9 Concentration of Measure

9.1 Bounded differences inequality

Recall that the Chernoff bound allows to prove exponential tail bounds for sums of **independent** random variables. For example, if Z is a sum of n independent Bernoulli random variables, then

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-2t^2/n}.$$

In this chapter, we develop tools for proving similar tail bounds for other random variables that do not necessarily arise as a sum of independent random variables.

The next theorem says:

> **A Lipschitz function of many *independent* random variables is concentrated.**

We will prove the following important and useful result, known by several names: **McDiarmid's inequality**, **Azuma–Hoeffding inequality**, and **bounded differences inequality**.

Theorem 9.1.1 (Bounded differences inequality)
Let $X_1 \in \Omega_1, \ldots, X_n \in \Omega_n$ be **independent** random variables. Suppose $f : \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ satisfies

$$\left|f(x_1, \ldots, x_n) - f(x_1', \ldots, x_n')\right| \leq 1 \tag{9.1}$$

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* 9.1.2** (Coupon collector)**.** Let $s_1, \ldots, s_n \in [n]$ chosen uniformly and inde-

pendently at random. Denote the number of "missing" elements by

$$Z = |[n] \setminus \{s_1, \dots, s_n\}|.$$

Note that changing one of the $s_1, \dots, s_n$ changes Z by at most 1, so we have

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

with

$$\mathbb{E}Z = n\left(1 - \frac{1}{n}\right)^n \in \left[\frac{n-1}{e}, \frac{n}{e}\right].$$

Theorem 9.1.1 holds more generally allowing the bounded difference to depend on the coordinate.

Theorem 9.1.3 (Bounded differences inequality)
Let $X_1 \in \Omega_1, \dots, X_n \in \Omega_n$ be **independent** random variables. Suppose $f : \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ satisfies

$$\left|f(x_1, \dots, x_n) - f(x_1', \dots, x_n')\right| \leq c_i \tag{9.2}$$

whenever $(x_1, \dots, x_n)$ and $(x_1, \dots, x_n)$ differ only on the i-th coordinate. Here $c_1, \dots, c_n$ are constants. Then the random variable $Z = f(X_1, \dots, X_n)$ satisfies, for every $\lambda \geq 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq \lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right)$$

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.

9.2 Martingales concentration inequalities

Definition 9.2.1
A ***martingale*** is a random real sequence $Z_0, Z_1, \dots$ such that for every Z_n, $\mathbb{E}|Z_n| < \infty$ and

$$\mathbb{E}[Z_{n+1}|Z_0, \dots, Z_n] = Z_n.$$

(To be more formal, we should talk about filtrations of a probability space . . .)

***Example* 9.2.2** (Random walks with independent steps)**.** If $(X_i)_{i \geq 0}$ is a sequence of

independent random variables with $\mathbb{E}X_i = 0$ for all i, then the partial sums $Z_n = \sum_{i \le n} X_i$ is a Martingale.

***Example* 9.2.3** (Betting strategy)**.** Betting on a sequence of fair coin tosses. After round, you are allow to change your bet. Let Z_n be your balance after the n-th round. Then Z_n is always a martingale regardless of your strategy.

Originally, the term "martingale" referred to the betting strategy where one doubles the bet each time until the first win and then stop betting. Then, with probability 1, $Z_n = 1$ for all sufficiently large n. (Why does this "free money" strategy not actually work?)

The next example is especially important to us.

***Example* 9.2.4** (Doob martingale)**.** Let $X_1, \dots, X_n$ be a random sequence (not necessarily independent, though they often are independent in practice). Consider a function $f(X_1, \dots, X_n)$. Let Z_i be the expected value of f after "revealing" (exposing) $X_1, \dots, X_i$, i.e.,

$$Z_i = \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_i].$$

So Z_i is the expected value of the random variable $Z = f(X_1, \dots, X_n)$ after seeing the first i arguments, and letting the remaining arguments be random. Then $Z_0, \dots, Z_n$ is a martingale (why?). It satisfies $Z_0 = \mathbb{E}Z$ (a non-random quantity) and $Z_n = Z$ (the random variable that we care about), and thereby offering a way to interpolate between the two.

***Example* 9.2.5** (Edge-exposure martingale)**.** We can reveal the random graph $G(n, p)$ by first fixing an order on all unordered pairs of $[n]$ and then revealing in order whether each pair is an edge. For any graph parameter $f(G)$ we can produce a martingale $X_0, X_1, \dots, X_{\binom{n}{2}}$ where Z_i is the conditional expectation of $f(G(n, p))$ after revealing whether there are edges for first i pairs of vertices. See Figure 9.1 for an example.

***Example* 9.2.6** (Vertex-exposure martingale)**.** Similar to the previous example, except that we now first fix an order on the vertex set, and, at the i-th step, with $0 \le i \le n$, we reveal all edges whose endpoints are contained in the first i vertices. See Figure 9.1 for an example.

Sometimes it is better to use the edge-exposure martingale and sometimes it is better to use the vertex-exposure martingale. It depends on the application. There is a trade-off between the length of the martingale and the control on the bounded differences.

The main result is that a martingale with *bounded differences* must be concentrated. The following fundamental result is called Azuma's inequality or the Azuma–Hoeffding inequality.

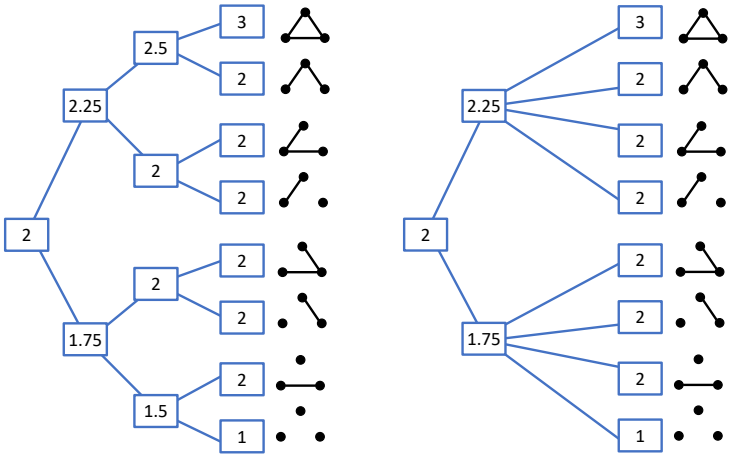


Figure 9.1: The edge-exposure martingale (left) and vertex-exposure martingale (right) for the chromatic number of $G(n, 1/2)$ with $n = 3$. The martingale is obtained by starting at the leftmost point, and splitting at each branch with equal probability.

Theorem 9.2.7 (Azuma's inequality)

Let $Z_0, Z_1, \dots, Z_n$ be a martingale satisfying

$$|Z_i - Z_{i-1}| \le 1 \quad \text{for each } i \in [n].$$

Then for every $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \ge \lambda\sqrt{n}) \le e^{-\lambda^2/2}.$$

Note that this is the same bound that we derived in Chapter 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 9.2.8 (Azuma's inequality)

Let $Z_0, Z_1, \dots, Z_n$ 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 can be used to recover the bounded differences inequality (Theorems 9.1.1 and 9.1.3) up to a usually unimportant constant in the exponent. To obtain the exact statement of Theorem 9.1.3, we state the following

strengthening of Azuma's inequality. (You are welcome to ignore the next statement if you do not care about the constant factor in the exponent — and really, you should not care.)

Theorem 9.2.9 (Azuma's inequality for Doob martingales)
Consider a Doob martingale $Z_i = \mathbb{E}[f(X_1, \dots, X_n) | X_1, \dots, X_i]$ as in Example 9.2.4. Suppose, conditioned on any value of $(X_1, \dots, X_{i-1})$, the possibilities for Z_i lies in an interval of length c_i (here c_i is non-random, but the location of the interval may depend on $X_1, \dots, X_{i-1}$). Then for any $\lambda > 0$,

$$\mathbb{P}(Z_n - Z_0 \geq \lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

***Remark* 9.2.10.** Applying the inequality to the martingale with terms $-Z_n$, we obtain the following lower tail bound:

$$\mathbb{P}(Z_n - Z_0 \leq -\lambda) \leq \exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

And we can put them together as

$$\mathbb{P}(|Z_n - Z_0| \geq \lambda) \leq 2\exp\left(\frac{-2\lambda^2}{c_1^2 + \cdots + c_n^2}\right).$$

***Remark* 9.2.11.** Theorem 9.2.8 is a special case of Theorem 9.2.9, since we can take $(X_1, \dots, X_n) = (Z_1 \dots, Z_n)$ and $f(X_1, \dots, X_n) = X_n$. Note that the $|Z_i - Z_{i-1}| \leq c_i$ condition in Theorem 9.2.8 implies that Z_i lies in an interval of length $2c_i$ if we condition on $(X_1, \dots, X_{i-1})$.

Lemma 9.2.12 (Hoeffding's lemma)
Let X be a real random variable contained in an interval of length ℓ. Suppose $\mathbb{E}X = 0$. Then

$$\mathbb{E}[e^X] \leq e^{\ell^2/8}.$$

Proof. Suppose $X \in [a, b]$ with $a \leq 0 \leq b$ and $b - a = \ell$. Then since e^x is convex, using a linear upper bound on the interval $[a, b]$, we have (note that RHS below is linear in x)

$$e^x \leq \frac{b - x}{b - a} e^a + \frac{x - a}{b - a} e^b, \quad \text{for all } x \in [a, b].$$

Since $\mathbb{E}X = 0$, we obtain

$$\mathbb{E}e^X \le \frac{b}{b-a}e^a + \frac{-a}{b-a}e^b.$$

Let $p = -a/(b-a)$. Then $a = -p\ell$ and $b = (1-p)\ell$. So

$$\log \mathbb{E}e^X \le \log\left((1-p)e^{-p\ell} + pe^{(1-p)\ell}\right) = -p\ell + \log(1-p+pe^\ell).$$

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 9.2.9. Let $t \ge 0$ be some constant to be decided later. Conditional on any values of $(X_1, \dots, X_{i-1})$, the random variable $Z_i - Z_{i-1}$ has mean zero and lies in an interval of length c_i. So Lemma 9.2.12 gives

$$\mathbb{E}[e^{t(Z_i - Z_{i-1})} | X_1, \dots, X_{i-1}] \le e^{t^2 c_i^2/8}.$$

Then the moment generating function satisfies

$$\begin{aligned}
\mathbb{E}[e^{t(Z_n - Z_0)}] &= \mathbb{E}\left[e^{t(Z_i - Z_{i-1})} e^{t(Z_{i-1} - Z_0)}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{t(Z_i - Z_{i-1})} \,\middle|\, X_1, \dots, X_{i-1}\right] e^{t(Z_{i-1} - Z_0)}\right] \\
&= e^{t^2 c_n^2/8}\mathbb{E}\left[e^{t(Z_{i-1} - Z_0)}\right].
\end{aligned}$$

Iterating, we obtain

$$\mathbb{E}\left[e^{t(Z_n - Z_0)}\right] \le e^{t^2(c_1^2 + \cdots c_n^2)/8}.$$

By Markov,

$$\mathbb{P}(Z_n - Z_0 \ge \lambda) \le e^{-t\lambda}\mathbb{E}\left[e^{t(Z_n - Z_0)}\right] \le e^{-t\lambda + \frac{t^2}{8}(c_1^2 + \cdots c_n^2)}.$$

Setting $t = 4\lambda/(c_1^2 + \cdots + c_n^2)$ yields the theorem. □

Now we apply Azuma's inequality to deduce the bounded differences inequality.

Proof of the bounded differences inequality (Theorem 9.1.3). Consider the Doob mar-

tingale $Z_i = \mathbb{E}[f(X_1, \dots, X_n)|X_1, \dots, X_i]$. The hypothesis of Theorem 9.1.3 implies that the hypothesis of Theorem 9.2.9 is satisfied. The same conclusion then follows. □

***Remark* 9.2.13.** Azuma's inequality (Theorem 9.2.9) is more versatile than (Theorem 9.1.3). For example, while changing X_i might change $f(X_1, \dots, X_n)$ by a lot in the worst case over all possible $(X_1, \dots, X_n)$, it might not change it by much in expectation over random choices of $(X_{i+1}, \dots, X_n)$. And so the c_i in Theorem 9.2.9 could potentially be smaller than in Theorem 9.1.3. This will be useful in some applications, including one that we will see later in the chapter.

9.3 Chromatic number of random graphs

Concentration of the chromatic number

Even before Bollobás (1988) showed that $\chi(G(n, 1/2)) \sim \frac{n}{2\log_2 n}$ whp (Theorem 8.3.2), using the bounded difference inequality, it was already known that the chromatic number of a random graph must be concentrated in a $O(\sqrt{n})$ window around its mean. The following application shows that one can prove concentration around the mean without even knowing where is the mean!

Theorem 9.3.1 (Shamir and Spencer 1987)
For every $\lambda \geq 0$, the chromatic number of a random graph $Z = \chi(G(n, p))$ 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 coordinates correspond to edges whose larger coordinate is i (cf. the vertex-exposure martingale Example 9.2.6). If two graphs G and G' differ only in edges incident to one vertex v, then $|\chi(G) - \chi(G')| \leq 1$ since, given a proper coloring of G using $\chi(G)$ colors, one can obtain a proper coloring of G' using $\chi(G) + 1$ colors by using a new color for v. Theorem 9.1.3 implies the result. □

***Remark* 9.3.2** (Non-concentration of the chromatic number)**.** Heckel (2021) showed that the $\chi(G(n, 1/2))$ is *not* concentrated on any interval of length n^c for any constant $c < 1/4$. This was the opposite of what most experts believed in. It has been conjectured that width of the window of concentrations fluctuates between $n^{1/4+o(1)}$ to $n^{1/2+o(1)}$ depending on n.

Clique number, again

Previously in Section 8.3, we used Janson inequalities to prove the following exponentially small bound on the probability that $G(n, 1/2)$ has small clique number. This was a crucial step in the proof of Bollobás' theorem (Theorem 8.3.2) that $\chi(G(n, 1/2)) \sim n/(2\log_2 n)$ whp. Here we give a different proof using the bounded difference inequality instead of Janson inequalities. The proof below in fact was the original approach of Bollobás (1988).

Lemma 9.3.3 (Same as Lemma 8.3.3)
Let $k_0 = k_0(n) \sim 2\log_2 n$ be the largest positive integer so that $\binom{n}{k_0} 2^{-\binom{k_0}{2}} \geq 1$. Then

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

A naive approach might be to estimate the number of k-cliques in G (this is the approach taken with Janson inequalities. The issue is that this quantity can change too much when we modify one edge of G. We will use a more subtle function on graphs. Note that we only care about whether there exists a k-clique or not.

Proof. Let $k = k_0 - 3$. Let $Y = Y(G)$ be the maximum number of edge-disjoint set of k-cliques in G. Then as a function of G, Y changes by at most 1 if we change G by one edge. (Note that the same does not hold if we change G by one vertex, e.g., when G consists of many k-cliques glued along a common vertex.)

So by the bounded differences inequality, for $G \sim G(n, 1/2)$,

$$\mathbb{P}(\omega(G) < k) = \mathbb{P}(Y = 0) \leq \mathbb{P}(Y - \mathbb{E}Y \leq -\mathbb{E}Y) \leq \exp\left(-\frac{2(\mathbb{E}Y)^2}{\binom{n}{2}}\right). \tag{9.1}$$

It remains to show that $\mathbb{E}Y \geq n^{2-o(1)}$. Create an auxiliary graph $\mathcal{H}$ whose vertices are the k-cliques in G, with a pair of k-cliques adjacent if they overlap in at least 2 vertices. Then $Y = \alpha(\mathcal{H})$. We would like to lower bound the independence number of this graph based on its average degree. Here are two ways to proceed:

1. Recall the Caro–Wei inequality (Corollary 2.3.5): for every graph H with average degree $\overline{d}$, 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.)

By a second moment argument (details again omitted, like in the proofs of Theorem 4.4.2 and Lemma 8.3.3), we have, with probability $1 - o(1)$, that the number of k-cliques in G is

$$|V(\mathcal{H})| \sim \mathbb{E}|V(\mathcal{H})| = \binom{n}{k} 2^{-\binom{k}{2}} = n^{3-o(1)}$$

and the number of unordered pairs of edge-overlapping k-cliques in G is

$$\mathbb{E}|E(\mathcal{H})| = n^{4-o(1)}.$$

Thus, with probability $1 - o(1)$, we can apply either of the above lower bounds on independent sets to obtain

$$\mathbb{E}Y \gtrsim \mathbb{E}\frac{|V(\mathcal{H})|^2}{|E(\mathcal{H})|} \gtrsim \mathbb{E}\frac{n^{6-o(1)}}{|E(\mathcal{H})|} \geq \frac{n^{6-o(1)}}{\mathbb{E}|E(\mathcal{H})|} = n^{2-o(1)}.$$

Together with (9.1), this completes the proof that $\mathbb{P}(\omega(G) < k) = e^{-n^{2-o(1)}}$. □

Chromatic number of sparse random graphs

Let us show that $G(n,p)$ is concentrated on a constant size window if p is small enough.

Theorem 9.3.4 (Shamir and Spencer 1987)
Let $\alpha > 5/6$ be fixed. Then for $p < n^{-\alpha}$, $\chi(G(n,p))$ is concentrated on four values with probability $1 - o(1)$. That is, there exists $u = u(n,p)$ such that, as $n \to \infty$,

$$\mathbb{P}(u \le \chi(G(n,p)) \le u + 3) = 1 - o(1).$$

Proof. It suffices to show that for all $\varepsilon > 0$, there exists $u = u(n,p,\varepsilon)$ so that, provided $p < n^{-\alpha}$ and n is sufficiently large,

$$\mathbb{P}(u \le \chi(G(n,p)) \le u + 3) \ge 1 - 3\varepsilon.$$

Let u be the least integer so that

$$\mathbb{P}(\chi(G(n,p)) \le u) > \varepsilon.$$

Now we make a clever choice of a random variable.

Let $G \sim G(n,p)$. Let $Y = Y(G)$ denote the minimum size of a subset $S \subseteq V(G)$ such that $G - S$ is u-colorable.

Note that Y changes by at most 1 if we change the edges around one vertex of G. Thus, by applying Theorem 9.1.1 with respect to vertex-exposure (Example 9.2.6), we have

$$\mathbb{P}(Y \le \mathbb{E}Y - \lambda\sqrt{n}) \le e^{-2\lambda^2}$$
$$\text{and} \qquad \mathbb{P}(Y \ge \mathbb{E}Y + \lambda\sqrt{n}) \le e^{-2\lambda^2}.$$

We choose $\lambda = \lambda(\varepsilon) > 0$ so that $e^{-2\lambda^2} = \varepsilon$.

First, we use the lower tail bound to show that $\mathbb{E}Y$ must be small. We have

$$e^{-2\lambda^2} = \varepsilon < \mathbb{P}(\chi(G) \le u) = \mathbb{P}(Y = 0) = \mathbb{P}(Y \le \mathbb{E}Y - \mathbb{E}Y) \le \exp\left(\frac{-2(\mathbb{E}Y)^2}{n}\right).$$

Thus

$$\mathbb{E}Y \le \lambda\sqrt{n}.$$

Next, we apply the upper tail bound to show that Y is rarely large. We have

$$\mathbb{P}(Y \ge 2\lambda\sqrt{n}) \le \mathbb{P}(Y \ge \mathbb{E}Y + \lambda\sqrt{n}) \le e^{-2\lambda^2} = \varepsilon.$$

Each of the following three events occur with probability at least $1 - \varepsilon$, for large enough n,

- By the above argument, there is some $S \subseteq V(G)$ with $|S| \le 2\lambda\sqrt{n}$ and $G - S$

may be properly u-colored.

- By the next lemma, one can properly 3-color $G[S]$.
- $\chi(G) \geq u$ (by the minimality of u at the beginning of the proof).

Thus, with probability at least $1 - 3\varepsilon$, all three events occur, and so we have $u \leq \chi(G) \leq u + 3$. □

Lemma 9.3.5

Fix $\alpha > 5/6$ and C. Let $p \leq n^{-\alpha}$. Then with probability $1 - o(1)$ every subset of at most $C\sqrt{n}$ vertices of $G(n,p)$ can be properly 3-colored.

Proof. Let $G \sim G(n,p)$. Assume that G is not 3-colorable. Choose minimum size $T \subseteq V(G)$ so that the induced subgraph $G[T]$ is not 3-colorable.

We see that $G[T]$ has minimum degree at least 3, since if $\deg_{G[T]}(x) < 3$, then $T - x$ cannot be 3-colorable either (if it were, then can extend coloring to x), contradicting the minimality of T.

Thus $G[T]$ has at least $3|T|/2$ edges. The probability that G has some induced subgraph on $t \leq C\sqrt{n}$ vertices and $\geq 3t/2$ edges is, by a union bound, (recall $\binom{n}{k} \leq (ne/k)^k$)

$$\leq \sum_{t=4}^{C\sqrt{n}} \binom{n}{t} \binom{\binom{t}{2}}{3t/2} p^{3t/2} \leq \sum_{t=4}^{C\sqrt{n}} \left(\frac{ne}{t}\right)^t \left(\frac{te}{3}\right)^{3t/2} n^{-3t\alpha/2}$$

$$\leq \sum_{t=4}^{C\sqrt{n}} \left(O(n^{1-3\alpha/2}\sqrt{t})\right)^t \leq \sum_{t=4}^{C\sqrt{n}} \left(O(n^{1-3\alpha/2+1/4})\right)^t.$$

The sum is $o(1)$ provided that $\alpha > 5/6$. □

***Remark* 9.3.6.** Theorem 9.3.4 was subsequently improved (by a refinement of the above techniques) by Łuczak (1991) and Alon and Krivelevich (1997). We now know that the chromatic number of $G(n, n^{-\alpha})$ has two-point concentration for all $\alpha > 1/2$.

9.4 Isoperimetric inequalities: a geometric perspective

We shall explore the following connection, which are two sides of the same coin:

Probability	*Geometry*
Concentration of Lipschitz functions	Isoperimetric inequalities

Milman recognized the importance of the ***concentration of measure phenomenon***, which he heavily promoted in the 1970's. The subject 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.

Euclidean space

The classic isoperimetric theorem in $\mathbb{R}^n$ says that that among all subset of $\mathbb{R}^n$ of given volume, the ball has the smallest surface volume. (The word "isoperimetric" refers to fixing the perimeter; equivalently we fix the surface area and ask to maximize volume.) This result (at least in two-dimensions) was known to the Greeks, but rigorous proofs were only found in towards the end of the nineteenth century.

Let (X, d_X) be a metric space. Let $A \subseteq X$. For any $x \in X$, write $d_X(x, A) := \inf_{a \in A} d_X(x, a)$ for the distance from x to A. Denote the set of all points within distance t from A by

$$A_t := \{x \in X : d_X(x, A) \le t\} \tag{9.1}$$

This is also known as the ***radius-t neighborhood of A***. One can visualize A_t by "expanding" A by distance t.

Theorem 9.4.1 (Isoperimetric inequality in Euclidean space)
Let $A \subseteq \mathbb{R}^n$ be a measurable set, and let $B \subseteq \mathbb{R}^n$ be a ball $\operatorname{vol}(A) = \operatorname{vol}(B)$. Then, for all $t \ge 0$,

$$\operatorname{vol} A_t \ge \operatorname{vol} B_t.$$

***Remark* 9.4.2.** A clean way to prove the above inequality is via the Brunn–Minkowski theorem.

Classically, the isoperimetric inequality is stated as (here ∂A is the boundary of A)

$$\operatorname{vol}_{n-1} \partial A \ge \operatorname{vol}_{n-1} \partial B.$$

These two formulations are equivalent. Indeed, assuming Theorem 9.4.1, we have

$$\begin{aligned} \operatorname{vol}_{n-1} \partial A = \left.\frac{d}{dt}\right|_{t=0} \operatorname{vol}_n A_t &= \lim_{t \to 0} \frac{\operatorname{vol} A_t - \operatorname{vol} A}{t} \\ &\ge \lim_{t \to 0} \frac{\operatorname{vol} B_t - \operatorname{vol} B}{t} = \operatorname{vol}_{n-1} \partial B. \end{aligned}$$

Conversely, we can obtain the neighborhood version from the boundary version by integrating (noting that B_t is always a ball).

The cube

We have an analogous result in the $\{0,1\}^n$ with respect to Hamming distance.In Hamming cube, ***Harper's theorem*** gives the exact result. Below, for $A \subseteq \{0,1\}^n$, we write A_t as in (9.1) for $X = \{0,1\}^n$ and d_X being the Hamming distance.

Theorem 9.4.3 (Isoperimetic inequality in the Hamming cube; Harper 1966)
Let $A \subseteq \{0,1\}^n$. Let $B \subseteq \{0,1\}^n$ be a Hamming ball with $|A| \geq |B|$. Then for all $t \geq 0$,

$$|A_t| \geq |B_t|.$$

***Remark* 9.4.4.** The above statement is tight when A has the same size as a Hamming ball, i.e., when $|A| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k}$ for some integer k. Actually, more is true. For any value of $|A|$ and t, the size of A_t is minimized by taking A to be an initial segment of $\{0,1\}^n$ according to the *simplicial ordering*: first sort by Hamming weight, and for ties, sort by lexicographic order. For more on this topic, particularly extremal set theory, see the book *Combinatorics* by Bollobás (1986).

Combined with the 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 9.4.5 (Rapid expansion from half to $1 - \varepsilon$)
Let $t > 0$. For every $A \subseteq \{0,1\}^n$ with $|A| \geq 2^{n-1}$, we have

$$|A_t| > (1 - e^{-2t^2/n})2^n.$$

Proof. Let $B = \{x \in \{0,1\}^n : \text{weight}(x) < n/2\}$, so that $|B| \leq 2^{n-1} \leq |A|$. Then by Harper's theorem (Theorem 9.4.3),

$$|A_t| \geq |B_t| = |\{x \in \{0,1\}^n : \text{weight}(x) < n/2 + t\}| > (1 - e^{-2t^2/n})2^n$$

by the Chernoff bound. □

In fact, using the above, we can deduce that even if we start with a small fraction (e.g., 1%) of the cube, and expand it slightly, then we would cover most of the cube.

Theorem 9.4.6 (Rapid expansion from ε to $1-\varepsilon$)
Let $\varepsilon > 0$ and $C = \sqrt{2\log(1/(\varepsilon))}$. If $A \subseteq \{0,1\}^n$ with $|A| \geq \varepsilon 2^n$, then

$$\left|A_{C\sqrt{n}}\right| \geq (1-\varepsilon)2^n.$$

First proof via Harper's isoperimetric inequality. Let $t = \sqrt{\log(1/\varepsilon)n/2}$ so that $e^{-2t^2/n} = \varepsilon$. Applying Theorem 9.4.5 to $A' = \{0,1\}^n \setminus A_t$, we see that $|A'| < 2^{n-1}$ (or else $\left|A'_t\right| > (1-\varepsilon)2^n$, so A'_t would intersect A, which is impossible since the distance between A and A' is greater than t). Thus $|A_t| \geq 2^{n-1}$, and then applying Theorem 9.4.5 yields $|A_{2t}| \geq (1-\varepsilon)2^n$. □

Let us give another proof of Theorem 9.4.6 without using Harper's exact isoperimetric theorem in the Hamming cube, and instead use the bounded differences inequality that we proved earlier.

Second proof via the bounded differences inequality. Pick a uniform random $x \in \{0,1\}^n$ and let $X = \operatorname{dist}(x, A)$. Note that X changes by at most 1 if a single coordinate of x is changed. Applying the bounded differences inequality, Theorem 9.1.1, we have the lower tail

$$\mathbb{P}(X - \mathbb{E}X \leq -\lambda) \leq e^{-2\lambda^2/n} \quad \text{for all } \lambda \geq 0$$

We have $X = 0$ if and only if $x \in A$, so

$$\varepsilon \leq \mathbb{P}(x \in A) = \mathbb{P}(X = 0) = \mathbb{P}(X - \mathbb{E}X \leq -\mathbb{E}X) \leq e^{-2(\mathbb{E}X)^2/n}.$$

Thus

$$\mathbb{E}X \leq \sqrt{\frac{\log(1/\varepsilon)n}{2}} = \frac{C\sqrt{n}}{2}.$$

Now we apply the upper tail of the bounded differences inequality

$$\mathbb{P}(X - \mathbb{E}X \geq \lambda) \leq e^{-2\lambda^2/n} \quad \text{for all } \lambda \geq 0$$

to yield

$$\mathbb{P}(x \notin A_{C\sqrt{n}}) = \mathbb{P}(X > C\sqrt{n}) \leq \mathbb{P}\left(X \geq \mathbb{E}X + \frac{C\sqrt{n}}{2}\right) \leq \varepsilon. \qquad \square$$

Isoperimetry versus concentration

The above two proofs illustrate the link between geometric isoperimetric inequalities and probabilistic concentration inequalities. Let know now state a simple result that

formalizes this connection.

Definition 9.4.7 (Lipschitz functions)
Given two metric spaces (X, d_X) and (Y, d_Y), we say that a function $f: X \to Y$ is ***C-Lipschitz*** if

$$d_Y(f(x), f(x')) \le C d_X(x, x') \qquad \text{for all } x, x' \in X.$$

So the bounded differences inequality applies to Lipschitz functions with respect to the Hamming distance. In particular, it tells us that if $f: \{0,1\}^n \to \mathbb{R}$ is 1-Lipschitz (with respect to the Hamming distance on $\{0,1\}^n$), it must be concentrated around its mean with respect to the uniform measure on $\{0,1\}^n$:

$$\mathbb{P}(|f - \mathbb{E}f| \ge n\lambda) \le 2e^{-2n\lambda^2}.$$

So f is *almost constant almost everywhere*. This is a counterintuitive high dimensional phenomenon.

Theorem 9.4.8 (Equivalence between notions of concentration of measure)
Let $t, \varepsilon \ge 0$. In a probability space $(\Omega, \mathbb{P})$ equipped with a metric. The following are equivalent:

(a) (Expansion/approximate isoperimetry) If $A \subseteq \Omega$ with $\mathbb{P}(A) \ge 1/2$, then

$$\mathbb{P}(A_t) \ge 1 - \varepsilon.$$

(b) (Concentration of Lipschitz functions) If $f: \Omega \to \mathbb{R}$ is 1-Lipschitz and $m \in \mathbb{R}$ satisfies $\mathbb{P}(f \le m) \ge 1/2$, then

$$\mathbb{P}(f > m + t) \le \varepsilon.$$

Remark* 9.4.9** (Median)**.** In (b), we often take m to be a ***median of f, which is defined to be a value such that $\mathbb{P}(f \ge m) \ge 1/2$ and $\mathbb{P}(f \le m) \ge 1/2$ (the median always exists but is not necessarily unique). For distributions with good concentration properties, the median and mean are usually close to each other. For example, we leave it as an exercise to check that if there is some m such that $\mathbb{P}(|f - m| \ge t) \le 2e^{-t^2/2}$ for all $t \ge 0$, then the mean and the medians of f all lie within $O(1)$ of m.

Proof. (a) $\implies$ (b): Let $A = \{x \in \Omega : f(x) \le m\}$. So $\mathbb{P}(A) \ge 1/2$. Since f is 1-Lipschitz, we have $f(x) \le m + t$ for all $x \in A_t$. Thus by (a)

$$\mathbb{P}(f > m + t) \le \mathbb{P}(\overline{A_t}) \le \varepsilon.$$

(b) $\implies$ (a): Let $f(x) = \operatorname{dist}(x, A)$ and $m = 0$. Then $\mathbb{P}(f \le 0) = \mathbb{P}(A) \ge 1/2$. Also f is 1-Lipschitz. So by (b),

$$\mathbb{P}(\overline{A_t}) = \mathbb{P}(f > m + t) \le \varepsilon. \qquad \square$$

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.

The sphere

We discuss analogs of the concentration of measure phenomenon in high dimensional geometry. This is rich and beautiful subject. An excellent introductory to this topic is the survey *An Elementary Introduction to Modern Convex Geometry* by Ball (1997).

Recall the isoperimetric inequality in $\mathbb{R}^n$ says:

> If $A \subseteq \mathbb{R}^n$ has the same measure as ball B, then $\operatorname{vol}(A_t) \ge \operatorname{vol}(B_t)$ for all $t \ge 0$.

Analogous exact isoperimetric inequalities are known in several other spaces. We already saw it for the boolean cube (Theorem 9.4.3). The case of sphere and Gaussian space are particularly noteworthy. The following theorem is due to Lévy (~1919).

Theorem 9.4.10 (Lévy's isoperimetric inequality on the sphere)
On a sphere in $\mathbb{R}^n$, let A be a measurable subset and B a spherical cap with $\operatorname{vol}_{n-1}(A) = \operatorname{vol}_{n-1}(B)$. Then for all $t \ge 0$,

$$\operatorname{vol}_{n-1}(A_t) \ge \operatorname{vol}_{n-1}(B_t).$$

We have the following upper bound estimate on the size of spherical caps.

Theorem 9.4.11 (Upper bound on spherical cap size)
Let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ be a uniform random unit vector in $\mathbb{R}^n$. Then for any $\varepsilon \ge 0$,

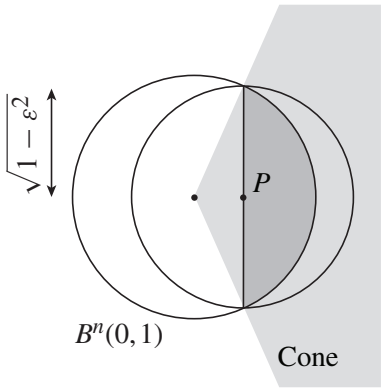
$$\mathbb{P}(x_1 \ge \varepsilon) \le e^{-n\varepsilon^2/2}.$$

The following proof (including figures) is taken from Tokz (2012), building on the method by Ball (1997).

Proof. Let C denote the spherical cap consisting of unit vectors x with $x_1 \geq \varepsilon$. Write $\widetilde{C}$ for the convex hull of C with the origin, i.e., the conical sector spanned by C. The idea is to contain $\widetilde{C}$ in a ball of radius $r \leq e^{-\varepsilon^2/2}$. Writing $B(r)$ for a ball of radius r in $\mathbb{R}^n$ so that, we have

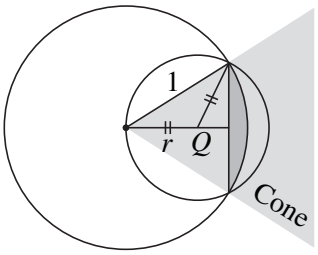
$$\frac{\operatorname{vol}_{n-1} C}{\operatorname{vol}_{n-1} S^{n-1}} = \frac{\operatorname{vol}_n \widetilde{C}}{\operatorname{vol}_n B_n(1)} = \frac{\operatorname{vol}_n B(r)}{\operatorname{vol}_n B(1)} = r^n \leq e^{-\varepsilon^2 n/2}.$$

Case 1: $\varepsilon \in [0, 1/\sqrt{2}]$.



As shown above, $\widetilde{C}$ is contained in a ball of radius $r = \sqrt{1-\varepsilon^2} \leq e^{-\varepsilon^2/2}$.

Case 2: $\varepsilon \in [1/\sqrt{2}, 1]$.



Then $\widetilde{C}$ is contained in a ball of radius r as shown above. Using similar triangles, we find that $r/(1/2) = 1/\varepsilon$. So $r = 1/(2\varepsilon) \leq e^{-\varepsilon^2/2}$, where final inequality is equivalent to $e^{x^2/2} \leq 2x$ for all $[1/\sqrt{2}, 1]$, which, by convexity, only needs to be checked at the endpoints of the interval. □

Combining the above two theorems, we deduce the following concentration of measure results.

Corollary 9.4.12 (Concentration of measure on the sphere)
Let A be a measurable subset of the unit sphere in $\mathbb{R}^n$, equipped with the metric inherited from $\mathbb{R}^n$. If $A \subseteq S^{n-1}$ has $\mathrm{vol}_{n-1}(A)/\mathrm{vol}_{n-1}(S^{n-1}) \geq 1/2$, then

$$\frac{\mathrm{vol}_{n-1}(A_t)}{\mathrm{vol}_{n-1}(S^{n-1})} \geq 1 - e^{-nt^2/4}.$$

***Remark* 9.4.13.** See §14 in Barvinok's notes for a proof of the sharper estimate with $e^{-nt^2/4}$ replaced by $\sqrt{\pi/8}e^{-nt^2/2}$, where now we are using the geodesic distance on the sphere.

Corollary 9.4.14 (Concentration of measure on the sphere)
Let S^{n-1} denote the unit sphere in $\mathbb{R}^n$. If $f\colon S^{n-1} \to \mathbb{R}$ is a 1-Lipschitz measurable function, then there is some real m so that, for the uniform measure on the sphere,

$$\mathbb{P}(|f - m| > t) \leq 2e^{-nt^2/4}.$$

Informally: *every Lipschitz function on a high dimensional sphere is almost constant almost everywhere.*

This is a rather counterintuitive high-dimensional phenomenon.

Gauss space

Another related setting is the ***Gauss space***, which is $\mathbb{R}^n$ equipped with the the probability measure γ_n induced by the Gaussian random vector whose coordinates are n iid standard normals, i.e., the normal random vector in $\mathbb{R}^n$ with covariance matrix I_n. Its probability density function of γ_n at $x \in \mathbb{R}^n$ is $(2\pi)^{-n}e^{-|x|^2/2}$. The metric on $\mathbb{R}^n$ is the usual Euclidean metric.

What would an isoperimetric inequality in Gauss space look like?

Although earlier examples of isoperimetric optimizers were all balls, for the Gauss space, the answer is actually a ***half-spaces***, i.e., points on one side of some hyperplane.

The Gaussian isoperimetric inequality, below, was first shown independently by Borell (1975) and Sudakov and Tsirel'son (1974).

Theorem 9.4.15 (Gaussian isoperimetric inequality)
If $A, H \subseteq \mathbb{R}^n$, H a half-space, and $\gamma(A) = \gamma(H)$, then $\gamma(A_t) \geq \gamma(H_t)$ for all $t \geq 0$, where γ is the Gauss measure.

If $H = \{x_1 \le 0\}$, then $H_t = \{x_1 \le t\}$, which has Gaussian measure $\ge 1 - e^{-t^2/2}$. Thus:

Corollary 9.4.16 (Concentration of measure for Gaussian vectors)
If $f : \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz, and Z is a vector of i.i.d. standard normals, then $X = f(Z)$ satisfies, for some m,

$$\mathbb{P}(|X - m| \ge t) \le 2e^{-t^2/2}.$$

Here is a rather handwavy explanation why the half-space is a reasonable answer.

Consider $\{-1, 1\}^{mn}$, where both m and n are large. Let us group the coordinates of $\{-1, 1\}^{mn}$ into block of length m. The sum of entries in each block (after normalizing by $\sqrt{m}$) approximates normal random variable by the central limit theorem.

In the Hamming cube, Harper's theorem tells us Hamming balls are isoperimetric optimizers. Since a Hamming ball in $\{-1, 1\}^{mn}$ is given by all points whose sum of coordinates is below a certain threshold, we should look at the analogous subset in the Gauss space, which would then consist of all points whose sum of coordinates is below a certain threshold. The set of all points whose of coordinate sum is below a certain threshold is half-space. Note also that the Gaussian measure is radially symmetric.

The sphere as approximately a sum of independent Gaussians. The Gauss space is a nice space to work with because a standard normal vector simultaneously possesses two useful properties (and it is essentially the only such random vector to have both properties):

(a) Rotational invariance

(b) Independence of coordinates

The squared-length of a random Gaussian vector is $Z_1^2 + \cdots + Z_n^2$ with iid $Z_1, \ldots, Z_n \in N(0, 1)$. It has mean n and a $O(\sqrt{n})$ window of concentration (e.g., by a straightforward adaptation of the Chernoff bound proof). Since $\sqrt{n + O(\sqrt{n})} = \sqrt{n} + O(1)$, the length of Gaussian vector is concentrated in a $O(1)$ window around $\sqrt{n}$ (the concentration can also be deduced from the above corollary for $f(x) = |x|$). So most of the distribution in the Gauss space lies within a constant distance of a sphere of radius $\sqrt{n}$. Due to rotational invariance, we see that a Gaussian distribution approximates the uniform distribution on sphere of radius $\sqrt{n}$ in high dimensions. In other words:

$$\text{random Gaussian vector} \approx \sqrt{n} \cdot \text{ random unit vector}.$$

Random Gaussian vectors often yield easier calculations due to coordinate independence, and so they often give an accessible way to analyze random unit vectors.

Note that how a *half-space* in the Gauss space intersect the sphere in a *spherical cap*, with both italicized objects being isoperimetric optimizers in their respective spaces.

Sub-Gaussian distributions

We introduce some terminology that captures notions we have seen so far. It will also be convenient for later discussions.

Definition 9.4.17 (Sub-Gaussian distribution)
We say that a random variable X is ***K-subGaussian about its mean*** if

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-t^2/K^2} \quad \text{for all } t \geq 0.$$

***Remark* 9.4.18.** This definition is not standard. Some places say σ^2-*subGaussian* for what we mean by σ-*subGaussian*.

Usually we will not worry about constant factors. Thus, saying that a family of random variables X_n is $O(K_n)$-subGaussian about its mean is the same as saying that there exist constant $C, c > 0$ such that

$$\mathbb{P}\left(|X_n - \mathbb{E}X_n| \geq t\right) \leq Ce^{-ct^2/K_n^2} \quad \text{for all } t \geq 0 \text{ and } n.$$

Also note that, up to changing the constants c, C, the definition does not change if we replace $\mathbb{E}X_n$ by a median of X_n above.

***Example* 9.4.19.** The concentration inequalities so far can be rephrased in terms of subGaussian distributions. Below is summary of results of the form: if X is a random point drawn from the given space, and f is a 1-Lipschitz function, then $f(X)$ is K-subGaussian.

space	distance	-subGaussian	reference
$\{0,1\}^n$	Hamming	$O(\sqrt{n})$	bounded diff. ineq. (Thm. 9.1.1)
S^{n-1}	Euclidean	$O(1/\sqrt{n})$	Lévy concentration (Cor. 9.4.14)
Gauss space $\mathbb{R}^n$	Euclidean	$O(1)$	Gaussian isoperimetric ineq. (Cor. 9.4.16)

The following lemma shows that for subGaussian random variables, it does not matter much if we define the tails around its median, mean, or root-mean-square.

Lemma 9.4.20 (Median vs. mean for subGaussian distributions)
There exists a constant $C > 0$ so that the following holds for any real random variable X satisfying, for some constants m and K,

$$\mathbb{P}(|X - m| \geq t) \leq 2e^{-t^2/K^2} \quad \text{for all } t \geq 0.$$

(a) Every median $\mathbb{M}X$ of X satisfies

$$|\mathbb{M}X - m| \leq CK.$$

(b) The mean of X satisfies

$$|\mathbb{E}X - m| \leq CK.$$

(c) For any $p \geq 1$, writing $\|X\|_p := (\mathbb{E}\,|X|^p)^{1/p}$ for the L^p norm of X,

$$\left|\|X\|_p - m\right| \leq CK\sqrt{p}.$$

(d) For every constant A there exists a constant $c > 0$ so that if $|m' - m| \leq AK$, then

$$\mathbb{P}(|X - m'| \geq t) \leq 2e^{-ct^2/K^2} \quad \text{for all } t \geq 0.$$

Proof. By considering X/K instead of X, we may assume that $K = 1$ for convenience.

(a) For any $t > \sqrt{2\log 2}$, we have $\mathbb{P}(|X - m| \geq t) \leq 2e^{-t^2} < 1/2$. So every median of X lies within $\sqrt{2\log 2}$ of m.

(b) We have

$$\begin{aligned}|\mathbb{E}X - m| \leq \mathbb{E}\,|X - m| &= \int_0^\infty \mathbb{P}(|X - m| \geq t)\,dt \\ &\leq \int_0^\infty 2e^{-t^2}\,dt = \sqrt{\pi}.\end{aligned}$$

(c) Using the triangle inequality on the L^p norm, we have

$$\begin{aligned}\left|\|X\|_p - m\right| \leq \|X - m\|_p = (\mathbb{E}\,|X - m|^p)^{1/p} &= \left(\int_0^\infty \mathbb{P}(|X - m|^p \geq t)\,dt\right)^{1/p} \\ \leq \left(\int_0^\infty 2e^{-t^{2/p}}\,dt\right)^{1/p} &= 2^{1/p}\Gamma\left(1 + \frac{p}{2}\right)^{1/p} = O(\sqrt{p}).\end{aligned}$$

(c) We can make c small enough so that $RHS = 2e^{-ct^2} \geq 1$ for $t \leq 2A$. For $t > 2A$,

we note that

$$\mathbb{P}(|X - m'| \geq t) \leq \mathbb{P}(|X - m| \geq t/2) \leq 2e^{-t^2/4}. \qquad \square$$

***Remark* 9.4.21** (Equivalent characterization of subGaussian distributions)**.** Given a real random variable X, if any of the below is true for some K_i, then the other conditions are true for some $K_j \leq CK_i$ for some absolute constant C.

(a) *(Tails)* $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/K_1^2}$ for all $t \geq 0$.

(b) *(Moments)* $\|X\|_{L^p} \leq K_2\sqrt{p}$ for all $p \geq 1$.

(c) *(MGF of X^2)* $\mathbb{E}e^{X^2/K_3^2} \leq 2$.

We leave the proofs as exercises. Also see §2.5.1 in the textbook *High-Dimensional Probability* by Vershynin (2018), which gives a superb introduction to the subject.

Johnson–Lindenstrauss Lemma

Given a set of N points in high-dimensional Euclidean space, the next result tells us that one can embed them in $O(\varepsilon^{-2} \log N)$ dimensions without sacrificing pairwise distances by more than $1 \pm \varepsilon$ factor. This is known as ***dimension reduction***. It is an important tool in many areas, from functional analysis to algorithms.

Theorem 9.4.22 (Johnson and Lindenstrauss 1982)
There exists a constant $C > 0$ so that the following holds. Let $\varepsilon > 0$. Let X be a set of N points in $\mathbb{R}^m$. Then for any $d > C\varepsilon^{-2} \log N$, there exists $f : X \to \mathbb{R}^d$ so that

$$(1 - \varepsilon)\,|x - y| \leq |f(x) - f(y)| \leq (1 + \varepsilon)\,|x - y| \quad \text{for all } x, y \in X.$$

***Remark* 9.4.23.** Here the requirement $d > C\varepsilon^{-2} \log N$ on the dimension is optimal up to a constant factor (Larsen and Nelson 2017).

We will take f to be $\sqrt{m/d}$ times an orthogonal projection onto a d-dimensional subspace chosen uniformly at random. The theorem then follows from the following lemma together with a union bound.

Lemma 9.4.24 (Random projection)
There exists a constant $C > 0$ so that the following holds. Let $m \geq d$ and let $P\colon \mathbb{R}^m \to \mathbb{R}^d$ denote the orthogonal projection onto the subspace spanned by the first d coordinates. Let z be a uniform random point on the unit sphere in $\mathbb{R}^m$. Let $y = Pz$ and $Y = |y|$. Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left|Y - \sqrt{\frac{d}{m}}\right| \geq t\right) \leq 2e^{-cmt^2}.$$

To prove the Theorem 9.4.22, for each pair of distinct points $x, x' \in X$, set

$$z = \frac{x - x'}{|x - x'|}, \qquad \text{so that } \sqrt{\frac{m}{d}}Y = \frac{|f(x) - f(x')|}{|x - x'|}.$$

Then the length of the projection of z onto a uniform random d-dimensional subspace has the same distribution as Y in the lemma. So setting $t = \varepsilon\sqrt{d/m}$, we find that

$$\mathbb{P}\left(\left|\sqrt{\frac{m}{d}}Y - 1\right| \geq \varepsilon\right) \leq 2e^{-c\varepsilon d} < 2N^{-cC}.$$

Provided that $C > 1/c$, we can take a union bound over all $\binom{N}{2} < N^2/2$ pairs of points of X to show that with some positive probability, the random f works.

Proof of the lemma. We have $z_1^2 + \cdots + z_n^2 = 1$ and each z_i has the same distribution, so $\mathbb{E}[z_i^2] = 1/m$ for each i. Thus

$$\mathbb{E}[Y^2] = \mathbb{E}\left[z_1^2 + \cdots + z_d^2\right] = \frac{d}{m}.$$

Note that P is 1-Lipschitz on the unit sphere. By Lévy's concentration measure theorem on the sphere, letting $\mathbb{M}Y$ denote the median of Y,

$$\mathbb{P}\left(|Y - \mathbb{M}Y| \geq t\right) \leq 2e^{-mt^2/4}.$$

The result then follows by Lemma 9.4.20, using that $\|Y\|_2 = \sqrt{d/m}$. □

Here is a cute application of Johnson–Lindenstrauss (this is related to a homework problem on the Chernoff bound).

Corollary 9.4.25
There is a constant $c > 0$ so that for every positive integer d, there is a set of $e^{c\varepsilon^2 d}$ points in $\mathbb{R}^d$ whose pairwise distances are in $[1 - \varepsilon, 1 + \varepsilon]$.

Proof. Applying Theorem 9.4.22 a regular simplex with unit edge lengths with N vertices in $\mathbb{R}^{N-1}$ to yield N points in $\mathbb{R}^d$ for $d = O(\varepsilon^{-2} \log N)$ and pairwise distances in $[1-\varepsilon, 1+\varepsilon]$. □

9.5 Talagrand's inequality

Talagrand (1995) developed a powerful concentration inequality. It is applicable to many combinatorial optimization problems on independent random inputs. The most general form of Talagrand's inequality can be somewhat difficult to grasp. So we start by discussing a special case with an easier geometric statement. Though, to obtain the full power of Talagrand's inequality with combinatorial consequences, we will need the full statement to be given later.

We omit the proof of Talagrand's inequality (see the Alon–Spencer textbook or Tao's blog post) and instead focus on explaining the theorem and its applications.

Distance to a subspace

We start with a geometrically motivated question.

Problem 9.5.1

Let V be a *fixed* d-dimensional subspace. Let $x \sim \mathrm{Unif}\{-1,1\}^n$. 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 $\sqrt{n-d}$?

Some easier special cases (codimension-1):

- If V is a 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 to $|Z|$ for $Z \sim N(0,1)$. In particular, it is $O(1)$-subGaussian.

- More generally, if for a hyperplane $V = \alpha^\perp$ for some unit vector $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$, one has $\operatorname{dist}(x, V) = |\alpha \cdot x|$. Note that flipping x_i changes $|\alpha \cdot x|$ by at most $2|\alpha_i|$. So by the bounded differences inequality Theorem 9.1.3, 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.

The bounded difference inequality applied to the function $x \in \{-1, 1\}^n \mapsto \operatorname{dist}(x, V)$, which is 2-Lipschitz (with respect to Hamming distance), gives

$$\mathbb{P}\left(|\operatorname{dist}(x, V) - \mathbb{E}\operatorname{dist}(x, V)| \geq t\right) \leq 2e^{-t^2/(2n)},$$

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).

If, instead of sampling $x \in \{-1, 1\}^n$, we took x to be a uniformly random point on the radius $\sqrt{n}$ sphere in $\mathbb{R}^n$ (which contains $\{-1, 1\}^n$), then Lévy concentration on the sphere (Corollary 9.4.14) implies that $\operatorname{dist}(x, V)$ is $O(1)$-subGaussian. Perhaps a similar bound holds when x is chosen from $\{-1, 1\}^n$?

Here is a corollary of Talagrand's inequality, which we will state in its general form later.

Theorem 9.5.2

Let V be a fixed d-dimensional subspace in $\mathbb{R}^n$. For uniformly random $x \in \{-1, 1\}^n$, one has

$$\mathbb{P}\left(|\operatorname{dist}(x, V) - \sqrt{n - d}| \geq t\right) \leq Ce^{-ct^2},$$

where $C, c > 0$ are some constants.

Convex Lipschitz functions of independent random variables

Let us now state Talagrand's inequality, first in a special case for convex functions, and then more generally. Below $\operatorname{dist}(\cdot,\cdot)$ means Euclidean distance.

Theorem 9.5.3 (Talagrand)
Let $A \subseteq \mathbb{R}^n$ be convex. Let $x \sim \operatorname{Unif}\{0,1\}^n$. Then for any $t \geq 0$,

$$\mathbb{P}(x \in A)\mathbb{P}(\operatorname{dist}(x, A) \geq t) \leq e^{-t^2/4}.$$

***Remark* 9.5.4.** (1) Note that A is a convex body in $\mathbb{R}^n$ and not simply a set of points in A.

(2) The bounded differences inequality gives us an upper bound of the form $e^{-ct^2/n}$, which is much worse than Talagrand's bound.

***Example* 9.5.5** (Talagrand's inequality fails for nonconvex sets)**.** Let

$$A = \left\{x \in \{0,1\}^n : \operatorname{wt}(x) \leq \frac{n}{2} - \sqrt{n}\right\}$$

(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 $\operatorname{dist}(y, A) \geq n^{1/4}$ (note that this is Euclidean distance and not Hamming distance). Using the central limit theorem, we have, for some constant $c > 0$ and sufficiently large n, for $x \sim \operatorname{Uniform}(\{-1,1\}^n)$, $\mathbb{P}(x \in A) \geq c$ and $\mathbb{P}(\operatorname{wt}(x) \geq n/2) \geq 1/2$, so the conclusion of Talagrand's inequality is false for $t = n^{1/4}$, in the case of this nonconvex A.

By an argument similar to our proof of Theorem 9.4.8 (the equivalence of notions of concentration of measure), one can deduce the following consequence.

Corollary 9.5.6 (Talagrand)
Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex and 1-Lipschitz function (with respect to Euclidean distance on $\mathbb{R}^n$). Let $x \sim \operatorname{Unif}\{0,1\}^n$. Then for any $r \in \mathbb{R}$ and $t \geq 0$,

$$\mathbb{P}(f(x) \leq r)\mathbb{P}(f(x) \geq r + t) \leq e^{-t^2/4}.$$

Remark* 9.5.7.** The proof below shows that the assumption that f is convex can be weakened to f being ***quasiconvex, i.e., $\{f \leq a\}$ is convex for every $a \in \mathbb{R}$.

Proof that Theorem 9.5.3 and Corollary 9.5.6 are equivalent. Theorem 9.5.3 implies Corollary 9.5.6: take $A = \{x : f(x) \leq r\}$. We have $f(x) \leq r + t$ whenever

$\operatorname{dist}(a, A) \le t$ since f is 1-Lipschitz. So $\mathbb{P}(f(x) \le r) = \mathbb{P}(x \in A)$ and $\mathbb{P}(f(x) \ge r + t) \le \mathbb{P}(\operatorname{dist}(x, A) \ge t)$.

Corollary 9.5.6 implies Theorem 9.5.3: $r = 0$ and take $f(x) = \operatorname{dist}(x, A)$, which is a convex function 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 9.5.8 (Talagrand)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and 1-Lipschitz function (with respect to Euclidean distance on $\mathbb{R}^n$). Let $x \sim \operatorname{Unif}(\{0,1\}^n)$. Then

$$\mathbb{P}(|f(x) - \mathbb{M}f(x)| \ge t) \le 4e^{-t^2/4}.$$

Proof. Setting $r = \mathbb{M}f(x)$ in Corollary 9.5.6 yields

$$\mathbb{P}(f(x) \ge \mathbb{M}f(x) + t) \le 2e^{-t^2/4}.$$

Setting $r = \mathbb{M}f(x) - t$ in Corollary 9.5.6 yields

$$\mathbb{P}(f(x) \le \mathbb{M}f(x) - t) \le 2e^{-t^2/4}.$$ □

Combining the two tail bounds yields the corollary.

Theorem 9.5.2 then follows. Indeed, Corollary 9.5.8 shows that $\operatorname{dist}(x, V)$ (which is a convex 1-Lipschitz function of $x \in \mathbb{R}^n$) is $O(1)$-subGaussian, which immediately implies Theorem 9.5.2.

***Example* 9.5.9** (Operator norm of a random matrix)**.** Let A be a random matrix whose entries are uniform iid from $\{-1, 1\}$. Viewing $A \mapsto \|A\|_{\mathrm{op}}$ 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 $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_{\mathrm{HS}}$, where the latter is the Hilbert–Schmidt norm, also known as the Frobenius norm, i.e., the ℓ_2-norm of the matrix entries). It follows by Talagrand's inequality (Corollary 9.5.8) that $\|A\|_{\mathrm{op}}$ is $O(1)$-subGaussian about its mean.

Convex distance

Talagrand's inequality has a much more general form, which has far-reaching combinatorial applications. We need 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

$$\boldsymbol{d_\alpha(x,y)} := \sum_{i : x_i \neq y_j} \alpha_i$$

For $A \subseteq \Omega$,

$$\boldsymbol{d_\alpha(x,A)} := \inf_{y \in A} d_\alpha(x, y).$$

Talagrand's ***convex distance*** between $x \in \Omega$ and $A \subseteq \Omega$ is defined by

$$\boldsymbol{d_T(x,A)} := \sup_{\substack{\alpha \in \mathbb{R}^n_{\ge 0} \\ |\alpha| = 1}} d_\alpha(x, A).$$

Here $|\alpha|$ denotes Euclidean length:

$$|\alpha|^2 := \alpha_1^2 + \cdots + \alpha_n^2.$$

***Example* 9.5.10** (Euclidean distance to convex hull)**.** If $A \subseteq \{0,1\}^n$ and $x \in \{0,1\}^n$, then $d_T(x, A)$ is the Euclidean distance from x to the convex hull of A.

Let us give another interpretation of convex distance. For $x, y \in \Omega$, let

$$\phi_x(y) = (1_{x_1 \neq y_1}, 1_{x_2 \neq y_2}, \dots, 1_{x_n \neq y_n}) \in \{0,1\}^n$$

be the vector of coordinatewise disagreements between x and y. Write

$$\phi_x(A) = \{\phi_x(y) : y \in A\} \subseteq \{0,1\}^n.$$

Then for any $\alpha \in \mathbb{R}^n_{\ge 0}$,

$$d_\alpha(x, A) = d_\alpha(\vec{0}, \phi_x(A)),$$

where the LHS is the weighted Hamming distance in Ω whereas the RHS takes place in $\{0,1\}^n$. Taking the supremum over $\alpha \in \mathbb{R}^n_{\ge 0}$ with $|\alpha| = 1$, and using the Example 9.5.10, we deduce

$$d_T(x, A) = \operatorname{dist}(\vec{0}, \operatorname{ConvexHull} \phi_x(A)).$$

The general form of Talagrand's inequality says the following. Note that it reduces to the earlier special case Theorem 9.5.3 if $\Omega = \{0,1\}^n$.

Theorem 9.5.11 (Talagrand's inequality: general form)
Let $A \subseteq \Omega = \Omega_1 \times \cdots \times \Omega_n$, with Ω equipped with a product probability measure. Let $x \in \Omega$ be chosen randomly with independent coordinates. Let $t \geq 0$. Then

$$\mathbb{P}(x \in A)\mathbb{P}(d_T(x, A) \geq t) \leq e^{-t^2/4}.$$

Let us see how Talagrand's inequality recovers a more general form of our geometric inequalities from earlier, extending from independent boolean random variables to independent bounded random variables.

Lemma 9.5.12 (Convex distance upper bounds Euclidean distance)
Let $A \subseteq [0,1]^n$ and $x \in [0,1]^n$. Then $\operatorname{dist}(x, \text{ConvexHull } A) \leq d_T(x, A)$.

Proof. For any $\alpha \in \mathbb{R}^n$, and any $y \in [0,1]^n$, we have

$$|(x - y) \cdot \alpha| \leq \sum_{i=1}^{n} |\alpha_i| \, |x_i - y_i| \leq \sum_{i : x_i \neq y_i} |\alpha_i| \, .$$

First taking the infimum over all $y \in A$, and then taking the supremum over unit vectors α, the LHS becomes $\operatorname{dist}(x, \text{ConvexHull } A)$ and the RHS becomes $d_T(x, A)$. □

Corollary 9.5.13 (Talagrand's inequality: convex sets and convex Lipschitz functions)
Let $x = (x_1, \ldots, x_n) \in [0,1]^n$ be independent random variables (not necessarily identical). Let $t \geq 0$. Let $A \subseteq [0,1]^n$ be a convex set. Then

$$\mathbb{P}(x \in A)\mathbb{P}(\operatorname{dist}(x, A) \geq t) \leq e^{-t^2/4}$$

where dist is Euclidean distance. Also, if $f \colon [0,1]^n \to \mathbb{R}$ is a convex 1-Lipschitz function, then

$$\mathbb{P}(|f - \mathbb{M}f| \geq t) \leq 4e^{-t^2/4}.$$

Here is a form of Talagrand's inequality that is useful for combinatorial applications. Below, one should think of $f(x)$ as the value of some optimization problem on some random input x. There is a hypothesis on how much $f(x)$ can change if we alter x. An example that we will examine in the next section is the length of the shortest tour through n random points in the unit square (the Euclidean traveling salesman problem).

Theorem 9.5.14 (Talagrand's inequality — functions with weighted certificates)
Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ equipped with the product measure. Let $f : \Omega \to \mathbb{R}$ be a function. Suppose for every $x \in \Omega$, there is some $\alpha(x) = (\alpha_1(x), \ldots, \alpha_n(x)) \in \mathbb{R}^n_{\geq 0}$ such that

$$f(y) \geq f(x) - \sum_{i : x_i \neq y_i} \alpha_i(x) \quad \text{for all } y \in \Omega.$$

Then, for every $t \geq 0$, (recall $|\alpha|^2 = \sum_{i=1}^n \alpha_i(x)^2$)

$$\mathbb{P}\left(|f - \mathbb{M}f| \geq t\right) \leq 4e^{-t^2/K^2} \quad \text{where } K = 2 \sup_{x \in \Omega} |\alpha(x)|.$$

***Remark* 9.5.15.** By considering $-f$ instead of f, we can change the hypothesis on f to

$$f(y) \leq f(x) + \sum_{i : x_i \neq y_i} \alpha_i(x) \quad \text{for all } y \in \Omega.$$

Note that x and y play asymmetric roles.

***Remark* 9.5.16.** The vector $\alpha(x)$ measures the resilience of $f(x)$ under changing some coordinates of x. It is important that we can choose a different weight $\alpha(x)$ for each x. In fact, if we do not let $\alpha(x)$ change with x, then Theorem 9.5.14 recovers the bounded differences inequality Theorem 9.1.3 up to an unimportant constant factor in the exponent of the bound.

Proof. Let $r \in \mathbb{R}$. Let $A = \{y \in \Omega : f(y) \leq r - t\}$. Consider an $x \in \Omega$ with $f(x) \geq r$. By hypothesis, there is some $\alpha(x) \in \mathbb{R}^n_{\geq 0}$ such that

$$d_{\alpha(x)}(x, y) \geq f(x) - f(y) \geq t \quad \text{for all } y \in A.$$

Taking infimum over $y \in A$, we find

$$|\alpha(x)| \, d_T(x, A) \geq t.$$

So

$$d_T(x, A) \geq \frac{t}{|\alpha(x)|} \geq \frac{2t}{K}.$$

And hence by Talagrand's inequality Theorem 9.5.11,

$$\mathbb{P}(f \leq r - t)\mathbb{P}(f \geq r) \leq \mathbb{P}(x \in A)\mathbb{P}\left(d_T(x, A) \geq \frac{2t}{K}\right) \leq e^{-t^2/K^2}.$$

Taking $r = \mathbb{M}f + t$ yields

$$\mathbb{P}(f \geq \mathbb{M}f + t) \leq 2e^{-t^2/K^2},$$

and taking $r = \mathbb{M}f$ yields

$$\mathbb{P}(f \leq \mathbb{M}f - t) \leq 2e^{-t^2/K^2}.$$

Putting them together yields the final result. □

Largest eigenvalue of a random matrix

Theorem 9.5.17

Let $A = (a_{ij})$ be an $n \times n$ symmetric random matrix with independent entries in $[-1, 1]$. Let $\lambda_1(X)$ 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 9.5.14. We would like to come up with a good choice of a weight vector $\alpha(A)$ for each matrix A so that for any other symmetric matrix B with $[-1, 1]$ entries,

$$\lambda_1(B) \geq \lambda_1(A) - \sum_{i \leq j : a_{ij} \neq b_{ij}} \alpha_{i,j}. \tag{9.1}$$

Note that in a random symmetric matrix we only have $n(n+1)/2$ independent random entries: the entries below the diagonal are obtained by reflecting the upper diagonal entries.

Let $v = v(A)$ be the unit eigenvector of A corresponding to the eigenvalue $\lambda_1(A)$. Then, by the Courant–Fischer characterization of eigenvalues,

$$v^\intercal A v = \lambda_1(A) \qquad \text{and} \qquad v^\intercal B v \leq \lambda_1(B).$$

Thus

$$\lambda_1(A) - \lambda_1(B) \leq v^\intercal (A - B) v \leq \sum_{i,j : a_{ij} \neq b_{ij}} |v_i||v_j| \left|a_{ij} - b_{ij}\right| \leq \sum_{i,j : a_{ij} \neq b_{ij}} 2|v_i||v_j|.$$

Thus (9.1) holds for the vector $\alpha(A) = (\alpha_{ij})_{i \leq j}$ defined by

$$\alpha_{ij} = \begin{cases} 4|v_i||v_j| & \text{if } i < j \\ 2|v_i|^2 & \text{if } i = j. \end{cases}$$

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 9.5.14 yields the result. □

***Remark* 9.5.18.** If A has mean zero entries, then a moments computation shows that $\mathbb{E}\lambda_1(A) = O(\sqrt{n})$ (the constant can be computed as well). A much more advanced fact is that, say for uniform $\{-1, 1\}$ entries, the true scale of fluctuation is $n^{-1/6}$, and when normalized, the distribution converges to something known as the Tracy–Widom distribution. This limiting distribution is "universal" in the sense that it occurs in many naturally occurring problems, including the next example.

Certifiable functions and longest increasing subsequence

An ***increasing subsequence*** of a permutation $\sigma = (\sigma_1, \dots, \sigma_n)$ is defined to be some $(\sigma_{i_1}, \dots, \sigma_{i_\ell})$ for some $i_1 < \cdots < i_\ell$.

Question 9.5.19

How well is the length X of the longest increasing subsequence of uniform random permutation concentrated?

While the entries of σ are not independent, we can generate a uniform random permutation by taking iid uniform $x_1, \dots, x_n \sim \mathrm{Unif}[0, 1]$ and let σ record the ordering of the x_i's. This trick converts the problem into one about independent random variables.

We leave it as an exercise to deduce that X is $\Theta(\sqrt{n})$ whp.

Changing one of the x_i's changes LIS by at most 1, so the bounded differences inequality tells us that X is $O(\sqrt{n})$-subGaussian. Can we do better?

The assertion that a permutation has an increasing permutation of length s can be checked by verifying s coordinates of the permutation. Talagrand's inequality tells us that in such situations the typical fluctuation should be on the order $O(\sqrt{\mathbb{M}X})$, or $O(n^{1/4})$ in this case.

Definition 9.5.20

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$. Let $A \subseteq \Omega$. We say that A is ***s-certifiable*** if for every $x \in A$, there exists a set $I(x) \subseteq [n]$ with $|I| \le s$ such that for every $y \in \Omega$ with $x_i = y_i$ for all $i \in I(x)$, one has $y \in A$.

For example, for a random permutation as earlier, having an increasing subsequence of length $\ge s$ is s-certifiable (namely by the indices of the length s increasing subsequence).

Theorem 9.5.21 (Talagrand's inequality for certifiable functions)
Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be equipped with a product measure. Let $f\colon \Omega \to \mathbb{R}$ 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 $A, B \subseteq \Omega$ be given by $A = \{x : f(x) \leq r - t\}$ and $B = \{y : f(y) \geq r\}$. For every $y \in B$, let $I(y) \subseteq [n]$ denote a set of $\leq s$ coordinates that certify $f \geq r$. Due to f being 1-Lipschitz, we see that every $x \in A$ disagrees with y on $\geq t$ coordinates of $I(y)$.

For every $y \in B$, let $\alpha(y)$ be the indicator vector for $I(y)$ normalized in length to a unit vector. Then for any $x \in A$,

$$d_\alpha(x, y) = \frac{|\{i \in I(y) : x_i \neq y_i\}|}{\sqrt{|I|}} \geq \frac{t}{\sqrt{s}}.$$

Thus $d_T(y, A) \geq t/\sqrt{s}$. Thus

$$\mathbb{P}(f \leq r - t)\mathbb{P}(f \geq r) \leq \mathbb{P}(A)\mathbb{P}(B) \leq \mathbb{P}(x \in A)\mathbb{P}(d_T(x, A) \geq t/\sqrt{s}) \leq e^{-t^2/(4s)}$$

by Talagrand's inequality (Theorem 9.5.11). □

Corollary 9.5.22 (Talagrand's inequality for certifiable functions)
Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be equipped with a product measure. Let $f\colon \Omega \to \mathbb{R}$ 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.

$$\mathbb{P}(f \le \mathbb{M}f - t) \le 2\exp\left(\frac{-t^2}{4\mathbb{M}f}\right).$$

Setting $r = \mathbb{M}f + t$, we obtain the upper tail

$$\mathbb{P}(X \ge \mathbb{M}f + t) \le 2\exp\left(\frac{-t^2}{4(\mathbb{M}f + t)}\right). \qquad \square$$

We can apply the above corollary to $[0,1]^n$ with f being the length of the longest subsequence. Then $f \ge r$ is r-certifiable. It is also easy to deduce that $\mathbb{M}f = O(\sqrt{n})$. The above tail bounds give us a concentration window of width $O(n^{1/4})$.

Corollary 9.5.23 (Longest increasing subsequence)
Let X be the length of the longest increasing subsequence of a random permutation of $[n]$. Then for every $\varepsilon > 0$ there exists $C > 0$ so that

$$\mathbb{P}(|X - \mathbb{M}X| \le Cn^{1/4}) \ge 1 - \varepsilon.$$

***Remark* 9.5.24.** The distribution of the length X of longest increasing subsequence of a uniform random permutation is now well understood through some deep results.

Vershik and Kerov (1977) showed that $\mathbb{E}X \sim 2\sqrt{n}$.

Baik, Deift, and Johansson (1999) showed that the correct scaling factor is $n^{1/6}$, and furthermore, $n^{-1/6}(X - 2\sqrt{n})$ converges to the Tracy–Widom distribution, the same distribution for the top eigenvalue of a random matrix.

9.6 Euclidean traveling salesman problem

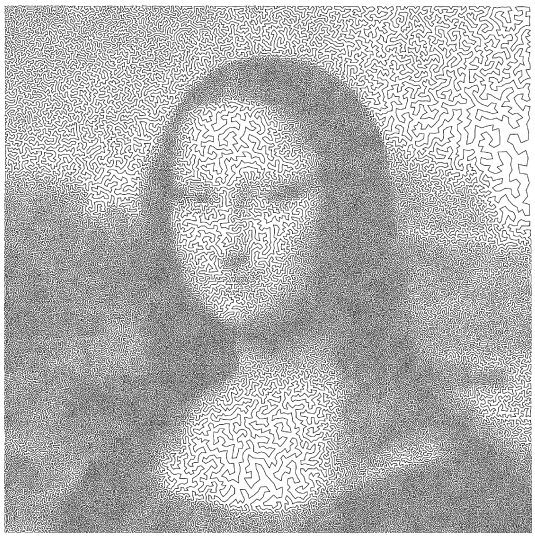
Given points $x_1, \dots, x_n \in [0,1]^2$, let $L(x_1, \dots, x_n) = L(\{x_1, \dots, x_n\})$ denote the length of the shortest tour through all given points and returns to its starting point. Equivalently, $L(x_1, \dots, x_n)$ is the minimum of

$$|x_{\sigma(1)} - x_{\sigma(2)}| + |x_{\sigma(2)} - x_{\sigma(3)}| + \cdots + |x_{\sigma(n)} - x_{\sigma(1)}|$$

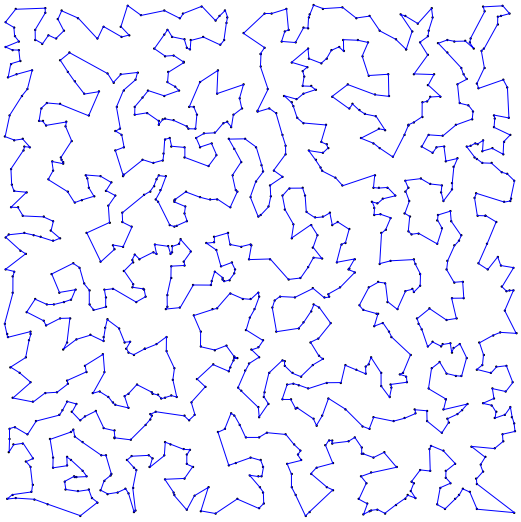
as σ ranges over all permutations of $[n]$. This Euclidean traveling salesman problem is NP-hard to solve exactly, although there is a $(1+\varepsilon)$-factor approximation algorithm with running polynomial time for any constant $\varepsilon > 0$ (Arora 1998).

Let

$$L_n = L(x_1, \dots, x_n) \quad \text{with i.i.d. } x_1, \dots, x_n \sim \mathrm{Unif}([0,1]^2)$$



The Mona Lisa TSP challenge.



A tour of 1000 random points.

Exercise: $\mathbb{E}L_n = \Theta(\sqrt{n})$

Beardwood, Halton, and Hammersley (1959) showed that whp $L_n/\sqrt{n}$ converges to some constant as $n \to \infty$ (the exact value of the constant is unknown).

We shall focus on the concentration of L_n.

We will present two methods that illustrate different techniques from this chapter.

Martingale methods

The following simple monotonicity property will be important for us: for any S and $x \in [0,1]^2$,

$$L(S) \le L(S \cup \{x\}) \le L(S) + 2\,\mathrm{dist}(x, S).$$

Here is the justification for the second inequality. Let y be the closest point in S to x. Consider a shortest tour through S of length $L(S)$. Let us modify this tour by first traversing through it, and when we hit y, we take a detour excursion from y to x and then back to y. The length of this tour, which contains $S \cup \{x\}$, is $L(S) + 2\,\mathrm{dist}(x, S)$, and thus the shortest tour through $S \cup \{x\}$ has length at most $L(S) + 2\,\mathrm{dist}(x, S)$.

If we simply apply the bounded difference inequality, we find that changing a single x_i might change $L(x_1, \dots, x_n)$ by $O(1)$ in the worse case, and so L_n is $O(\sqrt{n})$-subGaussian about its mean. This is a trivial result since L_n is typically $\Theta(\sqrt{n})$.

To do better, we apply Azuma's inequlality to the Doob martingale. The key observation is that the initially revealed points do not affect the conditional expectations by much even in the worst case.

Theorem 9.6.1 (Rhee and Talagrand 1987)
L_n is $O(\sqrt{\log n})$-subGaussian about its mean. That is,

$$\mathbb{P}(|L_n - \mathbb{E}L_n| \geq t) \leq \exp\left(\frac{-ct^2}{\log n}\right) \quad \text{for all } t > 0,$$

where $c > 0$ is some constant.

We need the following estimate.

Lemma 9.6.2
Let S be a set of k random points chosen independently and uniformly in $[0,1]^2$. For any (non-random) point $y \in [0,1]^2$, one has

$$\mathbb{E}\operatorname{dist}(y, S) \lesssim \frac{1}{\sqrt{k}}.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\operatorname{dist}(y,S) &= \int_0^{\sqrt{2}} \mathbb{P}(\operatorname{dist}(y,S) \geq t)\, dt \\
&= \int_0^{\sqrt{2}} \left(1 - \operatorname{area}\left(B(y,t) \cap [0,1]^2\right)\right)^k dt \\
&\leq \int_0^{\sqrt{2}} \exp\left(-k \operatorname{area}\left(B(y,t) \cap [0,1]^2\right)\right) dt \\
&\leq \int_0^{\infty} \exp\left(-\Omega(kt^2)\right) dt \lesssim \frac{1}{\sqrt{k}}.
\end{aligned}$$

□

Proof of Theorem 9.6.1. Let

$$L_{n,i}(x_1, \ldots, x_i) = \mathbb{E}\left[L_n(x_1, \ldots, x_n) \mid x_1, \ldots, x_i\right]$$

be the expectation of L_n conditional on the first i points (and averaging over the remaining $n - i$ points).

Claim. $L_{n,i}$ is $O\left(\frac{1}{\sqrt{n-i+1}}\right)$-Lipschitz with respect to Hamming distance.

We have

$$\begin{aligned}
L(x_1, \ldots, x_i, \ldots x_n) &\leq L(x_1, \ldots, x_i', \ldots x_n) + 2\operatorname{dist}(x_i, \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}) \\
&\leq L(x_1, \ldots, x_i, \ldots x_n) + \begin{cases} 2\operatorname{dist}(x_i, \{x_{i+1}, \ldots, x_n\}) & \text{if } i < n \\ O(1) & \text{if } i = n. \end{cases}
\end{aligned}$$

Taking expectation over $x_{i+1}, \dots, x_n$, and applying the previous lemma, we find that

$$L_{n,i}(x_1, \dots, x_i) \le L_{n,i}(x_1, \dots, x_{i-1}, x_i') + O\left(\frac{1}{\sqrt{n-i+1}}\right).$$

This proves the claim. Thus the Doob martingale

$$Z_i = \mathbb{E}\left[L_n(x_1, \dots, x_n) \mid x_1, \dots, x_i\right] = L_{n,i}(x_1, \dots, x_i)$$

satisfies

$$|Z_i - Z_{i-1}| \lesssim \frac{1}{\sqrt{n-i+1}} \quad \text{for each } 1 \le i \le n.$$

Now we apply Azuma's inequality (Theorem 9.2.8). Since

$$\sum_{i=1}^{n} \left(\frac{1}{\sqrt{n-i+1}}\right)^2 = O(\log n),$$

we deduce that $Z_N = L_n$ is $O(\sqrt{\log n})$-subGaussian about its mean. □

Talagrand's inequality

Using Talagrand's inequality, we will prove the following stronger estimate.

Theorem 9.6.3 (Rhee and Talagrand 1989)
L_n is $O(1)$-subGaussian about its mean. That is,

$$\mathbb{P}(|L_n - \mathbb{E}L_n| \ge t) \le e^{-ct^2} \quad \text{for all } t > 0,$$

where $c > 0$ is some constant.

***Remark* 9.6.4.** Rhee (1991) showed that this tail bound is sharp.

The proof below, following Steele (1997), applies the "space-filling curve heuristic."

A ***space-filling curve*** is a continuous surjection from $[0, 1]$ to $[0, 1]^2$. Peano (1890) constructed the first space-filling curve. Hilbert (1891) constructed another space-filling curve known as the ***Hilbert curve***. We will not give a precise description of the Hilbert curve here. Intuitively, the Hilbert curve is the pointwise limit of a sequence of piecewise linear curves illustrated in Figure 9.2. I recommend this 3Blue1Brown video on YouTube for a beautiful animation of the Hilbert curve along with applications.

We will only need the following property of the Hilbert space filling curve.

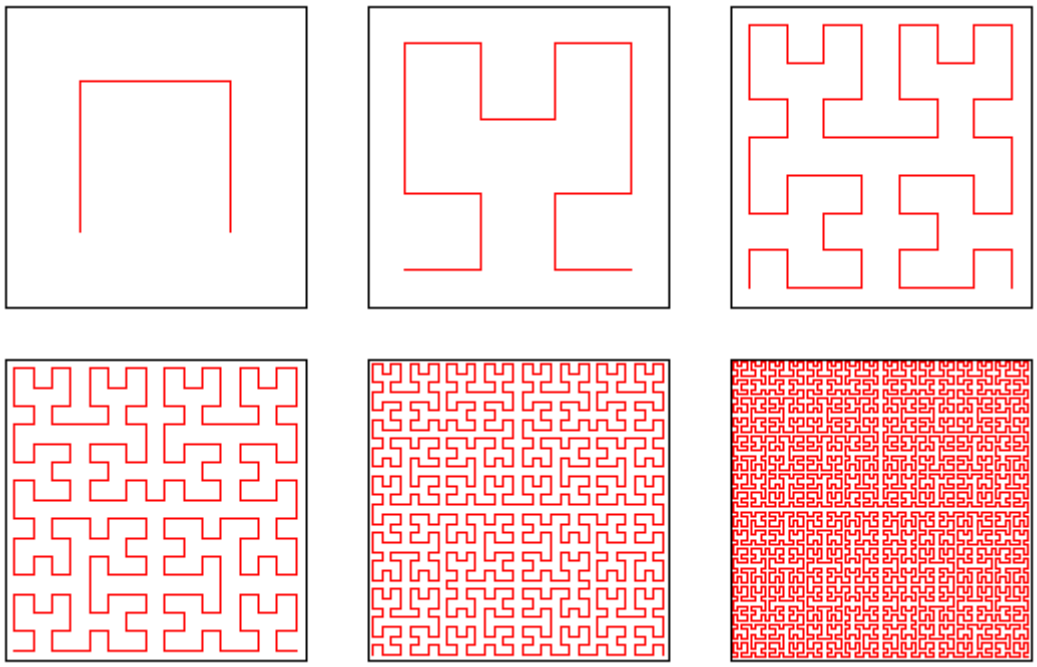


Figure 9.2: The Hilbert space-filling curve is the limit of discrete curves illustrated.

Definition 9.6.5 (Hölder continuity)

Given two metric spaces (X, d_X) and (Y, d_Y), we say that a map $f\colon X \to Y$ is ***Hölder continuous with exponent α*** if there is some constant C (depending on f) so that

$$d_Y(f(x), f(x')) \le C d_X(x, x')^\alpha \quad \text{for all } x, x' \in X.$$

***Remark* 9.6.6.** Hölder continuity with exponent 1 is the same as Lipschitz continuity. Often X has bounded diameter, in which case if f is Hölder continuous with exponent α, then it is so with any exponent $\alpha' < \alpha$.

Theorem 9.6.7

The Hilbert curve $H\colon [0,1] \to [0,1]^2$ is Hölder continuous with exponent $1/2$.

Proof sketch. The Hilbert space-filling curve H sends every interval of the form $[(i-1)/4^n, i/4^n]$ to a square of the form $[(j-1)/2^n, j/2^n] \times [(k-1)/2^n, k/2^n]$. Indeed, for each fixed n, the discrete curves eventually all have this property.

Let $x, y \in [0,1]$, and let n be the largest integer so that $x, y \in [(i-1)/4^n, (i+1)/4^n]$ for some integer i. Then $|x-y| = \Theta(4^{-n})$, and $|H(x) - H(y)| \lesssim 2^{-n}$. Thus $|H(x) - H(y)| \lesssim |x-y|^{1/2}$. □

***Remark* 9.6.8.** If a space filling space is Hölder continuous with exponent α, then $\alpha \le 1/2$. Indeed, the images of the intervals $[(i-1)/k, i/k]$, $i = 1, \dots, k$, cover the unit square, and thus one intervals must have image diameter $\gtrsim 1/\sqrt{k}$.

Lemma 9.6.9 (Space-filling curve heuristic)
Let $x_1, \dots, x_n \in [0,1]^2$. There is a permutation of σ of $[n]$ with (indices taken mod n)

$$\sum_{i=1}^{n} \left|x_{\sigma(i)} - x_{\sigma(i+1)}\right|^2 = O(1).$$

Proof. Order the points as they appear on the Hilbert space filling curve $H\colon [0,1] \to [0,1]^2$ (since H is not injective, there is more than one possible order). Then, there exist $0 \le t_1 \le t_2 \le \cdots \le t_n \le 1$ so that $H(t_i) = x_{\sigma(i)}$ for each i. Using that H is Hölder continuous with exponent $1/2$, we have

$$\sum_{i=1}^{n} \left|x_{\sigma(i)} - x_{\sigma(i+1)}\right|^2 = \sum_{i=1}^{n} |H(t_i) - H(t_{i+1})|^2 \lesssim \sum_{i=1}^{n} |t_i - t_{i+1}| \le 2. \qquad \square$$

***Remark* 9.6.10.** We leave it as an exercise to find an elementary proof of the lemma without invoking the existence of a space-filling curve. Hint: consider a finite approximation of the Hilbert curve.

Using Talagrand's inequality in the form of Theorem 9.5.14, to prove Theorem 9.6.3 that L_n is $O(1)$-subGaussian, it suffices to prove the following lemma.

Lemma 9.6.11
Let $\Omega = ([0,1]^2)^n$ be the space of n-tuples of points in $[0,1]^2$. There exists a map $\alpha\colon \Omega \to \mathbb{R}^n_{\ge 0}$ so that for all $x \in \Omega$, $\alpha(x) = (\alpha_1(x), \dots, \alpha_n(x)) \in \mathbb{R}^n_{\ge 0}$ satisfies

$$L(x) \le L(y) + \sum_{i : x_i \ne y_i} \alpha_i(x) \quad \text{for all } x, y \in \Omega \tag{9.1}$$

and

$$\sup_{x \in \Omega} \sum_{i=1}^{n} \alpha_i(x)^2 = O(1). \tag{9.2}$$

Proof. Let $x = (x_1, \dots, x_n) \in \Omega$, and let σ be the permutation of $[n]$ given by Lemma 9.6.9, the space-filling curve heuristic. Then σ induces a tour of $x_1, \dots, x_n$. Let $\alpha_i(x)$ equal twice the sum of the lengths of the two edges incident to x_i in this tour

(indices taken mod n):

$$\alpha_i(x) = 2\left(\left|x_i - x_{\sigma(\sigma^{-1}(i)+1)}\right| + \left|x_i - x_{\sigma(\sigma^{-1}(i)-1)}\right|\right).$$

Intuitively, this quantity captures "difficulty to serve" x_i.

Now we prove (9.1). First we take care of the first case when $x_i \neq y_i$ for all i: (9.1) follows from

$$L(x) \le \sum_{i=1}^{n} \left|x_{\sigma(i)} - x_{\sigma_{i+1}}\right| = \frac{1}{2}\sum_{i=1}^{n} \alpha_i(x).$$

Now suppose that $x_i = y_i$ for at least one i. Suppose we have a tour through y of length $L(y)$. Consider, for each i with $x_i \neq y_i$, the point x_i along with the two segments incident to x_i in the σ-induced tour through x (these are the "new edges"). Starting with an optimal tour through y, and by making various detours/excursions on the new edges, we can reach all the points of x, traversing each new edge at most twice. The length of the new tour is at most $L(y) + \sum_{i:x_i\neq y_i} \alpha_i(x)$. This proves (9.1).

Finally, it remains to prove (9.2). By Lemma 9.6.9,

$$\begin{aligned}\sum_{i=1}^{n} \alpha_i(x)^2 &\le 4\sum_{j=1}^{n} \left(\left|x_{\sigma(j)} - x_{\sigma(j+1)}\right| + \left|x_{\sigma(j)} - x_{\sigma(j+1)}\right|\right)^2 \\ &\lesssim \sum_{j=1}^{n} \left|x_{\sigma(j)} - x_{\sigma(j+1)}\right|^2 = O(1).\end{aligned}$$

□

Exercises

1. *Sub-Gaussian tails.* For each part, prove there is some constant $c > 0$ so that, for all $\lambda > 0$,
$$\mathbb{P}(|X - \mathbb{E}X| \ge \lambda\sqrt{\operatorname{Var} X}) \le 2e^{-c\lambda^2}.$$
   a) X is the number of triangles in $G(n, 1/2)$.
   b) X is the number of inversions of a uniform random permutation of $[n]$ (an *inversion* of $\sigma \in S_n$ is a pair (i, j) with $i < j$ and $\sigma(i) > \sigma(j)$).
2. Prove that for every $\varepsilon > 0$ there exists $\delta > 0$ and n_0 such that for all $n \ge n_0$ and $S_1, \dots, S_m \subset [2n]$ with $m \le 2^{\delta n}$ and $|S_i| = n$ for all $i \in [m]$, there exists a function $f\colon [2n] \to [n]$ so that $(1 - e^{-1} - \varepsilon)n \le |f(S_i)| \le (1 - e^{-1} + \varepsilon)n$ for all $i \in [m]$.
3. *Simultaneous bisections.* Fix Δ. Let $G_1, \dots, G_m$ with $m = 2^{o(n)}$ be connected graphs of maximum degree at most Δ on the same vertex set V with $|V| = n$. Prove that there exists a partition $V = A \cup B$ so that every G_i has $(1+o(1))e(G_i)/2$

edges between A and B.

4. ⋆ Prove that there is some constant $c > 0$ so that for every graph G with chromatic number k, letting S be a uniform random subset of V and $G[S]$ the subgraph induced by S, one has, for every $t \geq 0$,

$$\mathbb{P}(\chi(G[S]) \leq k/2 - t) \leq e^{-ct^2/k}.$$

5. ⋆ Prove that there is some constant $c > 0$ so that, with probability $1 - o(1)$, $G(n, 1/2)$ has a bipartite subgraph with at least $n^2/8 + cn^{3/2}$ edges.

6. Let $k \leq n/2$ be positive integers and G an n-vertex graph with average degree at most n/k. Prove that a uniform random k-element subset of the vertices of G contains an independent set of size at least ck with probability at least $1 - e^{-ck}$, where $c > 0$ is a constant.

7. ⋆ Prove that there exists a constant $c > 0$ so that the following holds. Let G be a d-regular graph and $v_0 \in V(G)$. Let $m \in \mathbb{N}$ and consider a simple random walk $v_0, v_1, \ldots, v_m$ where each v_{i+1} is a uniform random neighbor of v_i. For each $v \in V(G)$, let X_v be the number times that v appears among $v_0, \ldots, v_m$. For that for every $v \in V(G)$ and $\lambda > 0$

$$\mathbb{P}\left(\left|X_v - \frac{1}{d}\sum_{w \in N(v)} X_w\right| \geq \lambda + 1\right) \leq 2e^{-c\lambda^2/m}$$

   Here $N(v)$ is the neighborhood of v.

8. Prove that for every k there exists a $2^{(1+o(1))k/2}$-vertex graph that contains every k-vertex graph as an induced subgraph.

9. ⋆ *Tighter concentration of chromatic number*

   a) Prove that with probability $1 - o(1)$, every vertex subset of $G(n, 1/2)$ with at least $n^{1/3}$ vertices contains an independent set of size at least $c \log n$, where $c > 0$ is some constant.

   b) Prove that there exists some function $f(n)$ and constant C such that for all $n \geq 2$,

$$\mathbb{P}(f(n) \leq \chi(G(n, 1/2)) \leq f(n) + C\sqrt{n}/\log n) \geq 0.99.$$

10. Show that for every $\varepsilon > 0$ there exists $C > 0$ so that every $S \subseteq [4]^n$ with $|S| \geq \varepsilon 4^n$ contains four elements with pairwise Hamming distance at least $n - C\sqrt{n}$ apart.

11. *Concentration of measure in the symmetric group.* Let $U \subseteq S_n$ be a set of at least $n!/2$ permutations of $[n]$. Let U_t denote the set of permutations that can be obtained starting from some element of U and then applying at most t transpositions. Prove that

$$|U_t| \geq (1 - e^{-ct^2/n})n!$$

for every $t \geq 0$, where $c > 0$ is some constant.

Hint: Apply Azuma to a Doob martingale that reveals a random permutation

For the remaining exercises in this section, use Talagrand's inequality

12. Let Q be a subset of the unit sphere in $\mathbb{R}^n$. Let $\boldsymbol{x} \in [-1, 1]^n$ be a random vector with independent random coordinates. Let $X = \sup_{\boldsymbol{q} \in Q} \langle \boldsymbol{x}, \boldsymbol{q} \rangle$. Let $t > 0$. Prove that

$$\mathbb{P}(|X - \mathbb{M}X| \geq t) \leq 4e^{-ct^2}$$

where $c > 0$ is some constant.

13. *First passage percolation.* Prove that there are constants $c, C > 0$ so that the following holds. Let G be a graph, and let u and w be two distinct vertices with distance at most ℓ between them. Every edge of G is independently assigned some random weight in $[0, 1]$ (not necessarily uniform or identically distributed). The weight of a path is defined to be the sum of the weights of its edges. Let X be the minimum weight of a path from u to w using at most ℓ edges. Prove that there is some $m \in \mathbb{R}$ so that

$$\mathbb{P}(|X - m| \geq t) \leq Ce^{-ct^2/\ell}.$$

14. $\star$ *Second largest eigenvalue of a random matrix.* Let A be an $n \times n$ random symmetric matrix whose entries on and above the diagonal are independent and in $[-1, 1]$. Show that the second largest eigenvalue $\lambda_2(A)$ satisfies

$$\mathbb{P}(|\lambda_2(A) - \mathbb{E}\lambda_2(A)| \geq t) \leq Ce^{-ct^2},$$

for every $t \geq 0$, where $C, c > 0$ are constants.

Hint in white:

15. *Longest common subsequence.* Let $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_m)$ be two random sequences with independent entries (not necessarily identically distributed). Let X denote the length of the longest common subsequence, i.e., the largest k such that there exist $i_1 < \cdots < i_k$ and $j_1 < \cdots < j_k$ with $x_{i_1} = y_{j_1}, \ldots, x_{i_k} = y_{j_k}$.

Show that, for all $t \geq 0$,

$$\mathbb{P}(X \geq \mathbb{M}X + t) \leq 2\exp\left(\frac{-ct^2}{\mathbb{M}X + t}\right) \quad \text{and} \quad \mathbb{P}(X \leq \mathbb{M}X - t) \leq 2\exp\left(\frac{-ct^2}{\mathbb{M}X}\right)$$

where $c > 0$ is some constant.