theorem 6.1.5, 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.

Proof. In Theorem 6.1.5, set $x_i = 2\mathbb{P}(A_i)$ for each i. Then

$$x_i \prod_{j \in N(i)} (1 - x_j) \geq x_i \left(1 - \sum_{j \in N(i)} x_j\right) = 2\mathbb{P}(A_i)\left(1 - \sum_{j \in N(i)} 2\mathbb{P}(A_i)\right) \geq \mathbb{P}(A_i).$$

(The first inequality is by "union bound.") □

Proof of Lovász local lemma (general case). We will prove that

$$\mathbb{P}\left(A_i \,\middle|\, \bigwedge_{j \in S} \overline{A}_j\right) \leq x_i \quad \text{whenever } i \notin S \subseteq [n] \tag{6.1}$$

Once (6.1) has been established, we then deduce that

$$\begin{aligned}\mathbb{P}(\overline{A}_1 \cdots \overline{A}_n) &= \mathbb{P}(\overline{A}_1)\mathbb{P}\left(\overline{A}_2 \mid \overline{A}_1\right)\mathbb{P}\left(\overline{A}_3 \mid \overline{A}_1\overline{A}_2\right)\cdots\mathbb{P}\left(\overline{A}_n \mid \overline{A}_1 \cdots \overline{A}_{n-1}\right) \\ &\geq (1 - x_1)(1 - x_2)\cdots(1 - x_n),\end{aligned}$$

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

$$\mathbb{P}\left(A_i \,\middle|\, \bigwedge_{j \in S} \overline{A}_j\right) = \frac{\mathbb{P}\left(A_i \bigwedge_{j \in S_1} \overline{A}_j \,\middle|\, \bigwedge_{j \in S_2} \overline{A}_j\right)}{\mathbb{P}\left(\bigwedge_{j \in S_1} \overline{A}_j \,\middle|\, \bigwedge_{j \in S_2} \overline{A}_j\right)} \tag{6.2}$$

For the RHS of (6.2),

$$\text{numerator} \le \mathbb{P}\left(A_i \,\middle|\, \bigwedge_{j \in S_2} \overline{A}_j\right) = \mathbb{P}(A_i) \le x_i \prod_{j \in N(i)} (1 - x_i) \tag{6.3}$$

and, writting $S_1 = \{j_1, \ldots, j_r\}$,

$$\begin{aligned}
\text{denominator} &= \mathbb{P}\left(\overline{A}_{j_1} \,\middle|\, \bigwedge_{j \in S_2} \overline{A}_j\right) \mathbb{P}\left(\overline{A}_{j_2} \,\middle|\, \overline{A}_{j_1} \bigwedge_{j \in S_2} \overline{A}_j\right) \cdots \mathbb{P}\left(\overline{A}_{j_r} \,\middle|\, \overline{A}_{j_1} \cdots \overline{A}_{j_{r-1}} \bigwedge_{j \in S_2} \overline{A}_j\right) \\
&\ge (1 - x_{j_1}) \cdots (1 - x_{j_r}) \qquad \text{[by induction hypothesis]} \\
&\ge \prod_{j \in N(i)} (1 - x_i)
\end{aligned}$$

Thus (6.2) $\le x_i$, thereby finishing the induction proof of (6.1). □

6.2 Algorithmic local lemma

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 there exists some efficient algorithm to sample a good configuration in applications of the local lemma.

Moser (2009), during his PhD, achieved a breakthrough by coming up with the first efficient algorithmic version of the local lemma. Later, in a beautiful paper by Moser and Tardos (2010) extended the algorithm to a general framework for the local lemma.

The Moser–Tardos algorithm considers problems in the random variable model (Re-

mark 6.1.2). The algorithm is surprisingly simple.

Algorithm: Moser–Tardos local lemma algorithm

Initialize all the random variables;
while *there are violated events* **do**
    Pick an arbitrary violated event and resample its variables;

Theorem 6.2.1 (Moser and Tardos 2010)**.** If there are $x_1, \dots, x_n \in [0, 1)$ such that

$$\mathbb{P}(A_i) \le x_i \prod_{j \in N(i)} (1 - x_j) \quad \forall i \in [n],$$

then the above randomized algorithm resamples each A_i at most $x_i/(1 - x_i)$ times in expectation for each i.

Remark 6.2.2. The above theorem shows that the Moser–Tardos algorithm is an *Las Vegas* algorithm with polynomial expected runtime. A Las Vegas algorithm is a randomized algorithm that always terminates a successful result, but it might take a long time to terminate. Contrast this to a *Monte Carlo* algorithm, which runs in bounded time but may return a bad result with some small probability, and there may not be an efficient way to check whether the output is correct–e.g., randomly 2-coloring the edges of K_n to avoid a monochromatic $2 \log_2 n$-clique. A Las Vegas algorithm can be converted into a Monte Carlo algorithm by cutting off the algorithm after some time (significantly larger than the expected running time) and applying Markov's inequality to bound the probability of failure. On the other hand, there is in general no way to convert a Monte Carlo algorithm to a Las Vegas algorithm unless there is an efficient way to certify the correctness of the output of the algorithm.

Remark 6.2.3. The Moser–Tardos algorithm assumes the random variable model. Some assumption on the model is necessary since the problem can be computationally hard in general.

For example, let $q = 2^k$, and $f \colon [q] \to [q]$ be some fixed bijection. Let $y \in [q]$ be given. The goal is find x such that $f(x) = y$.

For each $i \in [k]$, let A_i be the event that $f(x)$ and y disagree on i-th bit. Then $A_1, \dots, A_k$ independent (check!). Also, $f(x) = y$ if and only if no event A_i occurs.

A trivial version of the local lemma (with empty dependency graph) guarantees the existence of some x such that $f(x) = y$.

However, finding x may be computationally hard for certain functions f. In fact, the existence of such one-way functions (easy to compute but hard to invert) is the bedrock of cryptography. A concrete example is $f \colon \mathbb{F}_q \to \mathbb{F}_q$ is given by $f(0) = 0$, and for $x \neq 0$, set

$f(x) = g^x$ for some multiplicative generator. Then inverting f is the **discrete logarithm problem**, which is believed to be computationally difficult.

6.3 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 6.3.1. 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. □

Corollary 6.3.2. 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 $k \geq 9$. □

Remark 6.3.3. 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 asymmetric form of the local lemma is insufficient (why is it insufficient? No bound on the the number of dependent events).

Theorem 6.3.4. Let H be a (non-uniform) hypergraph where every edge has size 3. Suppose

$$\sum_{f \in E(H) \setminus \{e\} : e \cap f \neq} 2^{-|f|} \leq \frac{1}{8}, \quad \text{for each edge } e,$$

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 $\mathbb{P}(A_e) = 2^{-|e|+1} \le 1/4$ since $|e| \ge 3$. Also, also

$$\sum_{f \in E(H)\setminus\{e\} : e \cap f \neq} \mathbb{P}(A_f) = \sum_{f \in E(H)\setminus\{e\} : e \cap f \neq} 2^{-|f|+1} \le 1/4.$$

Thus by Corollary 6.1.6 one can avoid all events A_e, and hence H is 2-colorable. □

Remark 6.3.5. A sign for when you should look beyond the symmetric local lemma is when there are bad events of very different nature (in particular, they have very different probabilities).

6.3.1 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 6.3.6. Let H be a (non-uniform) hypergraph on a possibly infinite vertex set, such that each edges is finite, has at least k vertices, and intersect at most d other edges. If $e2^{-k+1}(d+1) \le 1$, 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 an 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 $e \subset X$ 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 (so that open subsets are those defined by restriction to a finite set of coordinates).

For each finite subset X, let $C_X \subset [2]^V$ be the subset of colorings where no edge contained in X is monochromatic. Earlier from the local lemma we saw that $C_X \neq$. Furthermore,

$$C_{X_1} \cap \cdots \cap C_{X_\ell} \supseteq C_{X_1 \cup \cdots \cup X_\ell},$$

so $\{C_X : |X| < \infty\}$ is a collection of closed subsets of $[2]^V$ with the finite intersection property. Hence by compactness of $[2]^V$, we have $\bigcap_{X \subset V : |X| < \infty} \neq \emptyset$, and observe that any element of this intersection is a valid coloring of the hypergraph. □

Note that we may have $\mathbb{P}[\bigwedge_i A_i] = 0$ while $\bigwedge_i A_i \neq \emptyset$.

The same compactness argument tell us that: in the **random variable model** (Remark 6.1.2), **if it is possible to avoid every finite subset of bad events, then it is possible to avoid all bad events simultaneously**. (Again, one needs to be working in the random variable model for the compactness argument to work.

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 $c\colon \mathbb{R} \to [k]$. We say that $T \subset \mathbb{R}$ is **multicolored** with respect to c if all k colors appear in T

Question 6.3.7. For each k is there an m so that for every $S \subset \mathbb{R}$ 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 $m > (3+\epsilon)k \log k$ (and $k > k_0(\epsilon)$ sufficiently large).

Theorem 6.3.8. 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. By the compactness argument, it suffices to check the result for every finite $X \subset \mathbb{R}$.

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, and conclude by the symmetric version of the local lemma. □

6.4 Decomposing coverings

We say that a collection of disks in $\mathbb{R}^d$ is a **covering** is 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 a 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 6.4.1 (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 6.4.2. In $\mathbb{R}^d$, the same proof gives $\geq c_d 2^{k/d}$.

We will need the following combinatorial geometric fact:

Lemma 6.4.3. 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 $g(m+1) \leq g(m) + f(m) \leq g(m)$ by a similar argument as earlier. So $g(m) \leq f(m-1) + f(m-2) + \cdots + f(1) + g(0) \leq m^3$. □

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.

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. ■

With c sufficiently small, we have $e2^{-k+1}(d+1) \leq 1$. It then follows by Theorem 6.3.6

(local lemma + compactness argument) that this hypergraph is 2-colorable, which corresponds to a decomposition of the covering, a contradiction. □

6.5 Large independent sets

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 a certain structure on the independent set.

Theorem 6.5.1. Let $G = (V, E)$ be a graph with maximum degree Δ and let $V = V_1 \cup \cdots \cup V_r$ 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.

This example is instructive because it is not immediately obvious what to choose as bad events (even if you are already told to apply the local lemma).

We may assume that $|V_i| = k := \lceil 2e\Delta \rceil$ for each i, or else we can remove some vertices from V_i.

Pick $v_i \in V_i$ uniformly at random, independently for each i.

What do we choose as bad events $A_\bullet$? It turns out that some choices work better than others.

Attempt 1:
$A_{i,j} = \{v_i \sim v_j\}$ for each $1 \leq i < j \leq r$ where there is an edge between V_i and V_j
$\mathbb{P}(A_{i,j}) \leq \Delta/k$
Dependency graph: $A_{i,j} \sim A_{k,\ell}$ if $\{i,j\} \cap \{k,\ell\} \neq \emptyset$. 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:
$A_e = \{\text{both endpoints of } e \text{ are chosen}\}$ for each $e \in E$
$\mathbb{P}(A_e) = 1/k^2$
Dependency graph: $A_e \sim A_f$ if some V_i intersects both e and f. 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, and hence $\{v_1, \ldots, v_r\}$ is an independent set.

6.6 Directed cycles of length divisible by k

Theorem 6.6.1 (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.

(d-regular means in-degree and out-degree of every vertex is d)

Corollary 6.6.2. 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. □

More generally they proved:

Theorem 6.6.3 (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 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

$$\mathbb{P}(A_v) = (1 - 1/k)^\delta \leq e^{-\delta/k}.$$

The following is a valid dependency graph, noting that A_v only depends 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:

$A_v \sim A_w$ if $\{v\} \cup N^+(v)$ 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, as long as $e^{1-\delta/k}(1+\Delta+\delta\Delta)$, i.e.,

$$k \le \delta/(1+\log(1+\Delta+\delta\Delta)),$$

then we are done. This is almost, but not quite the result (though, for most application, we would be perfectly happy with such a bound).

The final trick is to notice that we actually have an even smaller dependency digraph:

A_v is independent of all A_w where $N^+(v)$ 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

$$ep(\delta\Delta+1) \le e^{1-\delta/k}(\delta\Delta+1) \le 1.$$

So we are done by the local lemma. □

6.7 Lopsided local lemma

In the dependency graph, intuitively, the neighbors of A_i consists of all the events dependent on A_i (again, same warning as earlier: it is insufficient to simply check for pairwise dependence). However, if the there is a positive dependence among the bad events—avoiding some bad events make it easier to avoid others—then perhaps it would actually make it easier to avoid all bad events. For example, in an extreme scenario, if several bad events are identical, so that they are perfectly positively correlated, then it is much easier to avoid them compared to avoiding independent bad events. In the opposite extreme, if several bad events are disjoint, then it would be harder to avoid all of them. Thus, intuitively, it seems reasonable that in the local lemma, we are primarily concerned about negative dependencies and but not positive dependencies among 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):

$$\text{numerator} \le \mathbb{P}\left(A_i \,\middle|\, \bigwedge_{j\in S_2} \overline{A}_j\right) = \mathbb{P}(A_i) \le x_i \prod_{j\in N(i)} (1-x_i).$$

If we had changed the middle $=$ to $\le$, the whole proof would remain valid. This observation allows us to weaken the independence assumption. Therefore we have the following

theorem.

Theorem 6.7.1 (Lopsided local lemma — Erdős and Spencer 1991)**.** Let $A_1, \dots, A_n$ be events. For each i, let $N(i) \subset [n]$ be such that

$$\mathbb{P}\left(A_i \,\middle|\, \bigwedge_{j \in S} \overline{A}_j\right) \leq \mathbb{P}(A_i) \quad \forall i \in [n] \text{ and } S \subseteq [n] \setminus (N(i) \cup \{i\}) \tag{6.4}$$

Suppose there exist $x_1, \dots, x_n \in [0, 1)$ such that

$$\mathbb{P}(A_i) \leq x_i \prod_{j \in N(i)} (1 - x_j) \quad \forall i \in [n].$$

Then with probability $\geq \prod_{i=1}^{n}(1 - x_i)$ none of the event A_i occur.

Like earlier, we also have a symmetric version that is easier to apply.

Corollary 6.7.2 (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**. Erdős and Spencer called it the **lopsidependency graph**, though I prefer "negative dependency graph" since it is more descriptive.

Remark 6.7.3. Here are several equivalent formulations of (6.4): for every $i \in [n]$ and $S \subseteq [n] \setminus (N(i) \cup \{i\})$,

- $\mathbb{P}\left(\overline{A}_i \,\middle|\, \bigwedge_{j \in S} \overline{A}_j\right) \geq \mathbb{P}(\overline{A}_i)$
- $\mathbb{P}\left(A_i \bigwedge_{j \in S} \overline{A}_j\right) \leq \mathbb{P}(A_i)\mathbb{P}\left(\bigwedge_{j \in S} \overline{A}_j\right)$

To put in words, each event is non-negatively dependent on its non-neighbors.

It may be slightly strange to think about at first, but to verify the validity of a *negative* dependency graph, we are actually checking *nonnegative* dependencies (against non-neighbors). Likewise, earlier, to verify a dependency graph, we need to check independence against non-neighbors.

Remark 6.7.4. From the proof of Theorem 6.7.1, we see that we can weaken the negative dependency hypothesis to

$$\mathbb{P}\left(A_i \,\middle|\, \bigwedge_{j \in S} \overline{A}_j\right) \leq x_i \prod_{j \in N(i)} (1 - x_j) \quad \forall i \text{ and } S \subseteq [n] \setminus N(i).$$

Though negative dependency is often easier to argue.

6.7.1 Random permutations and positive dependencies

Just like how most applications of the local lemma can be cast in terms of the the random variable model, which makes it easy to produce a valid dependency graph (by looking at shared random variables), a natural setting for applications of the lopsided local lemma is that of random permutations (and, by extending the domain, also random injections).

Here is a model problem: what is the probability that a random permutation π of $[n]$ has no fixed points? (Such permutations are called "derangements")

This problem can be solved exactly: using inclusion-exclusion, one can deduce that probability to be $\sum_{i=0}^{n}(-1)^i/i! = e^{-1} + o(1)$. Suppose that we did not know this answer.

Let A_i be the event that $\pi(i) = i$. It is easy to see that $\mathbb{P}(A_i) = 1/n$. If the events $A_1, \dots, A_n$ were independent, then we would deduce that with probability $(1 - 1/n)^n = 1/e + o(1)$ none of the A_i occur. But these events are not independent.

Intuitively, these events are positively dependent: having some fixed points makes it likes to see other fixed points. The next theorem makes this rigorous, so that we can deduce $\mathbb{P}(\overline{A}_1 \dots \overline{A}_n) \geq \mathbb{P}(\overline{A}_1) \cdots \mathbb{P}(\overline{A}_n) = (1 - 1/n)^n = 1/e - o(1)$, a lower bound that matches the truth.

It may be easier to visualize permutations as perfect matchings in the complete bipartite graph $K_{n,n}$. We will use these two interpretations interchangeably.

Theorem 6.7.5 (Positive dependence for random perfect matchings)**.** Let M be a perfect matching of $K_{n,n}$ chosen uniformly at random. For each matching F, let A_F denote the event that $F \subseteq M$.
Let $F_0, F_1, \dots, F_k$ be matchings such that no edge of F_0 shares a vertex with any edge from $F_1 \cup \cdots \cup F_k$. Then

$$\mathbb{P}\left(A_{F_0} \,\middle|\, \overline{A}_{F_1} \cdots \overline{A}_{F_k}\right) \leq \mathbb{P}(A_{F_0}).$$

In other words, if $\mathcal{F}$ is a set of matchings in $K_{n,n}$, then the following if a valid negative dependency graph on the events $\{A_F : F \in \mathcal{F}\}$: $A_{F_1} \sim A_{F_2}$ if F_1 and F_2 touch (i.e., some two edges coincide or share an endpoint).

Proof. By relabeling, we may assume that the edges of F_0 are $(1,1), (2,2), \dots, (t,t)$.

For each injection $\tau\colon [t] \to [n]$ (also viewed as a matching with edges $(1, \tau(1)), \dots, (t, \tau(t))$), let $\mathcal{M}_\tau$ denote the set of perfect matchings in $K_{n,n}$ containing τ but not containing any of $F_1, \dots, F_k$.

Let $\tau_0 \colon [t] \to [n]$ be the map sending i to i (i.e., the matching F_0). Then the LHS and RHS of the desired inequality $\mathbb{P}\left(A_{F_0} \,\middle|\, \overline{A}_{F_1} \cdots \overline{A}_{F_k}\right) \le \mathbb{P}(A_{F_0})$ can be rewritten as

$$\frac{|\mathcal{M}_{\tau_0}|}{\sum_\tau |\mathcal{M}_\tau|} \le \frac{1}{n(n-1)\cdots(n-t+1)},$$

where the sum is taken over all $n(n-1)\cdots(n-t+1)$ injections $\tau \colon [t] \to [n]$. Thus it suffices to prove that

$$|\mathcal{M}_{\tau_0}| \le |\mathcal{M}_\tau| \quad \text{for every injection } \tau \colon [t] \to [n].$$

To show this inequality, we construct an injection $\mathcal{M}_{\tau_0} \to \mathcal{M}_\tau$. Intuitively, this injection is obtained by permuting some of the vertices on the right-half of $K_{n,n}$ so that the matching τ_0 to taken to τ. Let us illustrate this idea in a simple case when $\tau(i) = t + i$ for each $i \in [t]$: we construct $\mathcal{M}_{\tau_0} \to \mathcal{M}_\tau$ by swapping, in $K_{n,n}$, the i-th vertex on the right-half with the $(t+i)$-th vertex on the right-half, for each $i \in [n]$.

More generally, extend $\tau \colon [t] \to [n]$ to a permutation σ on $[n]$ sending $\tau([t]) \setminus [t]$ to $[t] \setminus \tau([t])$ and otherwise leaving $[n] \setminus ([t] \cup \tau([t]))$ fixed as identity.

Then σ acts on the set of matchings in $K_{n,n}$ by permuting the right-endpoints. In particular, σ sends τ_0 to τ. Also σ permutes the set of perfect matchings of $K_{n,n}$.

It remains to show that if $M \in \mathcal{M}_{\tau_0}$, then its image σM lies in $\mathcal{M}_\tau$. By construction $\tau \subset \sigma M$. Suppose $F_i \subset \sigma M$ for some $i \in [k]$. Since F_i does not share any vertex with F_0, all the left-endpoints in F_i lie in $[n] \setminus [t]$. Since $(i, \tau(i))$ is an edge of σM, all the right-endpoints in F_i lie in $[n] \setminus ([t] \cup \tau([t]))$. It follows that $\tau F_i = F_i$, so that $F_i \subset M$, which contradicts $M \in \mathcal{M}_{\tau_0}$.

Thus σ induces an injection from $\mathcal{M}_{\tau_0}$ to $\mathcal{M}_\tau$. □

6.7.2 Latin square transversals

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:

1	2	3
2	3	1
3	1	2

(Name origin: The name Latin square was inspired by mathematical papers by Leonhard Euler (1707–1783), who used Latin characters as symbols)

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:

$$\begin{array}{ccc} \mathbf{1} & 2 & 3 \\ 2 & \mathbf{3} & 1 \\ 3 & 1 & \mathbf{2} \end{array}$$

Here is a famous open conjecture about Latin transversals. (Can you see why "odd" is necessary?)

Conjecture 6.7.6. Every odd order Latin square has a transversal.

The next result is the original application of the lopsided local lemma.

Theorem 6.7.7 (Erdős and Spencer 1991)**.** Every $n \times n$ array where every entry appears at most $n/(4e)$ times has a Latin transversal.

Proof. Let (m_{ij}) be the array. Pick a transversal uniformly at random. For each pair of equal entries $m_{ij} = m_{kl}$ in the array in distinct rows and distinct columns, consider the event $A_{ijkl} = A_{klij}$ that the transversal contains both locations (i, j) and (k, l). Then $\mathbb{P}(A_{ijkl}) = 1/(n(n-1))$. (By reinterpreting in the earlier language of matchings, A_{ijkl} is the event that the random perfect matchings contains the two edges (i, j) and (k, l), which are assigned identical edge-labels.)

By the earlier theorem, the following is a negative dependency graph: two pairs of entries are adjacent if they share some row or column, i.e., $A_{ijkl} \sim A_{i'j'k'l'}$ unless $|\{i, k, i', k'\}| = |\{j, l, j', l'\}| = 4$.

Let us count neighbors in this negative dependency graph. Given A_{ijkl}, there are at most $4n-4$ additional locations (x, y) that share a column or row with either of the two chosen entries (i, j) and (k, l). Once we have chosen (x, y), there are at most $n/(4e)-1$ choices for another (z, w) with $m_{xy} = m_{zw}$. Thus the maximum degree in this negative dependence graph is at most $(4n-4)\left(\frac{n}{4e} - 1\right) \leq \frac{n(n-1)}{e} - 1$. We can now apply the symmetric lopsided local lemma to conclude that with positive probability, none of the events A_{ijkl} occur. □

7 Correlation inequalities

Consider an Erdős–Rényi random graph $G(n, p)$. If we condition on it having a Hamiltonian cycle, intuitively, it seems that this conditioning would cause us to have more edges and thereby decreasing the likelihood that the random graph is planar. The main theorem of this chapter, the Harris–FKG inequality, makes this notion precise.

7.1 Harris–FKG inequality

Setup. We have n independent Bernoulli random variables $x_1, \dots, x_n$ (not necessarily identical, but independence is important).

An **increasing event** (or increasing property) A is defined by an upward closed subset of $\{0,1\}^n$ (an **up-set**), i.e.,

$$x \in A \text{ and } x \le y \text{ (coordinatewise)} \quad \implies \quad y \in A.$$

Examples in increasing properties of graphs:

- Having a Hamiltonian cycle
- Connected
- Average degree ≥ 4 (or: min degree, max degree, etc.)
- Having a triangle
- Not 4-colorable

Similarly, a **decreasing event** is defined by a downward closed collection of subset of $\{0,1\}^n$.

Note that $A \subset \{0,1\}^n$ is increasing if and only if its complement $\overline{A} \subset \{0,1\}^n$ is decreasing

The main theorem of this chapter is the following, which 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

$$\mathbb{P}(A \wedge B) \ge \mathbb{P}(A)\mathbb{P}(B)$$

Equivalently, we can write $\mathbb{P}(A \mid B) \ge \mathbb{P}(A)$.

Remark 7.1.2. Many of such inequalities were initially introduced for the problem of *percolations*, e.g., if we keep each edge of the infinite grid graph with vertex set $\mathbb{Z}^2$ with probability p, what is the probability that the origin is part of an infinite component (in which case we say that there is "percolation"). 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 for all $p > 1/2$. Thus the "bond percolation threshold" for $\mathbb{Z}^2$ is exactly $1/2$. Such exact results are extremely rare.

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 $f(x_1, \ldots, x_n)$ is **monotone increasing** if

$$x \leq y \text{ (coordinatewise)} \quad \Longrightarrow \quad f(x) \leq f(y).$$

Theorem 7.1.3 (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 $g = 1_B$.

Remark 7.1.4. The 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.

Proof. We use induction on n by integrating out the inequality one variable at a time. 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 $\mathbb{E}[fg] \geq \mathbb{E}[f]\mathbb{E}[g]$ (this is sometimes called Chebyshev's inequality/rearrangement inequality).

Now assume $n \geq 2$. Let $h = fg$. Define marginals $f_1, g_1, h_1 \colon \Omega_1 \to \mathbb{R}$ by

$$\begin{aligned}
f_1(y_1) &= \mathbb{E}[f|x_1 = y_1] = \mathbb{E}_{(x_2,\ldots,x_n)\in\Omega_2\times\cdots\times\Omega_n}[f(y_1, x_2, \ldots, x_n)], \\
g_1(y_1) &= \mathbb{E}[g|x_1 = y_1] = \mathbb{E}_{(x_2,\ldots,x_n)\in\Omega_2\times\cdots\times\Omega_n}[g(y_1, x_2, \ldots, x_n)], \\
h_1(y_1) &= \mathbb{E}[h|x_1 = y_1] = \mathbb{E}_{(x_2,\ldots,x_n)\in\Omega_2\times\cdots\times\Omega_n}[h(y_1, x_2, \ldots, x_n)],
\end{aligned}$$

Then f_1 and g_1 are 1-variable monotone increasing functions on Ω_1 (check!).

For every fixed $y_1 \in \Omega_1$, the function $(x_2, \ldots, x_n) \mapsto f(y_1, x_2, \ldots, x_n)$ is monotone increasing, and likewise with g. So applying the induction hypothesis for $n-1$, we have

$$h_1(y_1) \geq f_1(y_1)g_1(y_1). \tag{7.1}$$

Thus

$$\begin{aligned}
\mathbb{E}[fg] = \mathbb{E}[h] &= \mathbb{E}[h_1] \\
&\geq \mathbb{E}[f_1 g_1] && \text{[by (7.1)]} \\
&\geq (\mathbb{E}f_1)(\mathbb{E}g_1) && \text{[by the } n = 1 \text{ case]} \\
&= (\mathbb{E}f)(\mathbb{E}g).
\end{aligned}$$

□

Corollary 7.1.5. 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 $A_1, \ldots, A_k$ 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

$$\mathbb{P}(A\overline{B}) = \mathbb{P}(A) - \mathbb{P}(A\overline{B}) \overset{\text{Harris}}{\leq} \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(\overline{B}) = \mathbb{P}(A)\mathbb{P}(B).$$

The proof of the first inequality is similar. For the last inequality we apply the Harris inequality repeatedly. □

7.2 Applications to random graphs

7.2.1 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 $i, j, k \in [n]$, let A_{ijk} be the event that ijk is a triangle in $G(n,p)$. Then A_{ijk} is increasing, and

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \geq \mathbb{P}\left(\bigwedge_{i<j<k} \overline{A}_{ijk}\right) \geq \prod_{i<j<k} \mathbb{P}(\overline{A}_{ijk}) = (1-p^3)^{\binom{n}{3}}. \qquad \square$$

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

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \geq e^{-\Theta(n^3p^3)}.$$

Here is another lower bound

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \geq \mathbb{P}(G(n,p) \text{ is empty}) = (1-p)^{\binom{n}{2}} = e^{-\Theta(n^2p)}.$$

The bound from Harris is better when $p \ll n^{-1/2}$. Putting them together, we obtain

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \gtrsim \begin{cases} e^{-\Theta(n^3p^3)} & \text{if } p \lesssim n^{-1/2} \\ e^{-\Theta(n^2p)} & \text{if } n^{-1/2} \lesssim p \leq 0.99 \end{cases}$$

(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.

7.2.2 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)**.**

$$\mathbb{P}(\operatorname{maxdeg} G(n,1/2) \leq n/2) = (0.6102\cdots + o(1))^n$$

Instead of giving a proof, we consider an easier continuous model of the problem that motivates the numerical answer. Turning this continuous model paper into a rigorous proof about random graphs is more technical.

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 Z_{uv} to each edge uv of K_n.

Let $W_v = \sum_{u \neq v} Z_{uv}$, which models how much the degree of vertex v deviates from its expectation. In particular W_v is symmetric and mean 0, and $\mathbb{P}(W_v \leq 0)$.

The problem of estimating the probability that $\operatorname{maxdeg} G(n, 1/2) \leq n/2$ then should be modeled as

$$\mathbb{P}(\max_{v \in [n]} W_v \leq 0)$$

(Of course, other than intuition, there is no justification here that these two models actually mimic each other.)

Observe that $(W_v)_{v \in [n]}$ is a joint normal distribution, each coordinate has variance $n-1$ and pairwise covariance 1. So $(W_v)_{v \in [n]}$ has the same distribution as

$$\sqrt{n-2}(Z'_1, Z'_2, \ldots, Z'_n) + Z'_0(1, 1, \ldots, 1)$$

where $Z'_0, \ldots, Z'_n$ are iid standard normals.

Let Φ be the pdf and cdf of the standard normal $N(0, 1)$.

Thus

$$\mathbb{P}(\max_{v \in [n]} W_v \leq 0) = \mathbb{P}\left(\max_{i \in [n]} Z'_i \leq -\frac{Z'_0}{\sqrt{n-2}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \Phi\left(\frac{-z}{\sqrt{n-2}}\right)^n dz$$

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

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\max_{v\in[n]} W_v \leq 0) = \lim_{n\to\infty} \frac{1}{n} \log \int e^{nf(y)}\, dy = \max g = \log 0.6102\cdots .$$

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.

8.1 Probability of non-existence

Question 8.1.1. What is the probability that $G(n,p)$ is triangle-free?

As in indicated in the previous chapter, Janson inequalities will allow us upper bound such probabilities.

The following setup should be a reminiscent of both the second moment method as well as Lovász local lemma (the random variable model).

Setup 8.1.2. 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 events that occur. 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)

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

(note that (i,j) and (j,i) is each counted once).

The following inequality was proved by Janson, Łuczak, and Ruciński (1990).

Theorem 8.1.3 (Janson inequality I)**.** Assuming Setup 8.1.2,

$$\mathbb{P}(X = 0) \leq e^{-\mu + \Delta/2}.$$

Remark 8.1.4. When $\mathbb{P}(A_i) = o(1)$, Harris inequality gives us

$$\mathbb{P}(X=0) = \mathbb{P}(\overline{A}_1 \cdots \overline{A}_k) \geq \mathbb{P}(\overline{A}_1) \cdots \mathbb{P}(\overline{A}_k) = \prod_{i=1}^{k} (1 - \mathbb{P}(A_i)) = e^{-(1+o(1))\sum_{i=1}^{k} \mathbb{P}(A_i)} = e^{-(1+o(1))\mu}.$$

If furthermore $\Delta = o(\mu)$, then two bounds match to give $\mathbb{P}(X=0) = e^{-(1+o(1)\mu}$.

(Not Janson's original proof, which was by analytic interpolation. The following proof is by Boppana and Spencer (1989), with a modification by Warnke[1]. It has some similarities to the proof of Lovász local lemma)

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. Let

$$D_0 = \bigwedge_{j<i : j \not\sim i} \overline{A}_j \qquad \text{and} \qquad D_1 = \bigwedge_{j<i : j \sim i} \overline{A}_j$$

[1] Personal communication

Then

$$\begin{aligned} r_i &= \mathbb{P}(A_i|\overline{A}_1 \cdots \overline{A}_{i-1}) = \mathbb{P}(A_i|D_0D_1) = \frac{\mathbb{P}(A_iD_0D_1)}{\mathbb{P}(D_0D_1)} \\ &\geq \frac{\mathbb{P}(A_iD_0D_1)}{\mathbb{P}(D_0)} \\ &= \mathbb{P}(A_iD_1|D_0) \\ &= \mathbb{P}(A_i|D_0) - \mathbb{P}(A_i\overline{D}_1|D_0) \\ &= \mathbb{P}(A_i) - \mathbb{P}(A_i\overline{D}_1|D_0) \quad \text{[by independence]} \end{aligned}$$

Since A_i and $\overline{D}_1$ are both increasing events, and D_0 is a decreasing event, by Harris inequality (Corollary 7.1.5),

$$\mathbb{P}(A_i\overline{D}_1|D_0) \leq \mathbb{P}(A_i\overline{D}_1) = \mathbb{P}\left(A_i \wedge \bigvee_{j<i:j\sim i} A_j\right) \leq \sum_{j<i:j\sim i} \mathbb{P}(A_iA_j)$$

And the claim follows. □

In Setup 8.1.2 (as well as subsequent Janson inequalities by extension), one can actually allow A_i to be any increasing events, not simply events of the form $S_i \subseteq R$ (known as "principal up-sets").

Theorem 8.1.5 (Riordan and Warnke 2015)**.** Theorem 8.1.3 remains true if Setup 8.1.2 is modified as follows. The events A_i are allowed to any increasing events independent boolean random variables. We write $i \sim j$ if A_i and A_j are not independent (this is initially a pairwise condition, though see lemma below).

In most applications of Janson inequalities, it is easiest to work with principal up-sets. Note that Janson's inequality is false for general events.

Here to how to modify the above proof for work for arbitrary increasing events A_i. The only place we used independence is the "by independence" step above. The next statement shows that the this step remains valid for general up-sets.

Proposition 8.1.6. Let A and $B_1, \ldots, B_k$ be increasing events of independent boolean random variables. If A is independent of B_i for every $i \in [k]$, then A is independent of $\{B_1, \ldots, B_k\}$.

Proof. We first prove the statement for $k = 2$. Writing $B = B_1$ and $C = B_2$, we have

$$\mathbb{P}(A \cap (B \cap C)) + \mathbb{P}(A \cap (B \cup C)) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap C) = \mathbb{P}(A)(\mathbb{P}(B) + \mathbb{P}(C))$$

By Harris inequality, since $B \cap C$ and $B \cup C$ are increasing,

$$\mathbb{P}(A \cap (B \cap C)) \geq \mathbb{P}(A)\mathbb{P}(B \cap C) \quad \text{and} \quad \mathbb{P}(A \cap (B \cup C)) \geq \mathbb{P}(A)\mathbb{P}(B \cup C)$$

Summing the above two gives the previous equality, so the above two inequalities must be equalities. In particular, A is independent of $B \cap C$.

Since the intersection of two up-sets is an up-set, we see that A is independent of the intersection of any subset of $\{B_1, \ldots, B_k\}$, which then implies that A is independent of $\{B_1, \ldots, B_k\}$. □

Now let us return to the probability that $G(n,p)$ is triangle-free. In Setup 8.1.2, let $[N]$ with $N = \binom{n}{2}$ be the set of edges of K_n, and let $S_1, \ldots, S_{\binom{n}{3}}$ be 3-element sets where each S_i is the edge-set of a triangle. As in the second moment calculation in Section 4.1, we have

$$\mu = \binom{n}{3} p^3 \asymp n^3 p^3 \qquad \text{and} \qquad \Delta \asymp n^4 p^5.$$

(where Δ is obtained by considering all appearances of a pair of triangles glued along an edge).

If $p \ll n^{-1/2}$, then $\Delta = o(\mu)$, in which case Janson inequality I (Theorem 8.1.3 and Remark 8.1.4) gives the following.

Theorem 8.1.7. If $p = o(n^{-1/2})$, then

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) = e^{-(1+o(1))\mu} = e^{-(1+o(1))n^3p^3/6}.$$

Corollary 8.1.8. For a constant $c > 0$,

$$\lim_{n\to\infty} \mathbb{P}(G(n, c/n) \text{ is triangle-free}) = e^{-c^3/6}.$$

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

$$\mathbb{P}(G(n,p) \text{ is triangle-free}) \geq (1-p)^{\Theta(n^2)} = e^{-\Theta(n^2p)}$$

(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 8.1.9 (Janson inequality II)**.** Assuming Setup 8.1.2, 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 2.4.2), 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 $X_T := \sum_{i \in T} A_i$ 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 $T \subset [k]$ randomly by including every element with probability $q \in [0,1]$ independently. We have

$$\mathbb{E}\mu_T = q\mu \qquad \text{and} \qquad \mathbb{E}\Delta_T = q^2 \Delta$$

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. □

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 8.1.10. If $\mu \to \infty$ 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, now in the regime $p \gg n^{-1/2}$. We have

$$n^3p^3 \asymp \mu \ll \Delta \asymp n^4p^5.$$

So so for large enough n, Janson inequality II tells us

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \leq e^{-\mu^2/(2\Delta)} = e^{-\Theta(n^2p)}$$

Since

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) \geq \mathbb{P}(G(n, p) \text{ is empty}) \geq (1 - p)^{\binom{n}{2}} = e^{-\Theta(n^2p)}$$

where the final step assumes that p is bounded away from 1, we conclude that

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) = e^{-\Theta(n^2p)}$$

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 8.1.11. Suppose $p = p_n \leq 0.99$. Then

$$\mathbb{P}(G(n, p) \text{ is triangle-free}) = \begin{cases} \exp\left(-\Theta(n^2p)\right) & \text{if } p \gtrsim n^{-1/2} \\ \exp\left(-\Theta(n^3p^3)\right) & \text{if } p \lesssim n^{-1/2} \end{cases}$$

Remark 8.1.12. 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 8.1.2. 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):

$$\mathbb{P}(X = 0) \leq \exp\left(\frac{-\mu^2}{2(\mu+\Delta)}\right). \tag{8.1}$$

(Note that this form of the inequality conveniently captures Janson inequalities I & II.)

Proof. (Lutz Warnke[2]) Let $q \in [0,1]$. Let $T \subset [k]$ where each element is included with probability q independently.

Let $X_T = \sum_{i \in T} 1_{A_i}$. Note that this is the same as $\sum_i 1_{A_i} W_i$ where each $W_i \sim \text{Bernoulli}(q)$.

[2]Personal communication

We have

$$\mathbb{P}(X_T = 0|X) = (1-q)^X$$

Taking expectation and applying Janson inequality I to X_T, we obtain

$$\mathbb{E}[(1-q)^X] = \mathbb{P}(X_T = 0) \leq e^{-\mu' + \Delta'/2} = e^{-q\mu + q^2\Delta/2}$$

where

$$\mu' = q\mu \quad \text{and} \quad \Delta' = q^2\Delta.$$

By Markov's inequality,

$$\begin{aligned}
\mathbb{P}(X \leq \mu - t) &= \mathbb{P}\left((1-q)^X \leq (1-q)^{\mu - t}\right) \\
&\leq (1-q)^{-\mu+t}\mathbb{E}[(1-q)^X] \\
&\leq (1-q)^{-\mu+t}e^{-q\mu+q^2\Delta/2}.
\end{aligned}$$

It remains to show that there is a choice of q so that $RHS \leq \exp\left(\frac{-t^2}{2(\mu+\Delta)}\right)$.

Let $1-q = e^{-\lambda}$, $\lambda \geq 0$. Then

$$\lambda - \frac{\lambda^2}{2} \leq q \leq \lambda$$

So

$$\begin{aligned}
\mathbb{P}(X \leq -\mu + t) &\leq (1-q)^{\mu - t}e^{-q\mu + q^2\Delta/2} \\
&\leq \exp\left(\lambda(\mu - t) - \left(\lambda - \frac{\lambda^2}{2}\right)\mu + \lambda^2\frac{\Delta}{2}\right) \\
&= \exp\left(\lambda t - \frac{\lambda^2}{2}(\mu + \Delta)\right)
\end{aligned}$$

Setting $\lambda = 1/(\mu + \Delta)$ yields the result. □

Example 8.2.3 (Lower tails for triangle counts)**.** Let X be the number of triangles in $G(n,p)$. We have $\mu \asymp n^3p^3$ and $\Delta \asymp n^4p^5$. Fix a constant $\delta \in (0,1]$. Let $t = \delta\mathbb{E}X$. We have

$$\mathbb{P}(X \leq (1-\delta)\mathbb{E}X) \leq \exp\left(-\Theta\left(\frac{-\delta^2 n^6p^6}{n^3p^3 + n^4p^5}\right)\right) = \begin{cases} \exp\left(-\Theta_\delta(n^2p)\right) & \text{if } p \gtrsim n^{-1/2}, \\ \exp\left(-\Theta_\delta(n^3p^3)\right) & \text{if } p \lesssim n^{-1/2}. \end{cases}$$

The bounds are tight up to a constant in the exponent, since

$$\mathbb{P}(X \leq (1-\delta)\mathbb{E}X) \geq \mathbb{P}(X = 0) = \begin{cases} \exp\left(-\Theta(n^2p)\right) & \text{if } p \gtrsim n^{-1/2}, \\ \exp\left(-\Theta(n^3p^3)\right) & \text{if } p \lesssim n^{-1/2}. \end{cases}$$

Example 8.2.4 (No corresponding Janson inequality for upper tails)**.** Continuing with X being the number of triangles of $G(n,p)$, abased on the above lower tails, naively we might expect $\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) \leq \exp(-\Theta_\delta(n^2p))$, but actually this is false!

By planting a clique of size $\Theta(np)$, we can force $X \geq (1+\delta)\mathbb{E}X$. Thus

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) \geq p^{\Theta_\delta(n^2p^2)}$$

which is much bigger than $\exp(-\Theta(n^2p))$. The above is actually the truth (Kahn–DeMarco 2012 and Chatterjee 2012):

$$\mathbb{P}(X \geq (1+\delta)\mathbb{E}X) = p^{\Theta_\delta(n^2p^2)} \quad \text{if } p \gtrsim \frac{\log n}{n},$$

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 2019+)**.** Let X be the number of triangles in $G(n,p)$ with $p = p_n$ satisfying $n^{-1/2} \ll p \ll 1$,

$$-\log \mathbb{P}(X \geq (1+\delta)X) \sim \min\left\{\frac{\delta}{3}, \frac{\delta^{2/3}}{2}\right\} n^2p^2\log(1/p),$$

and for $n^{-1}\log n \ll p \ll n^{-1/2}$,

$$-\log \mathbb{P}(X \geq (1+\delta)X) \sim \frac{\delta^{2/3}}{2} n^2p^2\log(1/p).$$

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 frames that gave the above theorem for sufficiently slowly decaying $p \geq n^{-c}$.

For the corresponding problem for lower tails, 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 Clique and chromatic number of $G(n, 1/2)$

In Section 4.3, 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

$$\omega(G(n, 1/2)) \sim 2 \log_2 n \qquad \text{whp.}$$

Let use recall the proof using the second moment method. Let X denote the number of k-cliques in $G(n, 1/2)$. Then

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

Here $k = k_n$ depends on n.

If $\mu \to 0$, then Markov gives $X = 0$ whp.

If $\mu \to \infty$, then one checks that $\Delta \ll \mu^2$, so that Chebyshev's inequality gives $X > 0$ whp.

Let $k_0 = k_0(n)$ be the largest possible k so that $\mu_k \geq 1$. We have $\mu_{k_0} \geq 1 > \mu_{k_0+1}$ and

$$k_0 \sim 2 \log_2 n.$$

We have

$$\frac{\mu_{k+1}}{\mu_k} = n^{-1+o(1)} \quad \text{for } k \sim 2 \log_2 n$$

Thus $\omega(G(n, 1/2)) \sim 2 \log_2 n$ whp. In fact, this proof gives more, namely that the clique number is concentrated on at most two values

$$\omega(G(n, 1/2)) \in \{k_0 - 1, k_0\} \qquad \text{whp.}$$

Can two point concentration of $\omega(G(n, 1/2))$ really occur? (As opposed to being always concentrated on a single value with high probability.) It turns out that the answer is yes.

Theorem 8.3.1. Fix $\lambda \in (-\infty, \infty)$. Let $n_0(k)$ be the minimum n satisfying $\binom{n}{k}2^{-\binom{k}{2}} \geq 1$. Then, as $k \to \infty$, and for

$$n = n_0(k)\left(1 + \frac{\lambda + o(1)}{k}\right),$$

one has

$$\mathbb{P}(\omega(G(n,1/2)) = k-1) = e^{-e^{\lambda}} + o(1)$$
$$\text{and} \quad \mathbb{P}(\omega(G(n,1/2)) = k) = 1 - e^{-e^{\lambda}} + o(1).$$

Proof. Let X denote the number of k-cliques in $G(n,1/2)$. Using the notation of Setup 8.1.2 for Janson inequalities, one can check that

$$\mu = \binom{n}{k}2^{-\binom{k}{2}} \sim \left(1 + \frac{\lambda + o(1)}{k}\right)^k = e^{\lambda} + o(1)$$

and (details omitted)

$$\Delta \sim \mu^2 \frac{k^4}{n^2} + \mu \frac{2kn}{2^k} = o(1).$$

Then, by Harris inequality (lower bound) and Janson inequality I (upper bound), we have

$$e^{-(1+o(1))\mu} = (1 - 2^{-\binom{k}{2}})^{\binom{n}{k}} \leq \mathbb{P}(X = 0) \leq e^{-\mu + \Delta/2} = e^{-(1+o(1))\mu}.$$

Thus

$$\mathbb{P}(\omega(G(n,1/2)) < k) = \mathbb{P}(X = 0) = e^{-(1+o(1))\mu} = e^{-e^{\lambda}} + o(1).$$

At this point, we can use two-point concentration to conclude. Alternatively, note that $n_0(k) = 2^{(1+o(1))k/2}$, and thus $n = n_0(k-1)(1 + \frac{\lambda'}{k-1})$ for some $\lambda' \to \infty$, and so that the above bound also gives

$$\mathbb{P}(\omega(G(n,1/2)) < k-1) \leq e^{-e^{\lambda'}} + o(1) = o(1).$$

This again proves two-point concentration, and hence the conclusion. □

Thus one has genuine two-point concentration (i.e., with $\mathbb{P}(\omega(G(n,1/2)) = k_0)$ bounded away from 0 and 1) if

$$n = n_0(k)\left(1 + \frac{O(1)}{k}\right)$$

for some k. Noting that $n_0(k) = 2^{(1+o(1))k/2}$. The intervals $[n_0(k)(1 - K/k), n_0(k)(1 + K/k)]$ are disjoint for large enough k. We see that the number of integers n up to N with two-

points concentration is asymptotically

$$\sum_{k:n_0(k)\le N} O\left(\frac{n_0(k)}{k}\right) = O\left(\frac{N}{\log N}\right).$$

Thus for almost all integers we actually have one-point concentration.

The next statement tells us we have an exponentially small probability of having cliques of size $\sim 2\log_2 n$. This estimate will be important in the following theorem where we determine the chromatic number of $G(n, 1/2)$.

Theorem 8.3.2. Let $k_0 = k_0(n)$ be the largest possible k so that $\mu_k := \binom{n}{k}2^{-\binom{k}{2}} \ge 1$. Then

$$\mathbb{P}(\omega(G(n, 1/2)) < k_0 - 3) \le e^{-n^{2-o(1)}}$$

Note that there is a trivial lower bound of $2^{-\binom{n}{2}}$ coming from an empty graph.

Proof. We have $\mu_{k+1}/\mu_k = n^{-1+o(1)}$ whenever $k \sim k_0(n) \sim 2\log_2 n$.

Writing $k = k_0 - 3$ and using the notation of Setup 8.1.2 for Janson inequalities for X being the number of k-cliques, we have

$$\mu = \mu_k > n^{3-o(1)}.$$

One can check that (again details omitted on Δ; the second step uses $2^k = n^{2+o(1)}$),

$$\Delta \sim \mu^2\frac{k^4}{n^2} + \mu\frac{2kn}{2^k} \sim \mu^2\frac{k^4}{n^2}$$

So $\Delta > \mu$ for sufficiently large n, and we can apply Janson inequality II:

$$\mathbb{P}(X = 0) = \mathbb{P}(\omega(G(n, 1/2)) < k) \le e^{-\mu^2/(2\Delta)} < e^{-(1/2+o(1))n^2/k^4} = e^{-\Omega(n^2/(\log n)^4)}. \qquad \square$$

Since $G(n, 1/2)$ and its graph complement are identically distributed, and $\omega(G) = \alpha(\overline{G})$, the independence number α satisfies

$$\alpha(G(n, 1/2)) \sim 2\log_2 n \qquad \text{whp.}$$

It follows that the chromatic number of $G \sim G(n, 1/2)$ satisfies

$$\chi(G) \ge \frac{n}{\alpha(G)} \ge (1 + o(1))\frac{n}{2\log_2 n} \quad \text{whp.}$$

The following landmark remark of Bollobás pins down the asymptotics of the chromatic number of the random graph.

Theorem 8.3.3 (Bollobás 1988)**.** With probability $1 - o(1)$,

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

Proof. The lower bound proof was discussed before the theorem statement. For the upper bound we will give a strategy to properly color the graph with not too many 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. Instead, we will use the previous theorem to tell us that, in $G(n, 1/2)$, with high probability, every not-too-small 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 Theorem 8.3.2, for

$$k = k_0(m) \sim 2 \log_2 m \sim 2 \log_2 n,$$

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,

$$\mathbb{P}(\exists \text{ an } m\text{-vertex subset } S \text{ with } \alpha(G[S]) < k) < 2^n e^{-n^{2-o(1)}} = o(1).$$

Thus, with probability $1 - o(1)$ every m-vertex subset contains a k-vertex independent set. Assume that G has this property. Now we execute our strategy at the beginning of the proof:

- While $\geq m$ vertices remain:
  - Find an independent set of size k, and let it form its own color class
  - Remove these k vertices
- Color the remaining $< m$ vertices each with a new color.

Thus we obtain a proper coloring using at most

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

colors. □

9 Concentration of measure

Recall that Chernoff bound allows to prove exponential tail bounds for sums of **independent** random variables. For example, if Z is a sum of n Bernoulli random variables, then

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t\sqrt{n}) \leq 2e^{-2t^2/n}.$$

As a matter of terminology (which is convenient though we will largely not use), random variables Z that satisfy $\mathbb{P}(|Z| \geq t) \leq 2e^{-ct^2}$ for all $t \geq 0$ and constant $c > 0$ are called **sub-gaussian**. We usually are not too concerned about optimizing the constant c in the exponent of bound.

In this chapter, we develop tools for proving similar sub-gaussian tail bounds for other random variables that do not necessarily arise as a sum of independent random variables.

Here is the general principle:

> **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 9.0.1 (Bounded differences inequality)**.** Let $X_1 \in \Omega_1, \ldots, X_n \in \Omega_n$ be **independent** random variables. Suppose $f \colon \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ satisfies

$$|f(x_1, \ldots, x_n) - f(x_1', \ldots, x_n')| \leq 1 \tag{9.1}$$

whenever $(x_1, \ldots, x_n)$ and $(x_1', \ldots, x_n')$ differ on exactly one coordinate. Then the random variable $Z = f(X_1, \ldots, X_n)$ 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 $f(x_1, \ldots, x_n) = x_1 + \cdots + x_n$ to recover the Chernoff bound. The theorem tells us that the window of fluctuation of Z has length $O(\sqrt{n})$.

Example 9.0.2 (Coupon collector)**.** Let $s_1, \ldots, s_n \in [n]$ chosen uniformly and independently at random. Let

$$Z = |[n] \setminus \{s_1, \ldots, s_n\}|.$$

Then

$$\mathbb{E}Z = n\left(1-\frac{1}{n}\right)^n \in \left[\frac{n}{e}, \frac{n-1}{e}\right].$$

Note that changing one of the $s_1, \dots, s_n$ changes Z by at most 1, so we have

$$\mathbb{P}\left(|Z - n/e| \geq \lambda\sqrt{n}+1\right) \leq \mathbb{P}\left(|Z-\mathbb{E}Z| \geq \lambda\sqrt{n}\right) \leq 2e^{-2\lambda^2}.$$

Definition 9.0.3 (Lipschitz functions)**.** Given two metric spaces (X, d_X) and (Y, d_Y), we say that a function $f\colon X \to Y$ is **C-Lipschitz** if

$$d_Y(f(x), f(x')) \leq C d_X(x, x') \qquad \text{for all } x, x' \in X.$$

Then (9.2) says that $f\colon \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ is 1-Lipschitz with respect to the Hamming distance on $\Omega_1 \times \cdots \times \Omega_n$.

Note that while it may be tempting to think about the cases $\Omega_i = \{0, 1\}$, it will be crucial for us to consider more general Ω_i for our applications.

Theorem 9.0.1 holds more generally allowing the bounded difference to depend on the coordinate.

Theorem 9.0.4 (Bounded differences inequality)**.** Let $X_1 \in \Omega_1, \dots, X_n \in \Omega_n$ be **independent** random variables. Suppose $f\colon \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ satisfies

$$|f(x_1, \dots, x_n) - f(x_1', \dots, x_n')| \leq c_i \tag{9.2}$$

whenever $(x_1, \dots, x_n)$ and $(x_1, \dots, x_n)$ differ only on the i-th coordinate. Then the random variable $Z = f(X_1, \dots, X_n)$ satisfies, for every $\lambda \geq 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq \lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right)$$

and

$$\mathbb{P}(Z - \mathbb{E}Z \leq -\lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

We will prove these inequality using martingales.

9.1 Martingales concentration inequalities

Definition 9.1.1. A **martingale** is a random real sequence $Z_0, Z_1, \ldots$ such that for every Z_n, $\mathbb{E}|Z_n| < \infty$ and

$$\mathbb{E}[Z_{n+1}|Z_0, \ldots, Z_n] = Z_n.$$

(To be more formal, we should talk about filtrations of a probability space ...)

Example 9.1.2 (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 9.1.3 (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 9.1.4 (Doob martingale)**.** Given some underlying random variables $X_1, \ldots, X_n$ (not necessarily independent, though they often are independent in practice), and a function $f(X_1, \ldots, X_n)$. Let Z_i be the expected value of f after "revealing" (exposing) $X_1, \ldots, X_i$, i.e.,

$$Z_i = \mathbb{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_i].$$

So Z_i is the expected value of the random variable $Z = f(X_1, \ldots, X_n)$ after seeing the first i arguments, and letting the remaining arguments be random. Then $Z_0, \ldots, Z_n$ 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 9.1.5 (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 $X_0, X_1, \ldots, X_{\binom{n}{2}}$ 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 5 for an example.

Example 9.1.6 (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

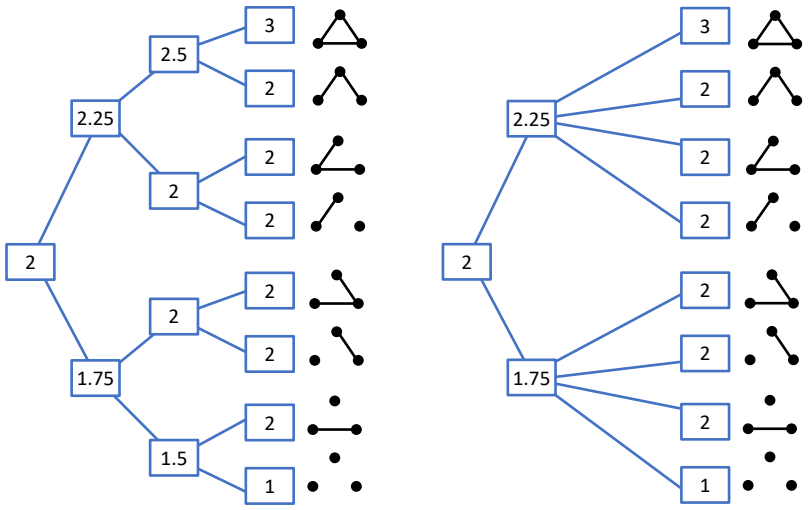


Figure 5: The edge-exposure martingale (left) and vertex-exposure martingale (right) for the chromatic number of $G(n, 1/2)$ with $n = 3$. The martingale is obtained by starting at the leftmost point, and splitting at each branch with equal probability.

reveal all edges whose endpoints are contained in the first i vertices. See Figure 5 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.

Theorem 9.1.7 (Azuma's inequality)**.** Let $Z_0, Z_1, \ldots, Z_n$ 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 Section 5 for $Z_n = X_1 + \cdots X_n$ 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 9.1.8 (Azuma's inequality)**.** Let $Z_0, Z_1, \ldots, Z_n$ be a martingale satisfying

$$|Z_i - Z_{i-1}| \le c_i \quad \text{for each } i \in [n].$$

For any $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \ge \lambda) \le \exp\left(\frac{-\lambda^2}{2(c_1^2 + \cdots + c_n^2)}\right).$$

The above formulations of Azuma's inequality recovers the bounded differences inequality Theorems 9.0.1 and 9.0.4 up to a (usually unimportant) constant in the exponent (details shortly). To obtain the exact statement of Theorem 9.0.4, we state the following strengthening of Azuma's inequality.

Theorem 9.1.9 (Azuma's inequality)**.** Let $Z_0, Z_1, \ldots, Z_n$ be a martingale such that, for each $i \in [n]$, conditioned on $(Z_0, \ldots, Z_{i-1})$, the random variable Z_i lies inside an interval of length c_i (the location of the interval may depend on $Z_0, \ldots, Z_{i-1}$). Then for any $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \ge \lambda) \le \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

Remark 9.1.10. Applying the inequality to the martingale with terms $-Z_n$, we obtain the following lower tail bound:

$$\mathbb{P}(Z_n - Z_0 \le -\lambda) \le \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

And we can put them together as

$$\mathbb{P}(|Z_n - Z_0| \ge \lambda) \le 2\exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

Lemma 9.1.11 (Hoeffding)**.** Let X be a real random variable contained in an interval of length ℓ. Suppose $\mathbb{E}X = 0$. Then

$$\mathbb{E}[e^X] \le e^{\ell^2/8}.$$

Proof. Suppose $X \in [a, b]$ with $a \le 0 \le b$ and $b - a = \ell$. Then since e^x is convex, using a linear upper bound on the interval $[a, b]$, we have

$$e^x \le \frac{b - x}{b - a}e^a + \frac{x - a}{b - a}e^b, \quad \forall x \in [a, b].$$

Thus

$$\mathbb{E}e^X \le \frac{b}{b-a}e^a + \frac{-a}{b-a}e^b.$$

Let $p = -a/(b-a)$. then $a = -p\ell$ and $b = (1-p)\ell$, we have

$$\log \mathbb{E}e^X \le \log\left((1-p)e^{-p\ell} + pe^{(1-p)\ell}\right) = -p\ell + \log(1-p+pe^{\ell}).$$

Fix $p \in [0,1]$. Let

$$\varphi(\ell) := -p\ell + \log(1-p+pe^{\ell}).$$

It remains to show that $\varphi(\ell) \le \ell^2/8$ for all $\ell \ge 0$, which follows from $\varphi(0) = \varphi'(0) = 0$ and $\varphi''(\ell) \le 1/4$ for all $\ell \ge 0$, as

$$\varphi''(\ell) = \left(\frac{p}{(1-p)e^{-p\ell}+p}\right)\left(1 - \frac{p}{(1-p)e^{-p\ell}+p}\right) \le \frac{1}{4},$$

since $t(1-t) \le 1/4$ for all $t \in [0,1]$. □

Proof of Theorem 9.1.9. By adding a constant to the sequence, we may assume that $Z_0 = 0$. Let

$$X_i = Z_i - Z_{i-1}$$

be the martingale difference. Let $t \ge 0$. Then the hypothesis together with Lemma 9.1.11 imply that

$$\mathbb{E}[e^{tX_i} | Z_0, \dots, Z_{i-1}] \le e^{t^2 c_i^2/8}.$$

Then the moment generating function satisfies

$$\mathbb{E}[e^{tZ_n}] = \mathbb{E}\left[e^{t(X_n+Z_{n-1})}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{tX_n} \;\middle|\; Z_0, \dots, Z_{n-1}\right] e^{tZ_{n-1}}\right] = e^{t^2c_n^2/8}\mathbb{E}[e^{tZ_{n-1}}].$$

Iterating, we obtain

$$\mathbb{E}[e^{tZ_n}] \le e^{t^2(c_1^2+\cdots c_n^2)/8}.$$

By Markov,

$$\mathbb{P}(Z_n \ge \lambda) \le e^{-t\lambda}\mathbb{E}[e^{tZ_n}] \le e^{-t\lambda + \frac{t^2}{8}(c_1^2+\cdots c_n^2)}.$$

Setting $t = 4\lambda/(c_1^2 + \cdots + c_n^2)$ yields the theorem. □

Let us use Azuma's inequality to prove the bounded difference inequality (Theorem 9.0.4), whose statement is copied below:

> Let $Z_0, Z_1, \dots, Z_n$ be a martingale such that, for each $i \in [n]$, conditioned on $(Z_0, \dots, Z_{i-1})$, the random variable Z_i lies inside an interval of length c_i (the

location of the interval may depend on $Z_0, \dots, Z_{i-1}$). Then for any $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

Proof of the Theorem 9.0.4. Consider the Doob martingale $Z_i = \mathbb{E}[Z|X_1, \dots, X_i]$.

By the Lipschitz condition, we see that for every $i \in [n]$ and fixed $x_1 \in \Omega_1, \dots, x_{i-1} \in \Omega_{i-1}$, we have

$$\max_{x_i \in \Omega_i} f(x_1, \dots, x_{i-1}, x_i, X_{i+1}, \dots, X_n) - \min_{x_i \in \Omega_i} f(x_1, \dots, x_{i-1}, x_i, X_{i+1}, \dots, X_n) \leq c_i$$

for every possible $X_{i+1}, \dots, X_n$, so that taking expectation of these random values shows that, conditioned on the values of $X_1, \dots, X_{i-1}$, there is an interval (possibly depending on $X_1, \dots, X_{i-1}$) of length c_i that Z_i lies in.

Since $Z_0 = \mathbb{E}Z$ and $Z_n = Z$, the desired bound follows from Azuma's inequality (Theorem 9.1.9).[3] □

9.2 Chromatic number of random graphs

9.2.1 Concentration of chromatic number

Even before Bollobás (1988) showed that $\chi(G(n, 1/2)) \sim \frac{n}{2\log_2 n}$ whp (Theorem 8.3.3), using the bounded difference inequality, it was already known that the chromatic number of a random graph must be concentrated in a $\omega(\sqrt{n})$ window around its mean. The following application shows that one can prove concentration around the mean without even knowing where is the mean!

Theorem 9.2.1 (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 $\Omega_2 \times \cdots \times \Omega_n$ where $\Omega_i = \{0, 1\}^{i-1}$ and its coordiantes correspond to edges whose larger coordinate is i (cf. the vertex-exposure martingale Example 9.1.6). 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 9.0.4 implies the result. □

[3] We are cheating somewhat here, since multiple instance of $(X_1, \dots, X_i)$ can correspond to the same $(Z_0, \dots, Z_i)$. To be more correct, we should restate Theorem 9.1.9 instead of a filtration based on the Doob martingale.

Remark 9.2.2 (Non-concentration of the chromatic number). Recently, a surprising breakthrough of Heckel (2019+) showed that the $\chi(G(n,1/2))$ is *not* concentrated on any interval of length $n^{1/4-\epsilon}$ for any constant $\epsilon > 0$. This was the opposite of what most experts believed in. Given the new realization, it seems reasonable to suspect that the length of the window of concentrations fluctuates between $n^{1/4+o(1)}$ to $n^{1/2+o(1)}$ depending on n.

9.2.2 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 8.3.3) 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).

Theorem 9.2.3 (Same as Theorem 8.3.2)**.** 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. Here, instead, we use a very clever and non-obvious choice of a Lipschitz function of graphs.

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 $G \sim G(n,1/2)$,

$$\mathbb{P}(\omega(G) < k) = \mathbb{P}(Y = 0) \leq \mathbb{P}(Y - \mathbb{E}Y \leq -\mathbb{E}Y) \leq \exp\left(-\frac{2(\mathbb{E}Y)^2}{\binom{n}{2}}\right). \tag{9.3}$$

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 $\overline{d}$, we have

$$\alpha(H) \geq \sum_{v \in V(H)} \frac{1}{1+d_v} \geq \frac{|V(H)|}{1+\overline{d}} = \frac{|V(H)|^2}{|V(H)| + 2\,|E(H)|}.$$

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

$$\alpha(H) \geq q\mathbb{E}\,|V(H)| - q^2\,|E(H)| \qquad \text{for every } q \in [0,1].$$

Provided that $|E(H)| \geq |V(H)|/2$, we can take $q = |V(H)|/(2\,|E(H)|) \in [0,1]$ and obtain

$$\alpha(H) \geq \frac{|V(H)|^2}{4\,|E(H)|} \qquad \text{if } |E(H)| \geq \frac{1}{2}\,|V(H)|\,.$$

(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.)

We have, with probability $1 - o(1)$, the number of k-cliques $|V(\mathcal{H})|$ satisfies

$$|V(\mathcal{H})| \sim \mu := \mathbb{E}\,|V(\mathcal{H})| = \binom{n}{k} 2^{-\binom{k}{2}} \geq n^{3-o(1)}$$

and the number of pairs of edge-overlapping k-cliques $|E(\mathcal{H})|$ satisfies

$$\mathbb{E}\,|E(\mathcal{H})| =: \frac{\Delta}{2} \sim \frac{\mu^2 k^4}{2n^2} \gg \mu$$

(details again omitted; this is the same first and second moment calculation as in Section 4.3 and Theorem 8.3.2.) Thus, with probability $1 - o(1)$, we can apply either of the above lower bounds on independent sets to obtain

$$\mathbb{E}Y \gtrsim \mathbb{E}\frac{\mu^2}{|E(\mathcal{H})|} \gtrsim \frac{\mu^2}{\Delta} \sim \frac{n^2}{k^4}.$$

Thus by (9.3), we obtain

$$\mathbb{P}(\omega(G) < k) \le \exp\left(-\frac{2(\mathbb{E}Y)^2}{\binom{n}{2}}\right) \le \exp\left(-\Omega\left(\frac{n^2}{k^8}\right)\right) = \exp\left(-\Omega\left(\frac{n^2}{(\log n)^8}\right)\right). \quad \square$$

9.2.3 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 9.2.4 (Shamir and Spencer 1987)**.** Let $\alpha > 5/6$ be fixed. Then for $p < n^{-\alpha}$, $\chi(G(n,p))$ is concentrated in four values with probability $1 - o(1)$, i.e., there exists $u = u(n,p)$ such that, as $n \to \infty$,

$$\mathbb{P}(u \le \chi(G(n,p)) \le u + 3) = 1 - o(1).$$

Proof. Let $\epsilon = \epsilon_n > 0$ and $\epsilon \to 0$ (we'll later choose it to be arbitrarily small). Let $u = u(n,p,\epsilon)$ be the least integer so that

$$\mathbb{P}(\chi(G(n,p)) \le u) > \epsilon.$$

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 $S \subset V(G)$ 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 9.0.1 with respect to vertex-exposure (Example 9.1.6), we have

$$\begin{aligned} &\mathbb{P}(Y \le \mathbb{E}Y - \lambda\sqrt{n}) \le e^{-2\lambda^2} \\ \text{and} \quad &\mathbb{P}(Y \ge \mathbb{E}Y + \lambda\sqrt{n}) \le e^{-2\lambda^2}. \end{aligned}$$

We choose $\lambda = \lambda(\epsilon) > 0$ so that $e^{-2\lambda^2} = \epsilon$.

First, we use the lower tail bound to show that $\mathbb{E}Y$ must be small. We have

$$e^{-2\lambda^2} = \epsilon < \mathbb{P}(\chi(G) \le u) = \mathbb{P}(Y = 0) = \mathbb{P}(Y \le \mathbb{E}Y - \mathbb{E}Y) \le \exp\left(\frac{-2(\mathbb{E}Y)^2}{n}\right)$$

so

$$\mathbb{E}Y \le \lambda\sqrt{n}.$$

Next, we apply the upper tail bound to show that Y is rarely large. We have

$$\mathbb{P}(Y \geq 2\lambda\sqrt{n}) \leq \mathbb{P}(Y \geq \mathbb{E}Y + \lambda\sqrt{n}) \leq e^{-2\lambda^2} = \epsilon.$$

Each of the following three events occur with probability at least $1 - \epsilon$, for large enough n,

- By the above argument, there is some $S \subset V(G)$ 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 $1 - 3\epsilon$, all three events occur, and so we have $u \leq \chi(G) \leq u + 3$. □

Lemma 9.2.5. 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 $T \subset V(G)$ 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 $\binom{n}{k} \leq (ne/k)^k$)

$$\leq \sum_{t=4}^{C\sqrt{n}} \binom{n}{t} \binom{\binom{t}{2}}{3t/2} p^{3t/2} \leq \sum_{t=4}^{C\sqrt{n}} \left(\frac{ne}{t}\right)^t \left(\frac{te}{3}\right)^{3t/2} n^{-3t\alpha/2}$$

$$\leq \sum_{t=4}^{C\sqrt{n}} \left(O(n^{1-3\alpha/2}\sqrt{t})\right)^t \leq \sum_{t=4}^{C\sqrt{n}} \left(O(n^{1-3\alpha/2+1/4})\right)^t$$

the sum is $o(1)$ provided that $\alpha > 5/6$. □

Remark 9.2.6. Theorem 9.2.4 was subsequently improved (by a refinement of the above techniques) by Łuczak (1991) and Alon and Krivelevich (1997), who showed two-point concentration for all $\alpha > 1/2$.

9.3 Isoperimetric inequalities: a geometric perspective

The bounded differences inequality (Theorem 9.0.1) tells that if $f\colon \{0,1\}^n \to \mathbb{R}$ is 1-Lipschitz (with respect to the Hamming distance on $\{0,1\}^n$), it must be concentrated around its mean:

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

Given that the maximum possible variation in f is n, the above concentration inequality says that f is *almost constant*, which should be somewhat counterintuitive.

It turns out that similar phenomenon occurs in other spaces not just the Hamming cube. In fact, it is really a general high dimensional geometric phenomenon. In this section, we explore this concentration of phenomenon from a geometric perspective, and explain how it relates to **isoperimetric inequalities**.

Recall the classic isoperimetric theorem in $\mathbb{R}^n$ It 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.)

Here is a slightly stronger formulation. Given a metric space (X, d_X) and a set $A \subset X$, we write

$$A_t := \{x \in X : d_X(x, A) := \min_{a\in A} d_X(x, a) \leq t\} \tag{9.4}$$

for set of all points within distance t from A. One can visualize by "expanding" A by distance t.

Theorem 9.3.1 (Isoperimetric inequality in Euclidean space)**.** Let $A \subset \mathbb{R}^n$ be a measurable set, and let $B \subset \mathbb{R}^n$ be a ball $\mathrm{vol}(A) = \mathrm{vol}(B)$. Then, for all $t \geq 0$,

$$\mathrm{vol}(A_t) \geq \mathrm{vol}(B_t).$$

Remark 9.3.2. One can recover the classic inequality on surface volumes $\mathrm{vol}_{n-1}(\delta A) \geq \mathrm{vol}_{n-1}(\delta B)$ by noting that

$$\mathrm{vol}_{n-1}(\delta A) = \frac{d}{dt}\Big|_{t=0} \mathrm{vol}_n(A_t). \lim_{t\to 0} \frac{\mathrm{vol}(A_t) - \mathrm{vol}(A)}{t} \geq \lim_{t\to 0} \frac{\mathrm{vol}(B_t) - \mathrm{vol}(B)}{t} = \mathrm{vol}_{n-1}(\delta B).$$

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 $A \subset \{0,1\}^n$, we write A_t as in (9.4) for $X = \{0,1\}^n$ and d_X being the Hamming distance.

Theorem 9.3.3 (Isoperimetic inequality in the Hamming cube; Harper 1966)**.** Let $A \subset \{0,1\}^n$. Let $B \subset \{0,1\}^n$ be a Hamming ball with $|A| \geq |B|$. Then for all $t \geq 0$,

$$|A_t| \geq |B_t|.$$

Remark 9.3.4. 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.

It is worth examining the sizes of the Hamming ball as a function of its radius.

Let

$$B(r) = \{x \in \{0,1\}^n : \text{weight}(x) \leq r\}$$

denote the Hamming ball of radius r. Using the central limit theorem, we find that, for every fixed $z \in \mathbb{R}$, as $n \to \infty$.

$$\frac{1}{2^n}\left|B\left(\frac{n}{2} + \frac{z\sqrt{n}}{2}\right)\right| = \frac{1}{2^n} \sum_{0 \leq i \leq \frac{n}{2} + \frac{z\sqrt{n}}{2}} \binom{n}{i} \sim \mathbb{P}_{Z \sim N(0,1)}(Z \leq t) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-x^2/2}\, dx.$$

Also, by Chernoff bound, we have

$$\frac{1}{2^n}\left|B\left(\frac{n}{2} + \frac{z\sqrt{n}}{2}\right)\right| \leq e^{-z^2/2} \qquad \text{if } z \leq 0$$

and

$$\frac{1}{2^n}\left|B\left(\frac{n}{2} + \frac{z\sqrt{n}}{2}\right)\right| \geq 1 - e^{-z^2/2} \qquad \text{if } z \geq 0.$$

Combined with the isoperimetic 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 9.3.5. Let $t > 0$. For every $A \subset \{0,1\}^n$ 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 9.3.3),

$$|A_t| \geq |B_t| = |\{x \in \{0,1\}^n : \text{weight}(x) < n/2 + t\}| > (1 - e^{-2t^2/n})2^n$$

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 9.3.6. Let $\epsilon > 0$. If $A \subset \{0,1\}^n$ with $|A| \geq \epsilon 2^n$, then

$$\left|A_{\sqrt{2\log(1/\epsilon)n}}\right| \geq (1-\epsilon)2^n.$$

First proof via isoperimetric inequality. Let $t = \sqrt{\log(1/\epsilon)n/2}$ so that $e^{-2t^2/n} = \epsilon$. Applying Theorem 9.3.5 to $A' = \{0,1\}^n \setminus A_t$, we see that $|A'| < 2^{n-1}$ (or else $|A'_t| > (1-\epsilon)2^n$, 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 9.3.5 yields $|A_{2t}| \geq (1-\epsilon)2^n$. □

Let us give another proof of Theorem 9.3.6 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 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 9.0.1, we have the lower tail

$$\mathbb{P}(X - \mathbb{E}X \leq -t) \leq e^{-2t^2/n}.$$

We have $X = 0$ if and only if $x \in A$, so

$$\epsilon \leq \mathbb{P}(x \in A) = \mathbb{P}(X - \mathbb{E}X \leq -\mathbb{E}X) \leq e^{-2(\mathbb{E}X)^2/n}.$$

Thus

$$\mathbb{E}X \leq \sqrt{\frac{\log(1/\epsilon)n}{2}}.$$

Now we apply the upper tail

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq e^{-2t^2/n}$$

with

$$t = \sqrt{2(\log(1/\epsilon)n} \geq 2\mathbb{E}X$$

to yield

$$\mathbb{P}(x \notin A_t) = \mathbb{P}(X > t) < \mathbb{P}\left(X \geq \mathbb{E}X + \sqrt{\frac{\log(1/\epsilon)n}{2}}\right) \leq \epsilon. \qquad \square$$

The above expansion/isoperimetry properties turn out to be actually equivalent to the concentration of Lipschitz function phenomenon we discussed earlier, as we show next. Milman recognized the importance of this **concentration of measure phenomenon**, which he heavily promoted in the 1970's. The subject was have 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.

Theorem 9.3.7 (Equivalence between notions of concentration of measure)**.** Let $t, \epsilon \geq 0$. In a probability space $(\Omega, \mathbb{P})$ equipped with a metric. The following are equivalent:

1. (Expansion/approximate isoperimetry) If $A \subset \Omega$ with $\mathbb{P}(A) \geq 1/2$, then

$$\mathbb{P}(A_t) \geq 1 - \epsilon.$$

2. (Concentration of Lipschitz functions) If $f \colon \Omega \to \mathbb{R}$ is 1-Lipschitz and $m \in \mathbb{R}$ satisfies $\mathbb{P}(f \geq m) \geq 1/2$ and $\mathbb{P}(f \leq m) \geq 1/2$ (i.e., m is a **median** of f), then

$$\mathbb{P}(f > m + t) \leq \epsilon.$$

Remark 9.3.8. There always exists a median, but it might not be unique. For example, for the uniform distribution on $\{0, 1\}$, any real number in the interval $[0, 1]$ is a valid median.

Proof. (a) $\Longrightarrow$ (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)

$$\mathbb{P}(f > m + \epsilon) \leq \mathbb{P}(\overline{A_t}) \leq \epsilon.$$

(b) $\Longrightarrow$ (a): Let $f(x) = \text{distance}(x, A)$ and $m = 0$. Since $\mathbb{P}(f \leq 0) = \mathbb{P}(A) \geq 1/2$ and $\mathbb{P}(f \geq 0) = 1$, m is a median. Also f is 1-Lipschitz. So by (b),

$$\mathbb{P}(\overline{A_t}) = \mathbb{P}(f > m + t) \leq \epsilon. \qquad \square$$

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, postively-curved manifolds, and the symmetric group.

Mean versus median. For a sub-gaussian random variable, very tight concentration (e.g., sub-gaussian), one can deduce that the mean and the median must be very close to each other.

Indeed, suppose there exist constants $C, \sigma > 0$ such that $\mathbb{P}(A_t) \leq Ce^{-(t/\sigma)^2}$ for all A with $\mathbb{P}(A) \geq 1/2$ and $t > 0$. Then for all 1-Lipschitz function f on Ω and m a median of f, one has

$$|\mathbb{E}f - m| \leq \mathbb{E}|f - m| = \int_0^\infty \mathbb{P}(|f - m| \geq t)\, dt \leq \int_0^\infty 2Ce^{-(t/\sigma)^2}\, dt = C\sqrt{\pi}\sigma$$

It follows that, for all $t \geq 0$,

$$\mathbb{P}(f \geq \mathbb{E}f + (t + C\sqrt{\pi})\sigma) \geq \mathbb{P}(f \geq m + t\sigma) \leq Ce^{-(t/\sigma)^2}$$

and

$$\mathbb{P}(f \leq \mathbb{E}f - (t + C\sqrt{\pi})\sigma) \geq \mathbb{P}(f \leq m - t\sigma) \leq Ce^{-(t/\sigma)^2}.$$

Similarly, if we know that $\mathbb{P}(|f - \mathbb{E}f| \geq t) \leq Ce^{-(t/\sigma)}$ for all $t > 0$, then $\mathbb{P}(|f - \mathbb{E}f| \geq t) < 1/2$ for all $t > \sqrt{\log(2C)}\sigma$, from which we deduce that every median m satisfies $|\mathbb{E}f - m| \leq \sqrt{\log(2C)}\sigma$.

There can indeed exist an order σ difference between the mean and the median in the setup above. For example, treating the cube as $\{-1, 1\}^n$, and taking

$$f(x_1, \ldots, x_n) = \max\{x_1 + \cdots + x_n, 0\},$$

we see that by the central limit theorem

$$\lim_{n\to\infty} \frac{|\mathbb{E}f - \text{median}(f)|}{\sqrt{n}} = \mathbb{E}_{Z\sim N(0,1)}[\max\{Z, 0\}] = \frac{1}{\sqrt{2\pi}}.$$

9.3.1 The sphere and Gauss space

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:

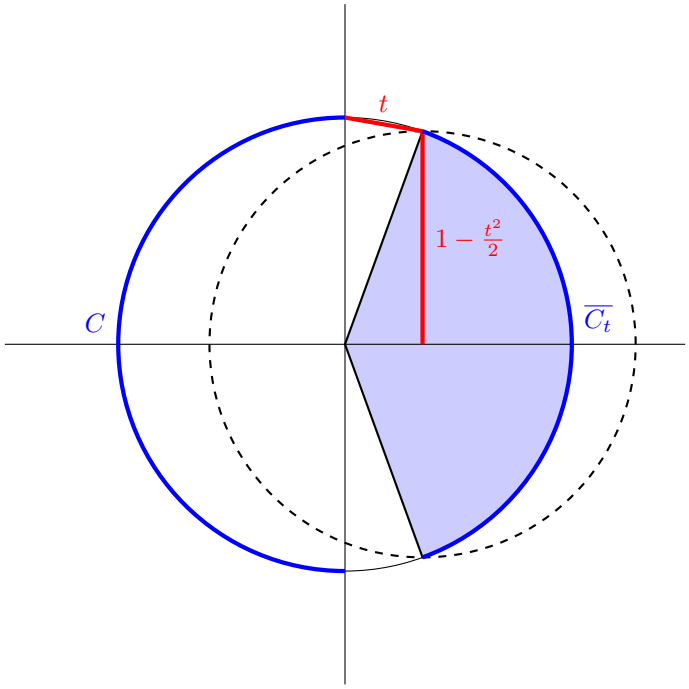
> If $A \subset \mathbb{R}^n$ has the same measure as ball B, then $\text{vol}(A_t) \geq \text{vol}(B_t)$ for all $t \geq 0$.

Analogous exact isoperimetric inequalities are known in several other spaces. We already saw it for the boolean cube (Theorem 9.3.3). The case of sphere and gaussian space are particularly noteworthy. The following theorem is due to Lévy.

Theorem 9.3.9 (Spherical isopeimetric inequality)**.** Inside S^{n-1} (equipped with the natural measure and distance), let A be a subset and B a spherical cap with $\mathrm{vol}_{n-1}(A) = \mathrm{vol}_{n-1}(B)$. Then for all $t \geq 0$,

$$\mathrm{vol}_{n-1}(A_t) = \mathrm{vol}_{n-1}(B_t).$$

Suppose C is a hemisphere in $S^{n-1} \subset \mathbb{R}^n$. Let us estimate $\mathrm{vol}_{n-1}(C)$. As in the diagram below, in the planar cross-section, the chord of length t subtends an angle of $\theta = 2\arcsin(t/2)$, so the vertical bolded segment has length $\cos\theta = 1 - 2\sin^2(\theta/2) = 1 - t^2/2$.



By considering the fraction of the ball subtended by $\overline{C_t}$ (i.e., the shaded wedge-sector above), which is contained in the smaller dashed ball or radius $1 - t^2/2$, we see that

$$1 - \frac{\mathrm{vol}_{n-1}(C_t)}{\mathrm{vol}_{n-1}(S^{n-1})} = \frac{\mathrm{vol}_{n-1}(\overline{C_t})}{\mathrm{vol}_{n-1}(S^{n-1})} \leq \left(1 - \frac{t^2}{2}\right)^n \leq e^{-nt^2/2}.$$

Corollary 9.3.10 (Concentration of measure on a sphere)**.** There exists some constant $c > 0$ so that

- If $A \subseteq S^{n-1}$ has $\mathrm{vol}_{n-1}(A)/\mathrm{vol}_{n-1}(S^{n-1}) \geq 1/2$, then

$$\frac{\mathrm{vol}_{n-1}(A_t)}{\mathrm{vol}_{n-1}(S^{n-1})} \geq 1 - e^{-t^2 n/2}.$$

- If $f\colon S^{n-1} \to \mathbb{R}$ is 1-Lipschitz, then there is some real m (e.g., a median) so that

$$\mathbb{P}(|f - m| > t) \leq 2e^{-nt^2/2}.$$

Second statement may be interpreted as "every Lipschitz function on a high dimensional sphere is nearby constant almost everywhere"

Another related setting is the **Gauss space**, which is $\mathbb{R}^n$ equipped with the the probability measure 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 $(2\pi)^{-n}e^{-|x|^2/2}$ for $x \in \mathbb{R}^n$. Let λ denote the Gaussian measure on $\mathbb{R}^n$. The metric on $\mathbb{R}^n$ is the usual Euclidean metric.

What would an isoperimetric inequality in Gauss space look like?

A naive guess, inspired by $\mathbb{R}^n$, may be that disks minimize perimeter. But this is actually not the case. It turns out that the Hamming cube is a better model for the Gauss space. Indeed, 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.

Note that the Gaussian measure is radially symmetric. So the above heuristic (which can be made rigorous) suggests that for the Gaussian isoperimetric inequality, we should look for **half-spaces**, i.e., points on one side of some hyperplane. This is indeed the case, as first shownindependently by Borell (1975) and Sudakov and Tsirel'son (1974).

Theorem 9.3.11 (Gaussian isoperimetric inequality)**.** If $A, H \subset \mathbb{R}^n$, H a half-space, and $\lambda(A) = \lambda(H)$, then $\lambda(A_t) \geq \lambda(H_t)$ for all $t \geq 0$, where λ is the Gauss measure.

Consequently, if $\mathbb{P}(A) \geq 1/2$, then $\mathbb{P}(\overline{A_t}) \leq \mathbb{P}(Z_1 > t) \leq e^{-t^2/2}$. And, if $f\colon \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz, and z is a vector of iid standard normals, then $X = f(z)$ satisfies

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-t^2/2}$$

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

Furthermore, the length of a random gaussian vector is given by $\sqrt{Z_1^2 + \cdots + Z_n^2}$ for iid $Z_1, \ldots, Z_n \in N(0,1)$, which is concentrated around $\sqrt{n}$ (e.g., by a straight forward adaptation of Chernoff bound. In fact, since $\sqrt{n + O(\sqrt{n})} = \sqrt{n} + O(1)$, the length of gaussian vector has a $O(1)$-length window of typical fluctation). So most of the distribution in the gauss space lies lie to 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. Random gaussian vectors give us a convenient method to analyze the concentration of measure phenomenon on the sphere. (It should now be satisfying to see how half-spaces in the gauss space intersect the sphere in a spherical cap, and both objects are isoperimetric optimzers in their respective spaces).

9.3.2 Johnson–Lindenstrauss Lemma

The next theorem is a powerful in many areas. For example, it is widely used in computer science as a means of dimension reduction.

Theorem 9.3.12 (Johnson and Lindenstrauss 1982)**.** Let $s_1, \ldots, s_N \in \mathbb{R}^n$. Then there exists $s_1', \ldots, s_N' \in \mathbb{R}^m$ where $m = O(\epsilon^{-2} \log N)$ and such that, for every $i \neq j$,

$$(1 - \epsilon)|s_i - s_j| \leq |s_i' - s_j'| \leq (1 + \epsilon)|s_i - s_j|.$$

Remark 9.3.13. Here m is optimal up to a constant factor (Larsen and Nelson 2017).

The theorem is proved by obtaining the new points $s_j' \in \mathbb{R}^m$ by taking a projection onto a uniform random m-dimensional subspace (and the scaling by $\sqrt{n/m}$). We would like to know that these projects roughly preserve the length of vectors. Once we have the following lemma, set $s_i' = \sqrt{m/n} P s_i$, and we can apply the lemma to $z = s_i - s_j$ for every

pair (i, j) and apply the union bound to use that, with probability at least $1 - CN^2 e^{-c\epsilon^2 m}$, one has $(1-\epsilon)|s_i - s_j| \le |s'_i - s'_j| \le (1+\epsilon)|s_i - s_j|$ for all (i, j).

Lemma 9.3.14 (Random projection)**.** Let P be a projection from $\mathbb{R}^n$ onto a random m-dimensional subspace. Let $z \in \mathbb{R}^n$ (fixed) and $y = Pz$. Then

$$\mathbb{E}[|y|^2] = \frac{m}{n}|z|^2$$

and, with probability $\ge 1 - 2e^{-c\epsilon^2 m}$ for some constant $c > 0$,

$$(1-\epsilon)\sqrt{\frac{m}{n}}\,|z| \le |y| \le (1+\epsilon)\sqrt{\frac{m}{n}}\,|z|\,.$$

Proof. By rescaling we may assume that $|z| = 1$.

The distribution of $Y = |y|$ does not change if we instead fix P to be the orthogonal projection onto the subspace spanned by the first m coordinate vectors, and z vary uniformly over the unit sphere.

Writing $z = (z_1, \dots, z_n)$, by symmetry we have $\mathbb{E}[z_1^2] = \cdots = \mathbb{E}[z_n^2]$. Since $z_1^2 + \cdots + z_n^2 = 1$, we have $\mathbb{E}[z_i^2] = 1/n$ for each i. Thus

$$\mathbb{E}[Y^2] = \mathbb{E}[z_1^2 + \cdots + z_m^2] = \frac{m}{n}.$$

Since the map $z \mapsto |y|$ is 1-Lipschitz, by Lévy concentration (Corollary 9.3.10),

$$\mathbb{P}\left(|Y - \mathbb{E}Y| \ge t\right) \le 2e^{-nt^2/2}, \quad \text{for all } t \ge 0.$$

In particular, we have that

$$\mathbb{E}[Y^2] - (\mathbb{E}Y)^2 = \operatorname{Var} Y = \int_0^\infty \mathbb{P}\left(|Y - \mathbb{E}Y|^2 \ge t\right)\, dt \le \int_0^\infty 2e^{-nt/2}\, dt = \frac{4}{n}.$$

So

$$\sqrt{\frac{m-4}{n}} \le \mathbb{E}Y \le \sqrt{\frac{m}{n}}.$$

This implies that, for some constants $c > 0$,

$$\mathbb{P}\left(\left|Y - \sqrt{\frac{m}{n}}\right| \ge t\right) \le 2e^{-cnt^2}, \quad \text{for all } t \ge 0.$$

Setting $t = \epsilon\sqrt{m/n}$ yields the result. □

A cute application of Johnson–Lindenstrauss (this was a starred homework exercise where you were asked to prove it using the Chernoff bound).

Corollary 9.3.15. There is a constant $c > 0$ so that for every positive integer m, there is a set of $e^{c\epsilon^2 m}$ points in $\mathbb{R}^m$ whose pairwise distances are in $[1-\epsilon, 1+\epsilon]$.

Proof. Applying Theorem 9.3.12 to the the N coordinate vectors in $\mathbb{R}^N$ yields a set of N points in $\mathbb{R}^m$ for $m = O(\epsilon^{-2} \log N)$ with pairwise distances in $[1-\epsilon, 1+\epsilon]$. □

9.4 Talagrand inequality

9.4.1 Convex Lipschitz functions of independent random variables

Problem 9.4.1. Let V be a *fixed* d-dimensional subspace. Let $x \sim \mathrm{Unif}\{-1,1\}^n$. How well is $\mathrm{dist}(x, V)$ concentrated?

Let $P = (p_{ij}) \in \mathbb{R}^{n \times n}$ be the matrix giving the orthogonal projection onto $V^\perp$. We have $\operatorname{tr} P = \dim V^\perp = n - d$. Then

$$\mathrm{dist}(x, V)^2 = |x \cdot Px| = \sum_{i,j} x_i x_j p_{ij}.$$

So

$$\mathbb{E}[\mathrm{dist}(x, V)^2] = \sum_i p_{ii} = \operatorname{tr} P = n - d.$$

How well is $\mathrm{dist}(x, V)$ concentrated around $\sqrt{n-d}$?

We say that a random variable X is **K-subgaussian** if

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-t^2/K^2}.$$

Note that a K-subgaussian random variable typically has $O(K)$-fluctuation around its mean.

Let us start with some examples.

If V is some coordinate subspace, then $\mathrm{dist}(x, V)$ is a constant not depending on x.

If $V = (1, 1, \ldots, 1)^\perp$, then $\mathrm{dist}(x, V) = |x_1 + \cdots + x_n|/\sqrt{n}$ which converge $|Z|$ 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 $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$,

one has $\operatorname{dist}(x, V) = |\alpha \cdot x|$. Note that flipping x_i changes $|\alpha \cdot x|$ by at most $2|\alpha_i|$. So So the bounded differences inequality Theorem 9.0.4, for every $t \geq 0$,

$$\mathbb{P}(|\operatorname{dist}(x, V) - \mathbb{E}\operatorname{dist}(x, V)| \geq t) \leq 2\exp\left(\frac{-2t^2}{4(\alpha_1^2 + \cdots + \alpha_n^2)}\right) \leq 2e^{-t^2/2}.$$

So again $\operatorname{dist}(x, V)$ is $O(1)$-subgaussian.

What about higher codimensional subspaces V? Then

$$\operatorname{dist}(x, V) = \sup_{\substack{\alpha \in V^\perp \\ |\alpha|=1}} |\alpha \cdot x|.$$

It is not clear how to apply the bounded difference inequality to all such α in the above supremum simultaneously.

On the other hand, if we were to ignore the α's and simply apply the bounded difference inequality to the function $x \in \{-1, 1\}^n \mapsto \operatorname{dist}(x, V)$, then, since this function is 2-Lipschitz (with respect to Hamming distance), we obtain

$$\mathbb{P}\left(|\operatorname{dist}(x, V) - \mathbb{E}\operatorname{dist}(x, V)| \geq t\right) \leq 2e^{-nt^2/2},$$

showing that $\operatorname{dist}(x, V)$ is $O(\sqrt{n})$-subgaussian—but this is a pretty bad result, as $|\operatorname{dist}(x, V)| \leq \sqrt{n}$ (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).

Instead of sampling $x \in \{-1, 1\}^n$, if we had taking 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 would imply that

$$\mathbb{P}_{x \sim \text{Uniform}(\sqrt{n}S^{n-1})}(|\operatorname{dist}(x, V) - \mathbb{E}\operatorname{dist}(x, V)| \geq t) \leq 2e^{-t^2/2}.$$

So $\operatorname{dist}(x, V)$ is $O(1)$-subgaussian if x is chosen from the radius $\sqrt{n}$ sphere. Perhaps a similar bound holds when x is chosen from $\{-1, 1\}^n$?

Talagrand (1995) developed a powerful inequality that allows us to answer the above question. The most general form of Talagrand's inequality can be somewhat hard to grasp at first, though it has important combinatorial consequences. We begin with more concrete geometric special cases.

Theorem 9.4.2. Let V be a fixed d-dimensional subspace in $\mathbb{R}^n$. For uniformly random $x \in \{-1,1\}^n$, one has

$$\mathbb{P}(|\operatorname{dist}(x,V) - \sqrt{n-d}| \geq t) \leq 2e^{-ct^2}$$

where $c > 0$ is some constant.

Previously, the bounded differences inequality tells us that a Lipschitz function on $\{-1,1\}^n$ is $O(\sqrt{n})$-subgaussian.

Talagrand inequality tells us that a **convex** Lipschitz function in $\mathbb{R}^n$ is $O(1)$-subgaussian when restricted to the boolean cube. We give the precise statement below. We omit the proof of Talagrand's inequality (see Alon–Spencer textbook or Tao's blog post) and instead focus on explaining the theorem and how to apply it.

Below $\operatorname{dist}(\cdot,\cdot)$ means Euclidean distance. And $A_t = \{x : \operatorname{dist}(x,A) \leq t\}$.

Theorem 9.4.3 (Talagrand)**.** Let $A \subset \mathbb{R}^n$ be convex, and let $x \sim \operatorname{Unif}\{0,1\}^n$. Then for any $t > 0$,

$$\mathbb{P}(x \in A)\mathbb{P}(\operatorname{dist}(x,A) \geq t) \leq e^{-ct^2}$$

where $c > 0$ is some absolute constant.

Remark 9.4.4. (1) Note that A is a convex body in $\mathbb{R}^n$ and not simply a set of points in A. It may be useful to think of A as the convex hull of a set of points in $\{-1,1\}^n$. Then distance to A is not the distance to these vertices of the boolean cube, but rather distance to the convex body A.

(2) The bounded differences inequality gives us an upper bound of the form $e^{-ct^2/n}$, which is much better than Talagrand's bound.

Example 9.4.5 (Talagrand's inequality fails for nonconvex sets)**.** Let

$$A = \left\{x \in \{0,1\}^n : \operatorname{wt}(x) \leq \frac{n}{2} - \sqrt{n}\right\}$$

(here A is a discrete set of points and not their convex hull). Then for every $y \in \{0,1\}^n$ with $\operatorname{wt}(y) \geq n/2$, one has $\operatorname{dist}(y,A) \geq n^{1/4}$. Using the central limit theorem, we have, for some constant $c > 0$ and sufficiently large n, for $x \sim \operatorname{Uniform}(\{-1,1\}^n)$, $\mathbb{P}(x \in A) \geq c$ and $\mathbb{P}(\operatorname{wt}(x) \geq n/2) \geq 1/2$, so the above inequality is false for $t = n^{1/4}$.

By an argument similar to our proof of Theorem 9.3.7 (the equivalence of notions of concentration of measure), one can deduce the following consequence.

Corollary 9.4.6. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be convex and 1-Lipschitz (with respect to Euclidean distance on $\mathbb{R}^n$). Then for any $r \in \mathbb{R}$ and $t > 0$, for $x \sim \mathrm{Unif}\{0,1\}^n$

$$\mathbb{P}(f(x) \le r)\mathbb{P}(f(x) \ge r + t) \le e^{-ct^2}.$$

where $c > 0$ is some absolute constant.

Remark 9.4.7. The proof below shows that the assumption that f is convex can be weakened to f being **quasiconvex**, i.e., $\{f \le a\}$ is convex for every $a \in \mathbb{R}$.

The versions of Talagrand inequality, Theorem 9.4.3 and Corollary 9.4.6, are equivalent:

- Theorem 9.4.3 implies Corollary 9.4.6: take $A = \{x : f(x) \le r\}$. We have $f(x) \le r + t$ whenever $\mathrm{dist}(a, A) \le t$ since f is 1-Lipschitz. So $\mathbb{P}(f(x) \le r) = \mathbb{P}(x \in A)$ and $\mathbb{P}(f(x) \ge r + t) \le \mathbb{P}(\mathrm{dist}(x, A) \ge t)$.
- Corollary 9.4.6 implies Theorem 9.4.3: take $f(x) = \mathrm{dist}(x, A)$ which is convex since A is convex.

Let us write $\mathbb{M}X$ to be a **median** for the random variable X, i.e., a non-random real so that $\mathbb{P}(X \ge \mathbb{M}X) \ge 1/2$ and $\mathbb{P}(X \le \mathbb{M}X) \ge 1/2$.

Corollary 9.4.8. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be convex and 1-Lipschitz (with respect to Euclidean distance on $\mathbb{R}^n$). Let $x \sim \mathrm{Unif}(\{0,1\}^n)$. Then

$$\mathbb{P}(|f(x) - \mathbb{M}f(x)| \ge t) \le 2e^{-ct^2}$$

where $c > 0$ is an absolute constant.

Proof. Setting $r = \mathbb{M}f(x)$ in Corollary 9.4.6 yields

$$\mathbb{P}(f(x) \ge \mathbb{M}f(x) + t) \le 2e^{-ct^2},$$

and setting $r = \mathbb{M}f(x)$ in Corollary 9.4.6 yields

$$\mathbb{P}(f(x) \ge \mathbb{M}f(x) - t) \le 2e^{-ct^2}.$$ □

Putting the two inequalities together, and changing the constant c, yields the corollary.

As an immediate corollary, we deduce Theorem 9.4.2 regarding the distance from a random point $x \in \{-1,1\}^n$ to a d-dimensional subspace. The above corollary shows that $\mathrm{dist}(x, V)$ (which is a convex 1-Lipschitz function of $x \in \mathbb{R}^n$) is $O(1)$-subgaussian, which immediately

implies the result (see Lemma 9.3.14 for an example of how to argue the omitted step where we replaced $\mathbb{M}X$ by $\mathbb{E}X$ and then by $(\mathbb{E}X^2)^{1/2}$).

Example 9.4.9 (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 $\mathbb{R}^{n^2} \to \mathbb{R}$, we see that it is convex (since the operator norm is a norm) and 1-Lipschitz (using that $\|\cdot\|_{\text{op}} \le \|\cdot\|_{\text{HS}}$, 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 9.4.8) that f is $O(1)$-subgaussian.

9.4.2 Convex distance

Talagrand's inequality if much more general than what we saw earlier and can be applied to a wide variety of combinatorial applications. We need a 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).

Weighted hamming distance: given $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n_{\ge 0}$, $x, y \in \Omega$, we set

$$d_\alpha(x, y) = \sum_{i=1}^n \alpha_i 1_{x_i \ne y_i}.$$

and for $A \subset \Omega$,

$$d_\alpha(x, A) = \inf_{y \in A} d_\alpha(x, y)$$

Talagrand's **convex distance** between $x \in \Omega$ and $A \subset \Omega$ is defined by

$$d_T(x, A) = \sup_{\substack{\alpha \in \mathbb{R}^n_{\ge 0} \\ |\alpha| = 1}} d_\alpha(x, A)$$

(here $|\alpha|^2 = \alpha_1^2 + \cdots + \alpha_n^2$).

Example 9.4.10. If $A \subset \{0, 1\}^n$ and $x \in \{0, 1\}^n$, then $d_T(x, A)$ is the Euclidean distance from x to the convex hull of A.

To see why this is called a convex distance, note that to compute $d_T(x, A)$, we can convert Ω to $\{0, 1\}^n$ based on their agreement with x, i.e., let $\phi_x(y) \in \{0, 1\}^n$ be the vector whose i-th coordinate is 1 iff $x_i \ne y_i$. Then, $d_\alpha(x, A)$ in Ω equals to $d_\alpha(\vec{0}, \phi_x(A)) = \phi_x(A) \cdot \alpha$ in $\{0, 1\}^n$. Taking the supremum over α, we see, using the Example 9.4.10,

$$d_T(x, A) = \operatorname{dist}(\vec{0}, \operatorname{ConvexHull} \phi_x(A)).$$

The general form of Talagrand's inequality says the following. Note that it reduces to the earlier special case Theorem 9.4.3 if $\Omega = \{0,1\}^n$.

Theorem 9.4.11 (General form of Talagrand's inequality)**.** Let $A \subseteq \Omega = \Omega_1 \times \cdots \times \Omega_n$, with Ω equipped with a product probability measure. Let $t \geq 0$. We have

$$\mathbb{P}(A)\mathbb{P}(x \in \Omega : d_T(x, A) \geq t) \leq e^{-t^2/4}.$$

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 9.4.12 (Convex distance upper bounds Euclidean distance)**.** Let $A \subset [0,1]^n$ and $x \in [0,1]^n$. Then $\operatorname{dist}(x, \operatorname{ConvexHull} A) \leq d_T(x, A)$.

Proof. For any $\alpha \in \mathbb{R}^n$, and any $y \in [0,1]^n$, we have

$$|(x-y)\cdot\alpha| \leq \sum_{i=1}^n |\alpha_i|\,|x_i - y_i| \leq \sum_{i=1}^n |\alpha_i|\,1_{x_i \neq y_i}.$$

First taking the infimum over all $y \in A$, and then taking the supremum over unit vectors α, the LHS becomes $\operatorname{dist}(x, \operatorname{ConvexHull} A)$ and the RHS becomes $d_T(x, A)$. □

Corollary 9.4.13 (Convex functions of independent bounded random variables)**.** Let $x = (x_1, \ldots, x_n) \in [0,1]$ be independent random variables (not necessarily identical). Let $t \geq 0$. Let $A \subset [0,1]^n$ be a convex set. Then

$$\mathbb{P}(x \in A)\mathbb{P}(\operatorname{dist}(x, A) \geq t) \leq e^{-t^2/4}$$

where dist is Euclidean distance. Also, if $f\colon [0,1]^n \to \mathbb{R}$ is a convex 1-Lipschitz function, then

$$\mathbb{P}(|f - \mathbb{M}f| \geq t) \leq 4e^{-t^2/4}.$$

9.4.3 How to apply Talagrand's inequality

Theorem 9.4.14. Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ equipped with the product measure. Let $f : \Omega \to \mathbb{R}$ be a function. Suppose for every $x \in \Omega$, there is some $\alpha(x) \in \mathbb{R}^n_{\geq 0}$ such that for every $y \in \Omega$,

$$f(x) \leq f(y) + d_{\alpha(x)}(x, y).$$

Then, for every $t \geq 0$,

$$\mathbb{P}(|f - \mathbb{M}f| \geq t) \leq 4 \exp\left(\frac{-t^2}{4 \sup_{x \in \Omega} |\alpha(x)|^2}\right).$$

Remark 9.4.15. Note that we can use a different weight $\alpha(x)$ for each x. This will be important for applications. Intuitively, it says that the smallness (or, equivalently the largeness) of $f(x)$ can be "certified" using $\alpha(x)$.

Remark 9.4.16. By considering $-f$ instead of f, we can change the hypothesis on f to

$$f(x) \geq f(y) - d_{\alpha(x)}(x, y).$$

Note that x and y play asymmetric roles.

Remark 9.4.17 (Talagrand recovers bounded differences). By choosing a fixed $\alpha \in \mathbb{R}^n_{\geq 0}$ (not varying with x), we see that Theorem 9.4.14 recovers the bounded differences inequality Theorem 9.0.4 up to an unimportant constant factor in the exponent of the bound. The power of Talagrand's inequality is that we are allowed to vary $\alpha(x)$.

Proof. Let $r \in \mathbb{R}$. Let $A = \{y \in \Omega : f(y) \leq r - t\}$. For any $x \in \Omega$, by hypothesis, there is some $\alpha(x) \in \mathbb{R}^n_{\geq 0}$ such that, for all $y \in A$,

$$f(x) \leq f(y) + d_{\alpha(x)}(x, y) \leq r - t + d_{\alpha(x)}(x, y).$$

Taking infimum over $y \in A$, we find

$$f(x) \leq r - t + d_{\alpha(x)}(x, A) \leq r - t + |\alpha(x)|\, d_T(x, A).$$

Thus, if $f(x) \geq r$, then

$$d_T(x, A) \geq \frac{t}{|\alpha(x)|} \geq \frac{t}{\sup_x |\alpha(x)|} =: s$$

And hence by Talagrand's inequality Theorem 9.4.11,

$$\mathbb{P}(f \leq r - t)\mathbb{P}(f \geq r) \leq \mathbb{P}(A)\mathbb{P}(x \in \Omega : d_T(x, A) \geq s) \leq e^{-s^2/4}.$$

Taking $r = \mathbb{M}f + t$ yields

$$\mathbb{P}(f \geq \mathbb{M}f + t) \leq 2e^{-s^2/4}$$

and taking $r = \mathbb{M}f$ yields

$$\mathbb{P}(f \leq \mathbb{M} - t) \leq 2e^{-s^2/4}.$$

Putting them together yields the final result. □

9.4.4 Largest eigenvalue of a random matrix

Theorem 9.4.18. Let $A = (a_{ij})$ be an $n \times n$ symmetric random matrix with independent entries in $[-1, 1]$. Let $\lambda_1(X)$ 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 9.4.14. 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,

$$\lambda_1(A) \leq \lambda_1(B) + \sum_{i \leq j} \alpha_{i,j} 1_{a_{ij} \neq b_{ij}}. \tag{9.5}$$

(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,

$$v^\intercal A v = \lambda_1(A) \qquad \text{and} \qquad v^\intercal B v \leq \lambda_1(B).$$

We have

$$\lambda_1(A) = v^\intercal A v = v^\intercal B v + v^\intercal (A - B) v \leq \lambda_1(B) + \sum_{i,j} 2 \left|v_i\right| \left|v_j\right| 1_{a_{ij} \neq b_{ij}}$$

(since $|a_{ij} - b_{ij}| \leq 2$). Thus (9.5) holds for the vector $\alpha(A) = (\alpha_{ij})_{i \leq j}$ defined by

$$\alpha_{ij} = \begin{cases} 4 \left|v_i\right| \left|v_j\right| & \text{if } i < j \\ 2 \left|v_i\right|^2 & \text{if } i = j. \end{cases}$$

We have

$$\sum_{i\le j} \alpha_{ij}^2 \le 8 \sum_{i,j} |v_i|^2 |v_j|^2 = 8 \left(\sum_i |v_i|^2 \right)^2 = 8.$$

So Theorem 9.4.14 yields the result. □

Remark 9.4.19. The above method can be adapted to prove concentration of the k-th largest eigenvalue, which is not a convex function of A, so the previous method in Example 9.4.9 does not apply.

Remark 9.4.20. 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 called a Tracy–Widom distribution.

9.4.5 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 $i_1 < \cdots < i_\ell$.

Question 9.4.21. How well is the length X of the longest increasing subsequence (LIS) of uniform random permutation concentrated?

While the entries of σ are not independent, we can generate a uniform random permutation by taking iid uniform $x_1, \dots, x_n \sim \mathrm{Unif}[0, 1]$ and let σ record the ordering of the x_i's. This trick converts the problem into one about independent random variables.

The probability that there exists an increasing subsequence of length k is, by union bound, at most

$$\mathbb{P}(X \ge k) \le \frac{1}{k!}\binom{n}{k} \le \left(\frac{e}{k}\right)^k \left(\frac{ne}{k}\right)^k \le \left(\frac{e^2 n}{k^2}\right)^k.$$

It follows that $\mathbb{M}X = O(\sqrt{n})$.

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 $O(\sqrt{\mathbb{M}X})$, or $O(n^{1/4})$ in this case.

Definition 9.4.22. Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$. Let $A \subseteq \Omega$. We say that A is **s-certifiable** 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 $y \in A$.

Theorem 9.4.23. Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be equipped with a product measure. Let $f : \Omega \to \mathbb{R}$ 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 $A, B \subset \Omega$ be given by $A = \{x : f(x) \leq r - t\}$ and $B = \{y : f(y) \geq r\}$. To apply Talagrand's inequality, Theorem 9.4.11, it suffices to show that for every $y \in B$, one has $d_T(y, A) \geq t/\sqrt{s}$, i.e., there is some $\alpha(y) \in \mathbb{R}^n_{\geq 0}$ so that

$$d_\alpha(x, y) \geq t|\alpha(y)|/\sqrt{s} \qquad \forall x \in A.$$

Indeed, let $y \in B$, and let $I(y)$ be a set of s coordinates that certify $f(y) \geq r$. Let $\alpha(y)$ be the indicator vector for $I(y)$. Note that

$$d_\alpha(x, y) = |\{i \in I(y) : x_i \neq y_i\}|.$$

Every $x \in A$ disagrees with y on at least t coordinates of $I(y)$, or else one can change x by fewer than t coordinates to get x' that agrees with y on I, so that $f(x') \geq r$, which contradicts f being 1-Lipschitz as $f(x) \leq r - t$. It follows that

$$d_\alpha(x, y) \geq t = t|\alpha(y)|/\sqrt{s}.$$ □

Corollary 9.4.24. Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be equipped with a product measure. Let $f : \Omega \to \mathbb{R}$ 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 $t \geq 0$,

$$\mathbb{P}(f \leq r - t)\mathbb{P}(X \geq r) \leq \exp\left(\frac{-t^2}{4r}\right).$$

Setting $r = \mathbb{M}f$, we obtain the lower tail.

$$\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2\exp\left(\frac{-t^2}{4m}\right).$$

Setting $r = m + t$, we obtain the upper tail

$$\mathbb{P}(X \geq \mathbb{M}f + t) \leq 2\exp\left(\frac{-t^2}{4(\mathbb{M}f + t)}\right). \qquad \square$$

Corollary 9.4.25. Let X be the length of the longest increasing subsequence of a random permutation of $[n]$. Then for every $\epsilon > 0$ there exists $C > 0$ so that

$$\mathbb{P}(|X - \mathbb{M}X| \leq Cn^{1/4}) \geq 1 - \epsilon.$$

Remark 9.4.26. 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 correcting scaling is $n^{1/6}$, and, after under this normalization, $n^{-1/6}(X - 2\sqrt{n})$ converges to the Tracy–Widom distribution, the same distribution for the top eigenvalue of a random matrix.

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** 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).

Intuitively, $H(X)$ measures the amount of "surprise" in the randomness of X. Note that we always have

$$H(X) \geq 0.$$

A more rigorous interpretation of this intuition is given by the Shannon noiseless coding theorem, which 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.

Lemma 10.1.2 (Uniform bound)**.**

$$H(X) \leq \log_2 |\operatorname{support}(X)|,$$

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

$$H(X) = \sum_{s \in S} f(p_s) \leq |S| \, f\left(\frac{1}{|S|} \sum_{s \in S} p_s\right) = |S| \, f\left(\frac{1}{|S|}\right) = \log_2 |S| \, .$$

□

We write $H(X, Y)$ for the entropy of the joint random variables (X, Y), i.e., letting $Z = (X, Y)$,

$$H(X, Y) := H(Z) = \sum_{(x,y)} -\mathbb{P}(X = x, Y = y) \log_2 \mathbb{P}(X = x, Y = y).$$

Note that

$$H(X, Y) = H(X) + H(Y) \qquad \text{if } X \text{ and } Y \text{ are independent.}$$

Definition 10.1.3 (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).

Lemma 10.1.4 (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

$$\begin{aligned} H(X|Y) := \mathbb{E}_y[H(X|Y=y)] &= \sum_y -p(y) \sum_x p(x|y) \log_2 p(x|y) \\ &= \sum_{x,y} -p(x,y) \log_2 \frac{p(x,y)}{p(y)} \\ &= \sum_{x,y} -p(x,y) \log_2 p(x,y) + \sum_y p(y) \log_2 p(y) \\ &= H(X,Y) - H(Y). \end{aligned}$$

□

Intuitively, the conditional entropy $H(X|Y)$ measures the amount of additional information in X not contained in Y.

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 $H(X|Y,Z) = H(X|Z)$.

Lemma 10.1.5 (Subadditivity)**.** $H(X,Y) \leq H(X) + H(Y)$, and more generally,

$$H(X_1, \ldots, X_n) \leq H(X_1) + \cdots + H(X_n).$$

Proof.

$$\begin{aligned} H(X) + H(Y) - H(X,Y) &= \sum_{x,y} \left(-p(x,y) \log_2 p(x) - p(x,y) \log_2 p(y) + p(x,y) \log_2 p(x,y)\right) \\ &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x)p(y) f\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &\geq f(1) = 0 \end{aligned}$$

where $f(t) = t$ is convex.

More generally, by iterating the above inequality for two random variables, we have

$$\begin{aligned} H(X_1, \ldots, X_n) &\leq H(X_1, \ldots, X_{n-1}) + H(X_n) \\ &\leq H(X_1, \ldots, X_{n-2}) + H(X_{n-1}) + H(X_n) \\ &\leq \cdots \leq H(X_1) + \cdots + H(X_n). \end{aligned}$$

□

Remark 10.1.6. The nonnegative quantity

$$I(X;Y) := H(X) + H(Y) - H(X,Y)$$

is called *mutual information*. Intuitively, it measures the amount of common information between X and Y.

Lemma 10.1.7 (Dropping conditioning)**.** $H(X|Y) \leq H(X)$ and $H(X|Y,Z) \leq H(X|Z)$

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

$$H(X|Y, Z = z) \geq H(X|Z = z)$$

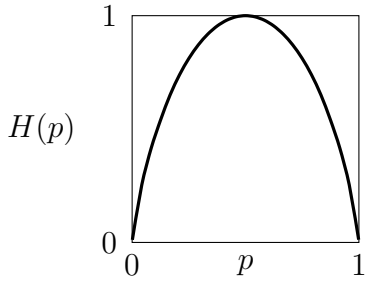
holds for every z, and taking expectation of z yields $H(X|Y,Z) \leq H(X|Z)$. □

Remark 10.1.8. The above inequality is often equivalently (why?) rephrased as the *data processing inequality*: $H(X|f(Y)) \geq H(X|Y)$ for any function f.

Here are some simple applications of entropy to **tail bounds**.

Let us denote the entropy of a Bernoulli random variable by

$$H(p) := H(\text{Bernoulli}(p)) = -p \log_2 p - (1-p) \log_2 (1-p).$$



Theorem 10.1.9. If $k \le n/2$, then

$$\sum_{0 \le i \le k} \binom{n}{i} \le 2^{H(k/n)n}.$$

Equivalently, the above inequality says that for $X \sim \text{Binomial}(n, 1/2)$, we have $\mathbb{P}(X \le k) \le 2^{(H(k/n)-1)n}$. This bound can be established using our proof technique for Chernoff bound by applying Markov's inequality to the moment generating function:

$$\sum_{0 \le i \le k} \binom{n}{i} \le \frac{(1+x)^n}{x^k} \qquad \forall x \in [0,1].$$

The infimum of the RHS over $x \in [0,1]$ is precisely $2^{(H(k/n)-1)n}$.

Now let us give a purely information theoretic proof. We can use the above theorem but let's do it from scratch to practice with entropy.

Proof. Let $(X_1, \ldots, X_n) \in \{0,1\}^n$ be chosen uniformly *conditioned* on $X_1 + \cdots + X_n \le k$. Then

$$\log_2 \sum_{0 \le i \le k} \binom{n}{i} = H(X_1, \ldots, X_n) \le H(X_1) + \cdots + H(X_n).$$

Each X_i is a Bernoulli with probability $\mathbb{P}(X_i = 1)$. Note that conditioned on $X_1 + \cdots + X_n = m$, one has $\mathbb{P}(X_i = 1) = m/n$. Varying over $m \le k \le n/2$, we find $\mathbb{P}(X_i = 1) \le k/n$, so $H(X_i) \le H(k/n)$. Hence

$$\log_2 \sum_{0 \le i \le k} \binom{n}{i} \le H(k/n)n. \qquad \square$$

Remark 10.1.10. 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

$$\frac{\log \mathbb{P}(X \le nq)}{n} \le -q \log \frac{q}{p} - (1-q) \log \frac{1-q}{1-p} \qquad \forall 0 \le q \le p$$

and

$$\frac{\log \mathbb{P}(X \ge nq)}{n} \le -q \log \frac{q}{p} - (1-q) \log \frac{1-q}{1-p} \qquad \forall p \le q \le 1.$$

10.2 Upper bound on the permanent and the number of perfect matchings

We define the **permanent** of $n \times n$ matrix A by

$$\operatorname{per} A := \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i,\sigma(i)}.$$

Formula for the permanent is simply that of the determinant without the extra sign factor:

$$\det A := \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i\sigma_i}.$$

We'll consider $\{0,1\}$-valued matrices. If A is the bipartite adjacency matrix of a bipartite graph, then $\operatorname{per} A$ is its 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)**.** Let $A = (a_{ij}) \in \{0,1\}^{n \times n}$, whose i-th row has sum d_i. Then

$$\operatorname{per} A \leq \prod_{i=1}^{n} (d_i!)^{1/d_i}$$

Note that equality is attained when A consists diagonal blocks of 1's (corresponding to perfect matchings in a bipartite graph of the form $K_{d_1,d_1} \sqcup \cdots \sqcup K_{d_t,d_t}$).

Proof. (Radhakrishnan 1997) Let σ be a uniform random permutation of $[n]$ conditioned on $a_{i\sigma_i} = 1$ for all $i \in [n]$. Then

$$\log_2 \operatorname{per} A = H(\sigma) = H(\sigma_1, \ldots, \sigma_n) = H(\sigma_1) + H(\sigma_2|\sigma_1) + \cdots + H(\sigma_n|\sigma_1, \ldots, \sigma_{n-1}).$$

We could have bounded $H(\sigma_i|\sigma_1, \ldots, \sigma_{i-1}) \leq H(\sigma_i) \leq \log_2 |\operatorname{support} \sigma_i| = \log_2 d_i$, but this step would be too lossy.

Here is a useful trick: **reveal the chosen entries in a uniform random order.**

Let $(\tau_1, \ldots, \tau_n)$ be a uniform random permutation of $[n]$. We have

$$H(\sigma) = H(\sigma_{\tau_1}) + H(\sigma_{\tau_2}|\sigma_{\tau_1}) + \cdots + H(\sigma_{\tau_n}|\sigma_{\tau_1}, \ldots, \sigma_{\tau_{n-1}}).$$

For now, consider the i-th row for a fixed i. Let $k \in [n]$ be the index with $\tau_k = i$.

After seeing $\sigma_{\tau_1}, \ldots, \sigma_{\tau_{k-1}}$, the expected number of remaining choices for σ_i is uniformly distributed in $[d_i]$ (since τ is uniform), so applying the uniform bound we have

$$H(\sigma_i|\sigma_{\tau_1}, \ldots, \sigma_{\tau_{k-1}}) \leq \mathbb{E}[\log_2 \text{support}(\sigma_i|\sigma_{\tau_1}, \ldots, \sigma_{\tau_{k-1}})] = \frac{\log_2 1 + \cdots \log_2 d_i}{d_i} = \frac{\log_2(d_i!)}{d_i}.$$

It follows that

$$\log_2 \operatorname{per} A = H(\sigma) \leq \sum_{i=1}^{n} \frac{\log_2(d_i!)}{d_i}$$

and the conclusion follows. □

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 bipartite 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. □

10.2.1 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 γ, while the lower bound $n!/2^{n-1}$ has been improved by a constant factor. 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 $i \to j$ is an edge of the tournament and 0 otherwise. Let d_i be the sum of the i-th row. Then $\operatorname{per} A$ counts the number of 1-factors (spanning disjoint unions of directed cycles) of the tournament. So by Brégman's theorem, we have

$$\text{number of Hamilton cycles} \le \operatorname{per} A \le \prod_{i=1}^{n} (d_i!)^{1/d_1}.$$

One can check (omitted) that the function $g(x) = (x!)^{1/x}$ is log-concave, i.e, $g(n)g(n+2) \ge g(n+1)^2$ for all $n \ge 0$. 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 $O(\sqrt{n} \cdot n!/2^n)$. □

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. □

10.3 Sidorenko's inequality

Given graphs F and G, a **graph homomorphism** from F to G is a map $\phi\colon V(F) \to V(G)$ of vertices that sends edges to edges, i.e., $\phi(u)\phi(v) \in E(G)$ for all $uv \in E(F)$.

Let

$$\hom(F, G) = \text{ the number of graph homomorphisms from } F \text{ to } G.$$

Define the **homomorphism density** (the *H-density in G*) by

$$\begin{aligned} t(F, H) &= \frac{\hom(F, G)}{v(G)^{v(F)}} \\ &= \mathbb{P}(\text{a uniform random map } V(F) \to V(G) \text{ is a graph homomorphism } F \to G) \end{aligned}$$

In this section, we are interested in the regime of fixed F and large G, in which case almost all maps $V(F) \to V(G)$ are injective, so that there is not much difference between homomorphisms and subgraphs. More precisely,

$$\hom(F, G) = \mathrm{aut}(F)(\#\text{copies of } F \text{ in } G \text{ as a subgraph}) + O_F(v(G)^{v(F)}).$$

where $\mathrm{aut}(F)$ is the number of automorphisms of F.

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 $n \to \infty$.

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.

Theorem 10.3.3 (Blakey and Roy 1965)**.** Sidorenko's conjecture holds if F is a tree.

Proof. We will construct a probability distribution μ on $\mathrm{Hom}(F, G)$, the set of all graph homomorphisms $F \to G$. Unlike earlier applications of entropy, here we are trying to prove a lower bound on $\hom(F, G)$ instead of an upper bound. So instead of taking μ to be a uniform distribution (which automatically has entropy $\log_2 \hom(F, G)$), we actually take μ to be carefully constructed distribution, and apply the upper bound

$$H(\mu) \leq \log_2 |\mathrm{support}\, \mu| = \log_2 \hom(F, G).$$

We are trying to show that

$$\frac{\hom(F,G)}{v(G)^{v(F)}} \geq \left(\frac{2e(G)}{v(G)^2}\right)^{e(F)}.$$

So we would like to find a probability distribution μ on $\mathrm{Hom}(F,G)$ satisfying

$$H(\mu) \geq e(F)\log_2(2e(G)) - (2e(F) - v(F))\log_2 v(G). \tag{10.1}$$

Let us explain the proof when F is a path on 4 vertices. The same proof extends to all trees F.

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

$$H(Z|X,Y) = H(Z|Y) \qquad \text{and} \qquad H(W|X,Y,Z) = H(W|Z)$$

and futhermore both quantities are equal to $H(Y|X)$ since XY, YZ, ZW are each distributed as a uniform random edge.

Thus

$$\begin{aligned}
H(X,Y,Z,W) &= H(X) + H(Y|X) + H(Z|X,Y) + H(W|X,Y,Z) && \text{[chain rule]}\\
&= H(X) + H(Y|X) + H(Z|Y) + H(W|Z) && \text{[conditional independence]}\\
&= H(X) + 3H(Y|X)\\
&= 3H(X,Y) - 2H(X) && \text{[chain rule]}\\
&\geq 3\log_2(2e(G)) - 2\log_2 v(G)
\end{aligned}$$

In the final step we used $H(X,Y) = \log_2(2e(G))$ since XY is uniformly distributed

among edges, and $H(X) \le \log_2 |\text{support}(X)| = \log_2 v(G)$. 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.)

This proof easily generalizes to all trees. □

Remark 10.3.4. See this MathOverflow discussions for the history as well as alternate proofs: https://mathoverflow.net/q/189222/

Theorem 10.3.5. 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.

From the previous proof (applied to a 2-edge path), we see that

$$H(X_1, Y_1, Y_2) \ge 2H(X_1, Y_1) - H(X_1) \ge 2\log_2(2e(G)) - \log_2 v(G).$$

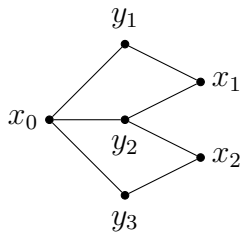
So we have

$$\begin{aligned}
&H(X_1, X_2, Y_1, Y_2) \\
&= H(Y_1, Y_2) + H(X_1, X_2 | Y_1, Y_2) && \text{[chain rule]} \\
&= H(Y_1, Y_2) + 2H(X_1 | Y_1, Y_2) && \text{[conditional independence]} \\
&= 2H(X_1, Y_1, Y_2) - H(Y_1, Y_2) && \text{[chain rule]} \\
&\ge 2(2\log_2(2e(G)) - \log_2 v(G)) - 2\log_2 v(G). && \text{[prev. ineq. and uniform bound]} \\
&= 4\log(2e(G)) - 4\log_2 v(G).
\end{aligned}$$

So we have verified (10.1) for $K_{2,2}$. □

Theorem 10.3.6 (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

$$
\begin{aligned}
&H(X_0, X_1, X_2, Y_1, Y_2, Y_3) \\
&= H(X_0, X_1, X_2 | Y_1, Y_2, Y_3) + H(Y_1, Y_2, Y_3) && \text{[chain rule]} \\
&= H(X_0 | Y_1, Y_2, Y_3) + H(X_1 | Y_1, Y_2, Y_3) + H(X_2 | Y_1, Y_2, Y_3) + H(Y_1, Y_2, Y_3) && \text{[conditional independence]} \\
&= H(X_0 | Y_1, Y_2, Y_3) + H(X_1 | Y_1, Y_2) + H(X_2 | Y_2, Y_3) + H(Y_1, Y_2, Y_3) && \text{[conditional independence]} \\
&= H(X_0, Y_1, Y_2, Y_3) + H(X_1, Y_1, Y_2) + H(X_2, Y_2, Y_3) - H(Y_1, Y_2) - H(Y_2, Y_3). && \text{[chain rule]}
\end{aligned}
$$

The proof of Theorem 10.3.3 actually lower bounds the first three terms:

$$
\begin{aligned}
H(X_0, Y_1, Y_2, Y_3) &\geq 3 \log_2(2e(G)) - 2 \log_2 v(G) \\
H(X_1, Y_1, Y_2) &\geq 2 \log_2(2e(G)) - \log_2 v(G) \\
H(X_2, Y_2, Y_3) &\geq 2 \log_2(2e(G)) - \log_2 v(G).
\end{aligned}
$$

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

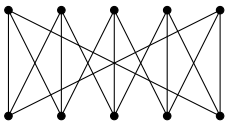
$$H(X_0, X_1, X_2, Y_1, Y_2, Y_3) \geq 7 \log_2(2e(G)) - 8 \log_2 v(G),$$

thereby verifying (10.1). □

To check that you understand the above proof, where did we use the assumption that F has vertex complete to the other part?

Many other graphs can be proved by extending this method.

The "smallest" open case of Sidorenko conjecture is when F is the following graph, often called the "Möbius graph", which is $K_{5,5}$ with a Hamilton cycle removed. (I think it is called the "Möbius graph" because it is the face-vertex incidence graph of the simplicial complex structure of the Möbius strip, built by gluing a strip of five triangles.)

Möbius graph $= K_{5,5} \setminus C_{10} =$ 

10.4 Shearer's lemma

Shearer's entropy lemma extends the subadditivity property of entropy. Befpre 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

$$\begin{aligned} H(X, Y) &= H(X) + H(Y|X) \\ H(X, Z) &= H(X) \qquad\quad + H(Z|X) \\ H(Y, Z) &= \qquad\quad H(Y) + H(Z|Y) \end{aligned}$$

Applying conditioning dropping, we see that their sum is at at least

$$2H(X, Y, Z) = 2H(X) + 2H(Y|X) + 2H(Z|X, Y).$$ □

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 $S \subset \mathbb{R}^3$ 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

$$2\log_2|S| = 2H(X,Y,Z) \leq H(X,Y)+H(X,Z)+H(Y,Z) \leq \log_2 \pi_{xy}(S)+\log_2 \pi_{xz}(S)+\log_2 \pi_{yz}(S).$$

□

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

$$\mathrm{vol}(S)^2 \leq \mathrm{area}(\pi_{xy}(S))\,\mathrm{area}(\pi_{xz}(S))\,\mathrm{area}(\pi_{yz}(S)).$$

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 $A_1, \ldots, A_s \subset [n]$ where each $i \in [n]$ appears in at least k sets A_j's. Writing $X_A := (X_i)_{i\in A}$,

$$kH(X_1, \ldots, X_n) \leq \sum_{j\in[s]} H(X_{A_j}).$$

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 $S \subset \mathbb{R}^n$,

$$|S|^{n-1} \leq \prod_{i=1}^{n} |\pi_i(S)|$$

Corollary 10.4.7. Let $A_1, \ldots, A_s \subset \Omega$ where each $i \in \Omega$ appears in at least k sets A_j. Then for every family $\mathcal{F}$ of subsets of Ω,

$$|\mathcal{F}|^k \leq \prod_{j \in [s]} \left|\mathcal{F}|_{A_j}\right|$$

where $\mathcal{F}|_A := \{F \cap A : F \in \mathcal{F}\}$.

Proof. Each subset of Ω corresponds to a vector $(X_1, \ldots, X_n) \in \{0,1\}^n$. Let $(X_1, \ldots, X_n)$ be a random vector corresponding to a uniform element of $\mathcal{F}$. Then

$$k \log_2 |\mathcal{F}| = kH(X_1, \ldots, X_n) \leq \sum_{j \in [s]} H(X_{A_j}) = \log_2 \left|\mathcal{F}|_{A_j}\right|. \qquad \square$$

10.4.1 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

$$|\mathcal{G}| \leq 2^{\binom{n}{2} - \frac{\binom{n}{2}}{r}} < 2^{\binom{n}{2}-2}.$$

□

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.

10.4.2 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? 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.

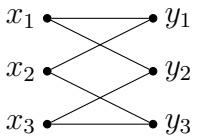
Alon (1991) proved an approximate version of this conjecture. Kahn (2001) proved it assuming the graph is bipartite. Zhao (2010) proved it in general.

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

$G =$ 

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

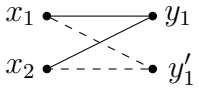
$$\begin{aligned}
2\log_2 i(G) &= 2H(X_1, X_2, X_3, Y_1, Y_2, Y_3) \\
&= 2H(X_1, X_2, X_3) + 2H(Y_1, Y_2, Y_3 | X_1, X_2, X_3) && \text{[chain rule]} \\
&\leq H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3) \\
&\quad + 2H(Y_1|X_1, X_2, X_3) + 2H(Y_2|X_1, X_2, X_3) + 2H(Y_3|X_1, X_2, X_3) && \text{[Shearer]} \\
&= H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3) \\
&\quad + 2H(Y_1|X_1, X_2) + 2H(Y_2|X_1, X_3) + 2H(Y_3|X_2, X_3) && \text{[conditional independence]}
\end{aligned}$$

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 $S \subset V(G)$, 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 indictor 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

$$\begin{aligned} H(X_{N(b)}) + dH(X_b|X_{N(b)}) &= H(X_{N(b)}) + H(X_b^{(1)}, \ldots, X_b^{(d)}|X_{N(b)}) \\ &= H(X_b^{(1)}, \ldots, X_b^{(d)}, X_{N(b)}) \\ &\leq \log_2 i(K_{d,d}) \end{aligned}$$

where $X_b^{(1)}, \ldots, X_b^{(d)}$ are conditionally independent copies of X_b given $X_{N(b)}$. Summing over all b yields the result. □

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 6), 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 $u_0 v_1$ 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 case of

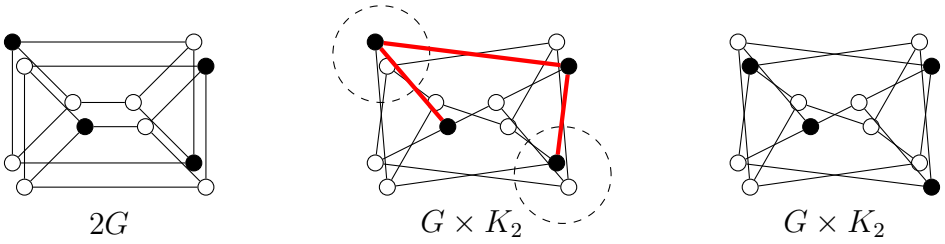


Figure 6: The bipartite swapping trick in the proof of Lemma 10.4.13: swapping the circled pairs of vertices (denoted A in the proof) fixes the bad edges (red and bolded), transforming an independent set of $2G$ into an independent set of $G \times K_2$.

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 $\phi\colon I(2G) \to I(G \times K_2)$. 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.) □

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

$$\hom(G, H) \leq \hom(K_{d,d}, H)^{n/(2d)}$$

Some important special cases:

- $\hom(G, \text{  }) = i(G)$, the number of independent sets of G;
- $\hom(G, K_q)$ = 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 = $  , then $\log_2 \hom(G, H)$ is the number of connected components of G, so that the maximizers of $\log_2 \hom(G, H)/v(G)$ 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 $G(n,p)$
  - Szemeŕedi'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

11.1.1 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 $\epsilon > 0$, there exists $C > 0$ such that the following holds.
For every n, there is a collection $\mathcal{C}$ of graphs on n vertices, with

$$|\mathcal{C}| \leq n^{Cn^{3/2}}$$

such that

(a) every $G \in \mathcal{C}$ has at most $(\frac{1}{4} + \epsilon)n^2$ edges, and

(b) every triangle-free graph is contained in some $G \in \mathcal{C}$.

Proof of upper bound of Theorem 11.0.2 (the number of n-vertex triangle-free graphs is $2^{n^2/4+o(n^2)}$. Let $\epsilon > 0$ be any real number (arbitrarily small). Let $\mathcal{C}$ be produced by Theorem 11.1.1.

Then every $G \in \mathcal{C}$ has at most $(\frac{1}{4} + \epsilon)n^2$ edges, and every triangle-free graph is contained in some $G \in \mathcal{C}$. Hence the number of triangle-free graphs is

$$|\mathcal{C}|\, 2^{(\frac{1}{4}+\delta)n^2} \leq 2^{(\frac{1}{4}+\epsilon)n^2 + O_\epsilon(n^{3/2}\log n)}.$$

Since $\epsilon > 0$ can be made arbitrarily small, the number triangle-free graphs on n vertices is $2^{(\frac{1}{4}+o(1))n^2}$. □

The same proof technique, with an appropriate container theorem, can be used to count H-free graphs.

We write $\mathrm{ex}(n, H)$ 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

$$\mathrm{ex}(n, H) = \left(1 - \frac{1}{\chi(H) - 1} + o(1)\right)\binom{n}{2}.$$

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 $\mathrm{ex}(n, H)$ 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 $2^{(1+o(1))\,\mathrm{ex}(n,H)}$.

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 $2^{O(\mathrm{ex}(n,H))}$.

11.1.2 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 $O(n^3p^3)$, whereas there are whp $n^2p/2$ edges, and $n^3p^3 \ll n^2p$, so we can remove $o(n^2p)$ 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 $\epsilon > 0$ be aribtrarily small. Let $\mathcal{C}$ be a set of containers for n-vertex triangle-free graphs in Theorem 11.1.1. For every $G \in \mathcal{C}$, $e(G) \leq \left(\frac{1}{4} + \epsilon\right) n^2$, so by an application of the Chernoff bound,

$$\mathbb{P}\left(e(G \cap G(n,p)) > \left(\frac{1}{4} + 2\epsilon\right) n^2 p\right) \leq e^{-\Omega_\epsilon(n^2 p)}$$

Since every triangle-free graph is contained in some $G \in \mathcal{C}$, by taking a union bound over $\mathcal{C}$, we see that

$$\begin{aligned}
&\mathbb{P}\left(G(n,p) \text{ has a triangle-free subgraph with } > \left(\frac{1}{4} + 2\epsilon\right) n^2 p \text{ edges}\right) \\
&\leq \sum_{G \in \mathcal{C}} \mathbb{P}\left(e(G \cap G(n,p)) > \left(\frac{1}{4} + 2\epsilon\right) n^2 p\right) \\
&\leq |\mathcal{C}|\, e^{-\Omega_\epsilon(n^2 p)} \\
&\leq e^{O_\epsilon(n^{3/2} \log n) - \Omega_\epsilon(n^2 p)} \\
&= o(1)
\end{aligned}$$

provided that $p \gg n^{-1/2} \log n$. □

11.2 Graph containers

Theorem 11.2.1. 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 C 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 **supersaturation** 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: an independent set $I \subset V$.

Output: a "fingerprint" $S \subset I$ of size $\leq 2\delta |V|/d$, and a container $C \supset I$ 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 independent sets $I, I' \subset V$ 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 $S \subset V$. 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 lemma, 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

$$S \colon \mathcal{I} \to 2^V \qquad \text{and} \qquad A \colon 2^V \to 2^V$$

(one only needs to define $A(\cdot)$ on sets in the image of S)
such that, for every $I \in \mathcal{I}$,

- $S(I) \subset I \subset S(I) \cup A(S(I))$
- $|S(I)| \leq 2\delta |V|/d$
- $|S(I) \cup A(S(I))| \leq (1-\delta)|V|$

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

$$\Delta_\ell(H) = \max_{A \subset V(H) : |A| = \ell} \text{the number of edges containing } A$$

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

$$\Delta_1(H) \leq cd \qquad \text{and} \qquad \Delta_2(H) \leq c\sqrt{d}.$$

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 $I \subset V(H)$, a fingerprint $S \subset I$ 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.