9 Concentration of measure

9.1 The geometric picture

Concentration of measure is an important concept in high-dimensional probability and geometry. We've shown examples of concentration of Lipschitz functions of many variables, and it turns out this concept is integrally connected to isoperimetry. For example, what's the largest area we can fence off with a given perimeter? This can be rephrased:

Problem 9.1

Given some constant volume V, what's the minimum possible surface area of that volume?

An example of a space we can work in is the **Hamming cube**: if we have an n-dimensional cube, and we label some number of points, what's the minimum size of the boundary? We can consider the **Hamming distance**, the number of differing coordinates between two vertices of the cube. It seems that we want to take some portion of the cube which is within some Hamming distance of a fixed point (this is a "ball"), which turns out to be true:

Theorem 9.2 (Harper)

If B is a (Hamming) ball, and the volume of A is equal to the volume of B, then the volume of A_t is at least the volume of B_t, where A_t is the set of all points within a distance t from a fixed point A.

What does this have to do with concentration of measure? We can prove an approximate version of Harper's theorem. For n very large, the distribution of Hamming distances looks like a normal distribution with width $\sqrt{n}$, so starting with a Hamming ball with ε area can be thought of as the set of points below $\frac{-t\sqrt{n}}{2}$ on the normal distribution.

Theorem 9.3

For every $\varepsilon > 0$, there exists a $t > 0$ such that for any subset $A \subset \{0,1\}^n$ of the Hamming cube with $|A| \geq \varepsilon 2^n$, $|A_{t\sqrt{n}}| \geq (1-\varepsilon)2^n$.

Proof. We're looking at a hypercube: pick a random vertex x in $\{0,1\}^n$ uniformly, and let X be the distance between x and the closest point in A. By the triangle inequality, this is 1-Lipschitz, and this is informative because $X = 0$ is the same as saying $x \in A$, which happens with probability at least ε. By the Azuma lower tail inequality, the probability that $X = 0$ is

$$\Pr(X \leq \mathbb{E}[X] - \mathbb{E}[X]) \leq \exp\left(-\frac{\mathbb{E}[X]^2}{2n}\right).$$

This gives an upper bound on the expectation of X: $\mathbb{E}[X] \leq \sqrt{2\log\left(\frac{1}{\varepsilon}\right)n}$, and now we use the upper tail estimate. That tells us that x shouldn't deviate too much: the probability $x \notin A_{t\sqrt{n}}$, where $t = 2\sqrt{2\log\left(\frac{1}{\varepsilon}\right)}$, is

$$\Pr\left(X > 2\sqrt{2\log\left(\frac{1}{\varepsilon}\right)n}\right) \leq \Pr\left(X > \mathbb{E}[X] + \sqrt{2\log\left(\frac{1}{\varepsilon}\right)n}\right) \leq \varepsilon.$$

(Rephrased, our variable is pretty large in expectation, and it is rarely very large.) So x is in $A_{t\sqrt{n}}$ with probability at least $1-\varepsilon$, as desired. □

This is actually a fairly general result, and we can go back and forth between the geometric and combinatorial interpretations of this statement.

Proposition 9.4

Let $t, \varepsilon > 0$ be real numbers, and let Ω be a probability space on which there exists a metric (such as the Hamming cube with the Hamming metric). Then the following are equivalent:

1. (Approximate isoperimetry) For all subsets $A \subset \Omega$ with $\Pr(A) \geq \frac{1}{2}$, then given a set
$$A_t = \{\omega : \text{dist}(\omega, A) \leq t\},$$
we have $\Pr(A_t) \geq 1 - \varepsilon$.
2. (Concentration of Lipschitz functions) For all functions $f : \Omega \to \mathbb{R}$ that are 1-Lipschitz – $|f(x) - f(y)| \leq \text{dist}_\Omega(x, y)$ – if we have a **median** $m \in \mathbb{R}$ such that $\Pr(f \leq m) \geq \frac{1}{2}$ and $\Pr(f \geq m) \geq \frac{1}{2}$, then
$$\Pr(f > m + t) \leq \varepsilon.$$

Note that this is concentration around the median, not the mean. We'll soon see that these aren't that different, though.

Proof. First let's show that (1) implies (2). Take the half of the probability space
$$A = \{\omega \in \Omega : f(\omega) \leq m\}.$$
This is at least half of our probability space by the definition of the median, and since f is 1-Lipschitz,
$$f(\omega) \leq m + t \ \forall \omega \in A_t.$$
Thus,
$$\Pr(f > m + t) \leq \Pr(\overline{A_t}) \leq \varepsilon$$
by condition (1).

The reverse implication (2) to (1) is not that hard either. We want to show that given any set A with half the space, its t-neighborhood consumes almost the whole space. The natural choice for our Lipschitz function f is the distance
$$f(\omega) = \text{dist}(\omega, A).$$
We pick $m = 0$, and now
$$\Pr(\overline{A_t}) = \Pr(\text{dist}(\omega, A) > t) \leq \varepsilon,$$
by condition (2). Now take the complement, $\Pr(A_t) = \Pr(f \leq t) \geq 1 - \varepsilon$, and we've shown condition (1). □

This can be useful, because sometimes it's more natural to think in terms of isoperimetry instead of functions (or vice versa).

9.2 Results about concentration: median versus mean

Let's look at another form of concentration of Lipschitz functions:

Proposition 9.5

If we have a 1-Lipschitz function $f : \{0,1\}^n \to \mathbb{R}$ (with respect to the Hamming metric), pick $\omega \sim \text{Unif}(\{0,1\}^n)$, and let $X = f(\omega)$ be our random variable. Then for all $s \in \mathbb{R}$, $t > 0$,

$$\Pr(X \le s)\Pr(X \ge s+t) \le e^{-t^2/(4n)}.$$

We should think about as taking "either s or $s+t$ to be a median:" then one of the terms becomes a constant $\frac{1}{2}$, and moving that to the other side gives the Gaussian-like tail for the other term.

Proof. We'll apply Azuma's inequality twice. Let μ be the mean of $X - s$ (we can always just shift the variable X so that $s = 0$). If $\mu < 0$, then by Azuma upper tail,

$$\Pr(X \le s)\Pr(X \ge s+t) \le \Pr(X \ge s+t) = \Pr(X - s - \mu \ge t - \mu) \le \exp\left(-\frac{(t-\mu)^2}{2n}\right) \le e^{-t^2/2n}$$

and we're done. Similarly, if $t - \mu < 0$,

$$\Pr(X \le s)\Pr(X \ge s+t) \le \Pr(X \le s) = \Pr(X - s - \mu \le -\mu) \le e^{-\mu^2/2n} \le e^{-t^2/2n},$$

and we're again done. So we're just left with the case where $\mu > 0$ and $t - \mu \ge 0$.

In this case, by Azuma's inequality (lower tail), we can say that

$$\Pr(X \le s) = \Pr(X - s - \mu \le -\mu)$$

and since $X - s - \mu$ is mean-zero and Lipschitz, this is at most $e^{-\mu^2/(2n)}$. On the other hand, by Azuma (upper tail),

$$\Pr(X \ge s+t) = \Pr(X - s - \mu \ge t - \mu) \le e^{(t-\mu)^2/(2n)}.$$

(Be careful here: we can really use this for $t - \mu \ge 0$, but otherwise we can just repeat the argument the other way around by starting with an upper tail argument instead.) Putting these together,

$$\Pr(X \le s)\Pr(X \ge s+t) \le \exp\left[-\frac{\mu^2 + (t-\mu)^2}{2n}\right] \le \exp\left[-\frac{t^2}{4n}\right]$$

by convexity, which is what we want. □

The next few sections are essentially about how to interpret these ideas. One way is to think about $A \subset \{0,1\}^n$ as a subset of the Boolean cube. If $\Pr(A) = \frac{|A|}{2^n}$ (uniform measure), then we have the following:

Corollary 9.6

Consider a uniform measure on the Boolean cube, and let $t > 0$. Then

$$\Pr(A)\Pr(\overline{A_t}) \le e^{-t^2/(4n)}.$$

In particular, if A is at least half the cube, and we expand it by some $c\sqrt{n}$, we get almost the entire cube:

$$\Pr(A_t) \ge 1 - 2e^{-t^2/(4n)}.$$

Earlier in the class, we were using the mean for concentration and other concepts, but now we have the median

instead: what relations are there between the mean and median? Suppose that we have a bound of the form

$$\Pr(\overline{A_t}) \leq Ce^{-(t/\sigma)^2}$$

for all A with $\Pr(A) \geq \frac{1}{2}$. (In this case, $\sigma \asymp \sqrt{n}$.) Given any 1-Lipschitz function f, and letting $X = f(\omega)$ (where ω is random in Ω), if we have a median m of X, then the difference between the mean and median is

$$|\mathbb{E}[X] - m| \leq \mathbb{E}|X - m| = \int_0^\infty \Pr(|X - m| \geq t)dt.$$

If we have sub-Gaussian tail bounds, this is

$$\leq \int_0^\infty 2Ce^{-(t/\sigma)^2} = C\sqrt{\pi}\sigma,$$

so the mean and median don't differ by more than a constant times σ. This is actually the tightest bound we can produce: consider the function $f : \{-1, 1\}^n \to \mathbb{R}$ defined by

$$f(x_1, \cdots, x_n) = |X_1 + \cdots + X_n|.$$

We can evaluate its mean and median by the Central Limit Theorem: $\frac{\mathbb{E}[X]}{\sqrt{n}}$ and $\frac{\text{med } X}{\sqrt{n}}$ converge to $\mathbb{E}|Z|$ and $\text{med}|Z|$, where Z is the standard normal – those are different constants! So the idea is that in general, we have

$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq C'e^{-(t/\sigma)^2},$$

and we have concentration around the mean as well - just possibly with a worse constant than the median version.

9.3 High-dimensional spheres

Most of our intuition about high-dimensional geometry is wrong! A good reference is Keith Ball's "An elementary introduction to modern convex geometry."

Theorem 9.7 (Isoperimetric inequality in $\mathbb{R}^n$)
If $A, B \subset \mathbb{R}^n$, B is a ball, and $\lambda(A) = \lambda(B)$ (they have the same measure), then

$$\lambda(A_t) \geq \lambda(B_t)$$

for all $t > 0$.

Basically, this is asking for the "smallest perimeter" among all sets of the same volume. We saw a version of this earlier by Harper: if $A, B \subset \{0, 1\}^n$ are subsets of the Hamming cube, $|A| = |B|$ and B is a Hamming ball, then $|A_t| \geq |B_t|$ for all $t > 0$.

Here's the most conceptual way to think about it: this is called **Steiner symmetrization** (alternatively shifting or compression) in the discrete case. The idea is to transform A to preserve the volume and decrease its perimeter. Cut it in half so we have half the volume on each side: if one side has smaller perimeter, then lose the worse side and reflect over the cut. If we can't do this, then every cut must cut the perimeter in two: then just show that such a shape must be a ball.

This isn't actually a proof though: we might need to do this infinitely many times, so there are some compactness issues with this idea. For the discrete setting, we just keep compressing our shape in some direction.

It turns out there's also a spherical isoperimetric inequality:

Theorem 9.8 (Levy)

On the unit sphere $S^{n-1} \subset \mathbb{R}^n$, use the arc distance (though it doesn't really matter). Then given two subsets $A, B \subset S^{n-1}$, where B is some spherical cap and $\lambda(A) = \lambda(B)$, we have $\lambda(A_t) \geq \lambda(B_t)$ for all $t > 0$.

This isn't easy to show, but remember that approximate isoperimetry is connected to concentration of measure. The counterintuitive thing is that distribution of measure is very different in high dimensions than in our ordinary 3-D space.

Fact 9.9

The volume of a t-neighborhood of a hemisphere C is almost everything:

$$\Pr(C_t) \geq 1 - e^{-ct^2 n},$$

where n is the number of dimensions.

We should not think of an n-dimensional shape as a very ball-like object: distribution of mass looks normal along any axis, with standard deviation $\frac{c}{\sqrt{n}}$. Also, most of the mass is near the surface rather than the middle!

Since there's a spherical isoperimetric inequality, we should also have an analogous statement about Lipschitz functions for the sphere:

Proposition 9.10

There exist absolute constants c, C (not dependent on n) such that given a function $f : S^{n-1} \to \mathbb{R}$ that is 1-Lipschitz,

$$\Pr(|f - \mathbb{E}[f]| > t) \leq Ce^{-ct^2 n}.$$

This can be rephrased as "every Lipschitz function is nearly constant nearly everywhere."

For cultural value, here's one more space that's good to mention: the Gauss space in $\mathbb{R}^n$ has the Euclidean metric, and we have the probability distribution

$$\vec{Z} = (Z_1, \cdots, Z_n) : Z_i \sim N(0, 1)$$

(with all Z_is independently distributed). We again have an isoperimetry theorem:

Theorem 9.11

If $A, B \subset \mathbb{R}^n$, and B is a "ball" with $\Pr(A) = \Pr(B)$, then

$$\Pr(A_t) \geq \Pr(B_t) \quad \forall t > 0.$$

A "ball" in Gauss space is intuitively supposed to look like a sphere of radius $\sqrt{n}$, because the probability density function is

$$f_n(x) = (2\pi)^{-n} e^{-|x|^2/2}.$$

The nice thing here is that Gaussian vectors are rotationally symmetric, and now the length of this vector can be written more simply:

$$|\vec{z}|^2 = z_1^2 + \cdots + z_n^2.$$

The expectation of $|\vec{z}|^2$ is n, and along with the spherical isoperimetry inequality, we now have a way to describe balls in the Gauss space: B should be some half-space. (It can have measure not equal to $\frac{1}{2}$ if we don't have the boundary of that half-space passing through the origin.)

9.4 Projections onto subspaces

The idea with our next section is that we want to represent a bunch of points in a smaller dimension without distorting the distances too much.

Theorem 9.12 (Johnson-Lindenstrauss Lemma)
Let $s_1, \cdots, s_N$ be points in $\mathbb{R}^d$. Then there exist $s_1', s_2', \ldots, s_N' \in \mathbb{R}^m$, where $m = O(\varepsilon^{-2} \log N)$, such that

$$(1-\varepsilon)|s_i - s_j| \leq |s_i' - s_j'| \leq (1+\varepsilon)|s_i - s_j|.$$

So we can approximately preserve distances up to a small multiplicative error.

Proof. Pick a random (orthogonal) projection onto an m-dimensional subspace (chosen uniformly at random). This projection is actually agnostic to the set of points $s_1, \cdots, s_N$. We claim that with positive probability, the desired outcome occurs. If we do this naively, everything gets smaller, so we'll scale by $\sqrt{\frac{n}{m}}$ to correct for that. Basically, our claim is that all the length ratios are generally preserved.

Lemma 9.13
Let P be a projection from $\mathbb{R}^n$ onto a random m-dimensional subspace, and let $z \in \mathbb{R}^n$ be some fixed vector. If we let $z' = Pz$ be a random variable, then $\mathbb{E}[|z'|^2] = \frac{m}{n}|z|^2$, and we have

$$(1-\varepsilon)\sqrt{\frac{m}{n}}|z| \leq |z'| \leq (1+\varepsilon)\sqrt{\frac{m}{n}}|z|$$

with probability at least $1 - e^{-c\varepsilon^2 m}$.

Proof of lemma. Note that fixing our vector and picking a random subspace is equivalent to fixing our projection P and choosing a random unit vector in $\mathbb{R}^n$. By rotational symmetry, we can make P the span of the first m basis elements $\{e_1, \cdots, e_m\}$, and thus any $z = (z_1, \cdots, z_n)$ corresponds to $z' = (z_1, \cdots, z_m, 0, \cdots, 0)$.

Note that $\mathbb{E}[|z'|^2] = \mathbb{E}[z_1^2 + \cdots + z_n^2]$: in general, it's easier to look at squared lengths than lengths. By symmetry, because all the z_i^2s have the same expectation, each z_i^2 has expected value $\frac{1}{n}$ (using linearity), so

$$\mathbb{E}|z'|^2 = \mathbb{E}[z_1^2 + \cdots + z_m^2] = \frac{m}{n}.$$

But now note that projection $z \to |z'|$ is a 1-Lipschitz function, so by Levy concentration (the isoperimetric inequality on the sphere), we know that

$$\Pr\left(\left||z'| - \sqrt{\frac{n}{m}}|z|\right| > \varepsilon\sqrt{\frac{n}{m}}\right) \leq \exp\left[-cn \cdot \frac{m}{n}\varepsilon^2\right] = \exp[-m\varepsilon^2].$$

(remember that mean and median are reasonably close because of Gaussian tails), which is exactly what the lemma claims. □

So now we can finish with a union bound: since everything happens with high probability, we can just say that the probability some pair (i, j) fails the distance check is at most

$$\leq N^2 e^{-c\varepsilon^2 m} < 1$$

as long as m is chosen to be $O(\varepsilon^{-2} \log N)$, and we have the desired result. □

9.5 What if we need stronger concentration?

Unfortunately, Azuma's inequality is not enough to solve all of our problems. Consider the following:

Problem 9.14

Let V be a fixed d-dimensional subspace (through the origin), and we pick a point $X \sim \text{Unif}(\{-1, 1\}^n)$. How well is the Euclidean distance $\text{dist}(x, V)$ concentrated?

We have n independent Boolean variables, so we have a Lipschitz function of x, which gives $\sqrt{n}$ concentration of our random variable X by Azuma's inequality. In particular, the probability that our variable is within $t\sqrt{n}$ of its mean decays like a Gaussian in t.

But the diameter of the cube itself is proportional to $\sqrt{n}$, so this is a pretty bad estimate!

Note that we can change the problem a bit, and Azuma does become pretty good. Specifically, if we pick V uniformly at random (then x can be either a fixed point or chosen randomly - it doesn't matter), Azuma's gives pretty good concentration. Alternatively, we could pick x uniformly at random from the sphere that goes through the vertices of the Boolean cube as well, and Azuma still yields reasonable concentration. These are the same by rotational symmetry, and in the second case, we're asking for the concentration of a Lipschitz function on a sphere. We know then that if f is Lipschitz on a $\sqrt{n}$-radius n-dimensional sphere, then

$$\Pr(|f - \mathbb{E}[f]| > t) \leq Ce^{-ct^2}.$$

So if we can get $O(1)$ concentration on the sphere, intuitively we should also be able to get it on the Boolean cube as well. We just haven't been able to do this with the methods introduced so far.

9.6 Talagrand's inequality: special case

As often happens with Euclidean distances, it's hard to calculate the mean of X, but analyzing X^2 is much easier. Let P be the projection operation onto the orthogonal complement of V, our d-dimensional subspace: then P is some matrix $\in \mathbb{R}^{n \times m}$, and we have

$$X^2 = \langle X, PX \rangle = \sum_{ij} x_i x_j p_{ij}.$$

Since the x_is are orthonormal, this just leads us to

$$\mathbb{E}[X^2] = \sum_i p_{ii} = \text{tr} P = n - d.$$

Notably, this expectation of X^2 does not depend on the orientation of V, though the distribution of X^2 does. So we should expect X to be concentrated about $\sqrt{n-d}$, and that gives us a center to work with. We're trying to claim that we have $O(1)$-concentration; specifically, we'd like to show that there is exponential decay with a constant deviation. To do this, we finally introduce the inequality we want:

Theorem 9.15 (Talagrand's inequality, simplified)

Let $A \subset \mathbb{R}^n$ be a **convex subset**, and let x be a uniform random point in the Boolean cube

$$x \sim \text{Unif}(\{0,1\}^n).$$

Then for all $t > 0$,

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

where we use the Euclidean distance.

Convexity here is extremely important. Talagrand is just not true otherwise - for example, consider A to be just the set of points in $\{0,1\}^n$ with weight (sum of entries) at least $\frac{n}{2}$. Then a random vertex is generally $O(\sqrt{n})$ away: specifically, there is probability at least $\frac{1}{4}$ that the weight of x is at most $\frac{n}{2} - c\sqrt{n}$ for some c.

Then the Euclidean distance is the square root of the Hamming distance on the Boolean cube, so the distance from x to A is on the order of $n^{1/4}$, which is not constant.

So what's really going on with this inequality? Given a convex set - for example, the convex hull of those same points in our Boolean cube - we're now measuring the distance to possibly some convex average of our vertices, and that distance is generally much smaller than if we were only allowed to use the vertices themselves.

Definition 9.16

Define a function f to be **quasi-convex** if all sets $\{f \leq a\}$ for $a \in \mathbb{R}$ are convex. (All convex functions are quasi-convex as well.)

Corollary 9.17

Let's say we have a function $f : \mathbb{R}^n \to \mathbb{R}$ that is quasi-convex and 1-Lipschitz with respect to the Euclidean distance: then for all $r \in \mathbb{R}, t > 0$, for x picked uniformly from the cube $\{0,1\}^n$,

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

This is a direct translation of the isoperimetric inequality. The theorem implies the corollary by letting A be the set of values $\{f \leq r\}$, which is convex if f is quasi-convex by definition. Since f is 1-Lipschitz by the triangle inequality, we have

$$f(x) \leq r + t \quad \forall x : \text{dist}(x, A) \leq t.$$

With this, we're now ready to answer our initial problem:

Theorem 9.18

Let V be a fixed d-dimensional subspace, and let $f(x) = \text{dist}(x, V)$. If we pick x uniformly on the cube $\{0,1\}^n$, then there exist constants $C, c > 0$ such that for all $t > 0$,

$$\Pr(|f - \mathbb{E}[f]| > t) \leq Ce^{-ct^2}.$$

Proof sketch. Let m be a median of f. Using Corollary 9.17, set $r = m$ to get the upper tail

$$\Pr(f \geq m + t) \leq 2e^{-t^2/4}.$$

Meanwhile, set $r = m - t$ to get the lower tail

$$\Pr(f \leq m - t) \leq 2e^{-t^2/4}.$$

We also mentioned that the median and the mean are very close for sub-Gaussian distributions. With some calculations, we can show that the median of X

$$\text{med}(X) = \sqrt{n-d} + O(1)$$

(with an absolute constant), or else we get inconsistency with tail bounds. Since $\mathbb{E}[f^2] = n - d$ and we have constant concentration, $\mathbb{E}[f]$ is also $\sqrt{n-d} + O(1)$, and thus we have constant deviation from the mean, as desired. □

So the whole point is that Talagrand is about **concentration of convex Lipschitz functions when evaluating at a random point of the Boolean cube.** We're not going to prove the inequality in class, because there are some tedious calculations involved. Instead, let's focus on combinatorial applications.

9.7 Random matrices

Let A be a random symmetric matrix with independent entries ± 1, where $a_{ij} = a_{ji}$. This can be thought of as being related to the adjacency matrix of a random graph.

It turns out the largest eigenvalue λ_1 is also the operator norm of A:

$$\lambda_1(A) = ||A||_{\text{op}}.$$

How well is this concentrated? We have about $O(n^2)$ variables, so Azuma's inequality gives something like $O(n)$ concentration about the mean. But this is pretty bad, because typically the largest eigenvalue

$$\lambda_1(A) \lesssim \sqrt{n},$$

so linear concentration doesn't really help at all. On the other hand, let's try to use Talagrand's inequality. We need to check a few things: consider the function $f : A \to ||A||_{\text{op}}$.

- Convexity comes from the fact that the operator norm is a norm, so we can use the triangle inequality.
- To show this function is 1-Lipschitz, we need

$$|f(x) - f(y)| \leq ||x - y||_2,$$

  where we're using the L^2 norm. This can be proved using Cauchy-Schwarz.

So now Talagrand's inequality tells us that we have constant-window concentration, independent of n. In other words, we've just showed that

$$\Pr(|\lambda_1(A) - \mathbb{E}(\lambda_1(A))| \geq t) \leq Ce^{-ct^2}$$

for some C, c, which decays like a Gaussian.

Fact 9.19

We actually know more about the concentration: it's actually $\Theta(n^{-1/6})$, and it converges to something called a Tracy-Widom distribution when normalized. Also, we know the mean of this distribution: the easiest way is to make the entries Gaussian instead of ± 1, but the answer is approximately $\sqrt{2n}$ regardless of the distribution.

As a sidenote, we can't actually use this method to prove the concentration of the second largest eigenvalue yet, since that's not convex as a function of our matrix entries. But the bottom line is that Talagrand's inequality is not just about the Boolean cube.

9.8 Talagrand's inequality in general

If we have a space $\Omega = \Omega_1 \times \cdots \times \Omega_n$, we may want to find the distance between two points.

Definition 9.20

Given a vector $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n_{\geq 0}$, define the **weighted Hamming distance**

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

This kind of distance is defined even if the individual Ω_is don't have metrics! So now if we have some subset $A \subset \Omega$ of our product space, we can define

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

This should still feel fairly familiar: for example, for $\Omega = \{0,1\}^n$, if we have a fixed α with $|\alpha| = 1$ (under the L^2 norm), we have

$$d_\alpha(x, y) = |\langle \alpha, 1_{x \neq y} \rangle|.$$

Azuma's inequality then tells us that for if we choose x uniformly on $\{0,1\}^n$, if we have a **fixed** α and subset $A \subset \{0,1\}^n$,

$$\Pr(|d_\alpha(x, A) - \mathbb{E}(d_\alpha(x, A))| \geq t) \leq 2e^{-t^2/2}.$$

(This is because the weights in Azuma's inequality satisfy $\sum c_i^2 = 1$ from the definition of α.) But Azuma only gives this to us for a fixed α: having this condition be true in all directions is much stronger, and that's what Talagrand's inequality tells us.

Definition 9.21

Define the **convex distance**

$$d_T(x, A) = \sup_{\substack{\alpha \in \mathbb{R}^n \\ |\alpha| = 1}} d_\alpha(x, A).$$

(Basically, choose the "worst" possible α: the one that separates x and A by the most.) This is easier to visualize if we think of $\Omega = \{0,1\}^n$ and $A \subset \Omega$ being a subset of that Boolean cube: $d_T(x, A)$ is then just the Euclidean distance from x to the convex hull of A. (In general, we do not need each coordinate to be limited to $\{0,1\}$, though: they can take on any set of values.)

Theorem 9.22 (Talagrand's inequality, general)

For any $A \subset \Omega = \Omega_1 \times \cdots \times \Omega_n$, let x be a random point in Ω with independent coordinates. Then

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

Here's another interpretation of the convex distance: we want to convert this to a "distance to convex hull" type argument even when we don't have a Boolean cube.

Definition 9.23

Define $U_A(x)$ to be the set of $s \in \{0,1\}^n$ such that there is some $y \in A$ so that $s_i = 1$ for all i with $x_i \neq y_i$. In other words, $s \in U_A(x)$ if the support of s contains the support of $x - y$ for some $y \in A$.

We can think of this as the "set of coordinates we need to change to get from x to A," ignoring the actual coordinates: notice that this is a subset of the Boolean cube even when A isn't. Notably, this is an increasing subset of the cube $\{0,1\}^n$.

Lemma 9.24

Letting dist be the Euclidean distance,

$$d_T(x, A) = \text{dist}(\vec{0}, \text{convex hull}(U_A(x))),$$

Proof. The left hand side is (by definition) the supremum over all weight vectors α of norm 1 of the α-distance between x and A, which is equivalent to looking at the closest point in A under this α: this is then

$$\sup_{\alpha} \inf_{y \in A} d_\alpha(x, y) = \sup_{\alpha} \inf_{y \in U_A(x)} (\alpha \cdot y).$$

Since A is convex, by von Neumann, we can swap the inf and sup as long as we extend to the convex hull

$$= \inf_{\substack{y \in \text{convex} \\ \text{hull}(U_A(x))}} \sup_{\alpha} (\alpha \cdot y) = \inf_{\substack{y \in \text{convex} \\ \text{hull}(U_A(x))}} |y|,$$

as desired. □

So how do we apply Talagrand? The idea is that we can adjust our α to favor certain coordinates and give us better bounds. α plays the role of a "certificate" that guarantees the existence of a small or large value. In particular, it follows from Talagrand that (rearranging)

$$\Pr(d_{\alpha(x)}(x, A) \geq t) \leq \frac{1}{\Pr(A)} e^{-t^2/4}.$$

Specifically, we can pick a different certificate α for each x:

Corollary 9.25

Let A, B be subsets of $\Omega = \Omega_1 \times \Omega_2 \cdots \times \Omega_n$. Suppose that for all $y \in B$, there exists an $\alpha = \alpha(y) \in \mathbb{R}^n_{\geq 0}$ so that for all $x \in A$,

$$d_\alpha(x, y) \geq t|\alpha|.$$

(This means the distance between A and B is large in the specific Talagrand sense.) Then

$$\Pr(A) \Pr(B) \leq e^{-t^2/4}.$$

To understand this, let's do another proof of the largest eigenvalue of our random matrix:

Proof. Let X be an $n \times n$ symmetric random matrix with independent entries in the interval $[-1, 1]$ (they can be distributed in any way, as long as they are independent and bounded). If we let $t > 0, M \in \mathbb{R}$, and we have the sets

$$A = \{X : \lambda_1(X) \leq M\}, B = \{X : \lambda_1(X) \geq M + t\},$$

we want to verify that for every matrix in B with large eigenvalue, we can certify this somehow: we pick some $\alpha(y)$ such that B is far away from A. Specifically, there exists some $\alpha \in \mathbb{R}^m$, where $m = \frac{n(n+1)}{2}$, such that $d_\alpha(x, y) \geq ct|\alpha|$ for all $x \in A$.

Let $\vec{v} \in \mathbb{R}^n$ be the top eigenvector corresponding to $\lambda_1(y)$. Then let

$$\alpha_{ij} = \begin{cases} v_i^2 & i = j \\ 2|v_i||v_j| & i \neq j; \end{cases}$$

the reason for doing this will become quickly apparent. By the Courant-Fischer characterization of the top eigenvector $\vec{v}$,

$$v^T Y v = \lambda_1(Y) \geq M + t,$$

and because X does not have large eigenvalue, we can set a contrasting bound for X:

$$v^T X v \leq \lambda_1(X) \leq M.$$

In particular, this means that we can use our eigenvector v to "separate" A and B:

$$t \leq v^T (X - Y) v,$$

and expanding out the difference as a bilinear form,

$$t \leq \sum_{i,j} v_i v_j (X_{ij} - Y_{ij}).$$

This is upper bounded by looking at only those where the two matrices differ in their entries:

$$\leq 2 \sum_{ij} |v_i||v_j| 1_{X_{ij} \neq Y_{ij}} \leq 2 d_\alpha(X, Y).$$

(Here, we used that the entries of X and Y are bounded by $[-1, 1]$.) Now note that the length of α is at most 2, and plug this into the corollary to get concentration. □

9.9 Increasing subsequences

Problem 9.26

Pick a uniformly random permutation $\sigma \in S_n$ of the first n integers. How long is the longest increasing subsequence?

Call this length X - it's important to note that we can skip entries (so subsequences don't need to be contiguous). For example, $5, 3, 1, 4, 6, 2, 7$ has longest increasing subsequence of length 4.

Our goal is to show that X is concentrated. Let's try to use the tools we have: first of all, let's try Azuma. We need independence of our underlying variables, so let's try to make our Ω_is independent: let $x_1, \cdots, x_n \sim \text{Unif}[0, 1]$ independently, and get a permutation of $[n]$ from the **relative orderings** of the x_is.

Then the length of the longest increasing subsequence changes by at most 1 if we change 1 coordinate, so it is 1-Lipschitz here. Azuma tells us that we have sub-Gaussian decay with a window size of $O(\sqrt{n})$.

How good is this? Let's do a first moment calculation to see the average size of X:

$$\Pr(X \geq k) \leq \binom{n}{k} \cdot \frac{1}{k!} \leq \frac{n^k}{(k!)^2},$$

since we pick any of the $\binom{n}{k}$ sequences of length k, and they have probability $\frac{1}{k!}$ of working.

So if $k = 100\sqrt{n}$, then this probability is $o(1)$, and thus we should expect the permutation to be no more than $c\sqrt{n}$ long. That means our concentration bound is bad! (In particular, any permutation of length n has either an increasing or decreasing sequence of length $\sqrt{n}$ by Pigeonhole.)

Let's see if Talagrand tells us anything better. The idea is that Talagrand is useful when we can "witness" rare events: showing such a sequence exists (that is, making the length **certificable**) doesn't use that many of the coordinates of $\vec{x}$. So Talagrand will actually tell us that we have fluctuations on the order of $O(\sqrt{x})$.

Here's that idea in more rigor:

Theorem 9.27

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$, and let $f : \Omega \to \mathbb{R}$ be a 1-Lipschitz function with respect to the Hamming distance. Suppose that we can verify

$$\{\omega : f(\omega) \geq r\}$$

by checking at most s coordinates. Then for every t,

$$\Pr(f(\omega) \leq r - t\sqrt{s}) \Pr(f(\omega) \geq r) \leq e^{-t^2/4}.$$

When we say **checking at most s coordinates** here, we specifically mean that with any ω with $f(\omega) \geq r$, there exists some subset $I \subset [n]$ with $|I| \leq S$ such that for all other ω' such that ω agrees with ω' on I, $f(\omega') \geq r$. In other words, knowing those s coordinates guarantees that our condition is true.

Proof. Let A, B be the sets

$$A : \{\omega : f(\omega) \leq r - t\sqrt{s}\}, B : \{\omega : f(\omega) \geq r\}.$$

Our goal is to check that for all $y \in B$, there exists $\alpha \in \mathbb{R}^n_{\geq 0}$ such that for all $x \in A$, $d_\alpha(x, y) \geq t|\alpha|$. (Basically, A is far away from B, even if we zoom in on I.)

But by definition of checking coordinates, there exists a set $I \subset [n]$ with $|I| \leq s$ for each $y \in B$. Here's the key: if we fix y and let $\alpha = 1_I$ (1 in the spots of I and 0 in the others), every $x \in A$ disagrees with y on at least $t\sqrt{s}$ coordinates of I (or else we could change x by less than $t\sqrt{s}$ coordinates and get x' to agree with y on I, meaning $f(x') \geq r$). This means that the weighted Hamming distance $d_\alpha(x, y) \geq t\sqrt{s} \geq t|\alpha|$, and now we can apply Talagrand's directly. □

Corollary 9.28

Given a 1-Lipschitz function on $f : \Omega = \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ (with respect to the Hamming distance), if $\{f \geq r\}$ can be verified by checking r coordinates, and m is a median of X, then for all t,

$$\Pr(X \leq m - t) \leq 2 \exp\left(-\frac{t^2}{4m}\right),$$

$$\Pr(X \geq m + t) \leq 2 \exp\left(-\frac{t^2}{4(m+t)}\right).$$

Proof. From the above theorem, renormalize t by a factor of $\sqrt{r}$: now we have

$$\Pr(f \leq r - t) \Pr(f \geq r) \leq e^{-t^2/(4r)}.$$

Setting $r = m$ gives the lower bound, and setting $R = m + t$ gets the lower bound. □

So now this applies directly to X for our increasing subsequence, since we can "witness" our event by just showing the subsequence itself. Note also that the median of X is $O(\sqrt{n})$, so now we know that

$$\Pr(|X - \mathbb{E}[X]| < s) = 1 - o(1) \text{ if } s \gg n^{1/4},$$

meaning we've found $\sqrt[4]{n}$-concentration.

But this is not the best possible result! In 1985, Vershik-Kerov showed that X is concentrated around $2\sqrt{n}$, and in fact, the limiting distribution was found by Baik-Deift-Johansson in 1999 to be

$$\frac{X - 2\sqrt{n}}{n^{1/6}} \to \text{Tracy-Widon distribution.}$$

(As we may remember, this is also the fluctuation of the top eigenvalue of a random matrix.)

18.218 Probabilistic Method in Combinatorics

Spring 2019