TOPICS IN HIGH-DIMENSIONAL PROBABILITY

NICHOLAS A. COOK*

Contents

Date: May 29, 2024.

Abstract. *These are day-to-day lecture notes for the graduate topics course Math 690-40 High-dimensional Probability, given at Duke in the spring semester of 2024. They have not been prepared for publication – in particular the references are incomplete, minimal effort has been made to keep notation uniform across all lectures, and some sections are intended as supplements to course references rather than stand-alone references. I am grateful to the students taking the course for pointing out errors as we go along!*

(From the course description:) This course aims to cover core topics in the theory of probability measures on high-dimensional Euclidean spaces as well as important applications. Topics will include the concentration of measure phenomenon, random matrices, suprema of random processes, hypercontractivity, entropy methods and Fourier analysis on the Boolean hypercube. We'll illustrate the theory with applications to areas such as graph theory, combinatorial optimization, high-dimensional statistics and compressed sensing, and statistical physics. A prior course in measure-theoretic probability would be ideal background.

0. Preliminaries

0.1. **General notation.** We often write $[n]$ for the discrete interval $\{1,\dots,n\}$. The set of all k-sets (i.e. sets of size k) in a set S is denoted $\binom{S}{k}$. For $p \in [1,\infty]$ the ℓ^p norms on $\mathbb{R}^n$ and $\mathbb{C}^n$ are denoted $\|x\|_p := (|x_1|^p + \cdots + |x_n|^p)^{1/p}$. The Euclidean inner product (dot product) on $\mathbb{C}^n$ is denoted $\langle x, y\rangle = \overline{x} \cdot y = \bar{x}_1 y_1 + \cdots + \bar{x}_n y_n$.

We use the notation $|A|$ for the cardinality of a finite set A. We also write $|J|$ for the length of an interval $J \subset \mathbb{R}$. There should be no confusion between these usages.

For a statement Q we write 1_Q for the Boolean variable that is 1 if and only if Q is true. When Q depends in a measurable way on a point ω in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we may denote the indicator variable $\mathbb{1}(Q) := 1_Q$ for typographical convenience.

0.2. **Asymptotic notation.** We will often write C, c (and C_0, c' etc.) for finite, positive universal constants, whose value may change from line to line. (C, C', C_0 etc. are understood to be sufficiently large – i.e. the statement will remain true if one replaces C by any larger value – and c, c', c_0 etc. are sufficiently small.) We may occasionally let these constants depend on a fixed parameter, and we'll warn the reader when this is the case.

We'll make use of the following standard asymptotic notation. For real quantities f, g that may depend on one or more parameters (such as the dimension n of a random vector or the variance σ^2 of a fixed distribution) we write $f = O(g)$ and $f \lesssim g$ to mean that $|f| \leq Cg$ for a sufficiently large universal constant $C > 0$ – that is, C is independent of all parameters. (It follows that g is non-negative.) If we allow C to depend on one or more parameters q then we indicate this with a subscript, writing $f = O_q(g)$, $f \lesssim_q g$.

When f is positive, $g = \Omega(f)$ and $g \gtrsim f$ mean $f \lesssim g$. We write $f = \Theta(g)$ and $f \asymp g$ to mean $f \lesssim g \lesssim f$.

While there's nothing asymptotic about the preceding notations *per se*, in practice they are used when one or more asymptotic parameters are present, e.g. when we are interested in the case that the dimension n of a random vector is large, and estimates $f = O(g)$ will only

be interesting when n is sufficiently large depending on the implicit constants.[1] For instance, an estimate

$$\mathbb{P}(E) \lesssim \exp(-cn) \tag{0.1}$$

on the probability of an event E is only nontrivial for $n \geq (\log C)/c$, where C is the implicit constant, since any probability is trivially bounded by 1.

Exercise 0.1. Suppose that for an event E depending on a positive integer parameter n we have an estimate

$$\mathbb{P}(E) \lesssim \exp(-n/K) \tag{0.2}$$

for some finite K. Show it follows that

$$\mathbb{P}(E) \leq 2\exp(-n/K') \tag{0.3}$$

for some $K' \lesssim K$.

We occasionally make use of the (more explicitly) asymptotic notation $o(\cdot), \omega(\cdot), \sim$. For real f, non-negative g, a real asymptotic parameter ε and $\varepsilon_0 \in [-\infty, \infty]$, we write $f = o_{\varepsilon \to \varepsilon_0}(g)$ to mean $f/g \to 0$ as $\varepsilon \to \varepsilon_0$ (and in particular that g is positive for all ε sufficiently close to ε_0). We will tend to suppress the subscript $\varepsilon \to \varepsilon_0$ when it is clear from the context (e.g. when the dimension n is tending to $+\infty$). When f is positive we may write $g = \omega_{\varepsilon \to \varepsilon_0}(f)$ to mean $f = o_{\varepsilon \to \varepsilon_0}(g)$. We write $f \sim_{\varepsilon \to \varepsilon_0} g$ to mean $f/g \to 1$ as $\varepsilon \to \varepsilon_0$. When the rate of convergence may depend on some other parameters q we write $f = o_{\varepsilon \to \varepsilon_0; q}(g)$ (or $f = o_q(g)$ when the asymptotic parameter is clear from the context), and $g = \omega_q(f)$, $f \sim_q g$ etc.[2]

Exercise 0.2. Show that for positive real number a, b we have

$$\max(a, b) \asymp a + b \tag{0.4}$$

and

$$\min\{a, b\} \asymp \frac{a}{1 + \frac{a}{b}}. \tag{0.5}$$

In particular, $\min\{a^2, a\} \asymp a^2/(1 + a)$.

As the reader will likely appreciate as we go through the course, the use of unspecified constants C, c and asymptotic notation $O(\cdot)$ can save a lot of ink and make arguments much easier to read and remember. The notation may be uncomfortable at first, but mastering it trains one to keep in mind which error terms we're really "fighting" and which parameter regimes we need to be careful of.

In short, our emphasis in this course will be on the fundamental ideas driving the basic results. While the pursuit of optimal explicit constants is both interesting and useful, it often introduces several extra details to the proofs that can distract from the fundamental ideas. For the sharp forms of some of the results we'll cover you can consult the course references, such as [BLM13, Led01].

[1] This point is perhaps worth clarifying since this notation is ubiquitous in the so-called *non-asymptotic theory* of random matrices, which will be partially covered in this course (for a nice survey see [RV10]). While many papers in that literature avoid the notation $O(\cdot), \lesssim$ and instead use only unspecified constants C, c etc. this choice is purely stylistic (usually their values are not unreasonable and can be extracted from the proofs without much effort).

[2] The notations $o(\cdot), \omega(\cdot), \sim$ mirror the notations $O(\cdot), \Omega(\cdot), \asymp$. It is tempting to analogously give meaning to the symbols $\ll, \gg$ to mirror $\lesssim, \gtrsim$, as is sometimes done in the literature, but we refrain from doing this to avoid confusion with their use in analytic number theory, where the *Vinogradov notation* $\ll$ is synonymous with our definition of $\lesssim$.

1. Jan 11: Introduction

Today we:

- See examples of the concentration of measure phenomenon on the high-dimensional sphere $\mathbb{S}^{d-1}$ and the discrete cube $\{-1,1\}^n$.
- Begin an application of concentration on the cube to prove the Johnson–Lindenstrauss lemma for dimension reduction of high-dimensional data.

1.1. **Concentration on the sphere.** Write $\mathbb{B}^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ for the closed Euclidean unit ball in $\mathbb{R}^d$, and $\mathbb{S}^{d-1}$ for its boundary. The ball of radius r is $r\mathbb{B}^d$, where we write $rA = \{ra : a \in A\}$ for the dilation of a set $A \subset \mathbb{R}^d$ by r. Let $v_d(r)$ be the d-dimensional volume (Lebesgue measure) of $r\mathbb{B}^d$. Note that $v_d(r) = v_d(1)r^d$. Let $s_{d-1}(r)$ be the $d-1$-dimensional surface measure of $r\mathbb{S}^{d-1}$ – that is,

$$s_{d-1}(r) = \lim_{\varepsilon \downarrow 0} \frac{\mathrm{vol}_d((r+\varepsilon)\mathbb{B}^d) - \mathrm{vol}_d(r\mathbb{B}^d)}{\varepsilon} = v_d'(r) = dv_d(1)r^{d-1}. \tag{1.1}$$

(You can check this and the following formulas for $d = 2, 3$.)

Proposition 1.1 (Volume of Euclidean d-ball)**.** *We have* $s_{d-1}(1) = 2\pi^{d/2}/\Gamma(d/2)$ *and* $v_d(1) = \frac{\pi^{d/2}}{\frac{d}{2}\Gamma(\frac{d}{2})}$.

Recall the gamma function is given by $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$. By change of variable we have

$$\Gamma(\alpha) = 2\int_0^\infty e^{-r^2}r^{2\alpha-1}dr.$$

which will be used in the proof.

Proof of Proposition 1.1. Recall that $I := \int_{-\infty}^\infty e^{-x^2}dx = \sqrt{\pi}$. Indeed, by squaring and changing to polar coordinates we get

$$I^2 = \int_{\mathbb{R}^2} e^{-x^2-y^2}dxdy = \int_0^{2\pi}\int_0^\infty e^{-r^2}rdrd\theta = \pi.$$

Hence,

$$\pi^{d/2} = I^d = \int_{\mathbb{R}^d} \exp(-x_1^2 - \cdots - x_d^2)dx_1\cdots dx_d = s_{d-1}(1)\int_0^\infty e^{-r^2}r^{d-1}dr = \frac{1}{2}s_{d-1}(1)\Gamma(\frac{d}{2}).$$

Rearranging yields the claimed formula for $s_{d-1}(1)$ The formula for $v_d(1)$ follows from this and (1.1). □

From Stirling's approximation for the gamma function we have

$$v_d(1) \sim \frac{1}{\sqrt{\pi d}}(\frac{2\pi e}{d})^{d/2}$$

as $d \to \infty$. To leading exponential order this is $\exp(-(1+o(1))\frac{1}{2}d\log d)$, which is very small compared to 1, the volume of the cube of side-length 1. This is a phenomenon not seen in dimensions 2 and 3, where the origin-centered cube of side-length 1 is entirely contained in the ball of radius 1. So in high dimensions, the cube "pokes through" the sphere in many

places (it has 2^d corners at distance $\sqrt{d/2}$ from the origin) while the centers of its $2d$ faces are always at distance $\frac{1}{2}$ from the origin. Apparently, most of its volume is out near the corners at scale $\sqrt{d}$. (We'll see quantitative versions of this later.)

Now we present a simple calculation that is our first example of the concentration of measure phenomenon in high dimensions.

Proposition 1.2 (Measure of spherical caps)**.** *Let μ_d be the normalized surface measure on $\mathbb{S}^{d-1}$, thus $\mu_d(\mathbb{S}^{d-1}) = 1$. For any $\varepsilon > 0$,*

$$\mu_d(\{x \in \mathbb{S}^{d-1} : |x_1| > \varepsilon\}) \leq 2\exp(-c\varepsilon^2 d). \tag{1.2}$$

Or to put it probabilistically: if $U = (U_1, \ldots, U_d) \sim \mu_d$ is a uniform random unit vector in $\mathbb{R}^d$, then

$$\mathbb{P}(|U_1| > \varepsilon) \leq 2\exp(-c\varepsilon^2 d) \qquad \forall \varepsilon \geq 0. \tag{1.3}$$

Thus, an ε-widening of the "equator" $\{x \in S^{d-1} : x_1 = 0\}$ contains all but an exponentially small (in d) proportion of the surface measure! And 99% of the surface measure is within distance $O(d^{-1/2})$ of the equator. We say that the surface measure "concentrates" near the equator. But of course by symmetry, we have that the surface measure concentrates near *any* fixed equator!

Note this implies that the intersection of the ε-neighborhoods of k orthogonal equators still has measure at least $1 - 2k\exp(-c\varepsilon^2 d)$ (by the union bound), which is still very close to 1 when ε is fixed (at say $\frac{1}{100}$) and k is sub-exponential in d. This is quite different from our low-dimensional experience, where the intersection of the ε-neighorhoods of two orthogonal equators of $\mathbb{S}^2$ – say "the" equator and the prime meridian – is a $O(\varepsilon)$-neighborhood of the north and south poles and thus has measure $O(\varepsilon^2)$.

Proof of Proposition 1.2. We'll use the probabilistic notation of (1.3). Clearly we may assume $\varepsilon \leq 1$ since the left hand side of (1.3) is zero otherwise. Since the left hand side of (1.3) is trivially bounded by 1, by lowering the constant c on the right hand side we may assume

$$\varepsilon \geq C/\sqrt{d} \tag{1.4}$$

for any fixed constant $C > 0$. We may similarly assume d is sufficiently large.

By considering a slice of the $d-1$-sphere at height y in the e_1 direction, we see that the density $f_{U_1}(y)$ for the distribution of U_1 is

$$\frac{s_{d-2}(\sqrt{1-y^2})}{s_{d-1}(1)} = \frac{s_{d-2}(1)}{s_{d-1}(1)}(1-y^2)^{(d-2)/2} \asymp d^{1/2}(1-y^2)^{(d-2)/2} \tag{1.5}$$

where the last estimate can be obtained from the formula from Proposition 1.1 (exercise).

By symmetry it suffices to show

$$\mathbb{P}(U_1 > \varepsilon) \leq \exp(-c\varepsilon^2 d) \tag{1.6}$$

for any $\varepsilon \leq 1$ as in (1.4). Integrating the density of U_1 gives

$$\begin{aligned}
\mathbb{P}(U_1 > \varepsilon) &\asymp \int_\varepsilon^1 d^{1/2}(1-y^2)^{(d-2)/2} dy \\
&\leq d^{1/2}\varepsilon^{-1} \int_\varepsilon^\infty y\exp(-(\tfrac{d}{2}-1)y^2) dy \\
&\lesssim \frac{1}{\varepsilon d^{1/2}} \exp(-(\tfrac{d}{2}-1)\varepsilon^2)
\end{aligned}$$

where in the second line we bounded $1 - y^2 \leq \exp(-y^2)$, extended the integral to $[\varepsilon, \infty)$, and inserted a factor $y/\varepsilon \geq 1$ in the integrand. Taking C in (1.4) sufficiently large (and assuming $d \geq 3$) we obtain the desired bound (1.6). □

1.2. **Concentration (and anti-concentration) on the discrete hypercube.** Now consider the discrete hypercube $\{-1,1\}^n$ in $\mathbb{R}^n$ with the normalized counting measure ν_n (we switch to n for the dimension). (Much of what we'll say extends with minor modification to the uniform measure on the solid cube $[-1,1]^n$.) Consider a random vector $X = (X_1, \ldots, X_n) \in \{-1,1\}^n$ with distribution ν_n. Equivalently, $X_1, \ldots, X_n$ are iid Rademacher variables.[3] We denote the unit vector in the all-ones direction

$$v := \frac{1}{\sqrt{n}}(1, \ldots, 1) \in \mathbb{S}^{n-1}.$$

We consider the distribution of $\langle X, v\rangle$, the projection of X to the all-ones direction.

From the law of large numbers we have $n^{-1/2}\langle X, v\rangle \to 0$ in probability as $n \to \infty$. Thus, while $\{-1,1\}^n$ has diameter $2\sqrt{n}$ in the direction of v, most of the mass is within $o(\sqrt{n})$ of the hyperplane $H_v = \{x \in \mathbb{R}^d : \langle x, v\rangle = 0\}$.

Moreover, the central limit theorem tells us that $\langle X, v\rangle \xrightarrow{d} G$, where $G \sim N(0,1)$ is a standard Gaussian variable. So most of the measure ν_n concentrates within distance $O(1) + o_{n\to\infty}(1)$ of H_v.

Remark 1.3 (An aside on anti-concentration and quantitative CLTs). The Berry–Esseen theorem gives a quantitative version of the CLT – in this setting it states that for any interval $J \subset \mathbb{R}$,

$$\mathbb{P}(\langle X, v\rangle \in J) = \mathbb{P}(G \in J) + O(n^{-1/2}). \tag{1.7}$$

Thus, the discrete random variable $\langle X, v\rangle$ is effectively smooth and Gaussian at scales much larger than $n^{-1/2}$.

As a preview, when we come to the topic of *anti-concentration* later in the course, we'll see that (1.7) in fact holds with v replaced by any suitably "generic" fixed $u \in \mathbb{S}^{n-1}$, and moreover, for most choices of u we actually get an improved error of size $O_K(n^{-K})$ for any fixed $K \geq 1$. Note that some "genericity" assumption is necessary as (1.7) clearly fails if we replace v with a standard basis vector.

Much of this course will explore how wide classes of random variables (generally functions of many independent variables) behave in some ways like Gaussians; in particular they display approximate versions of the following two nice properties of the Gaussian distribution:

(1) (Concentration). Gaussians have sub-Gaussian tails:

$$\mathbb{P}(|G| \geq t) \leq 2\exp(-ct^2).$$

(2) (Anti-concentration). Gaussians have bounded density:

$$\mathbb{P}(G \in J) = O(|J|) \qquad \text{for any interval } J \subset \mathbb{R}$$

where $|J|$ is the length (Lebesgue measure) of J.

[3] A *Rademacher variable* (or *random sign*) is a random variable that is uniform in $\{-1,1\}$.

Returning to the concentration of measure phenomenon for the measure space $(\{-1,1\}^n, \nu_n)$, a sharp quantitative form of the law of large numbers is provided by the following result (a special case of Hoeffding's inequality):

Theorem 1.4. *For any fixed unit vector $u \in \mathbb{S}^{n-1}$,*

$$\mathbb{P}(|\langle X, u\rangle| \geq t) \leq 2\exp(-ct^2) \qquad \forall t \geq 0. \tag{1.8}$$

Thus, 99% of the measure of the discrete cube (a set of diameter $\Theta(\sqrt{n})$) is contained within a $O(1)$-neighborhood of any fixed hyperplane through the origin.

We'll see a proof of Theorem 1.4 next time. For the remainder of this first lecture we turn to an important application.

1.3. **Application: dimension reduction for high-dimensional data.** Concentration of measure on the discrete hypercube can be used to establish the following:

Theorem 1.5 (Johnson–Lindenstrauss lemma)**.** *Let $x_1, \ldots, x_m$ be fixed (deterministic) points in $\mathbb{R}^N$. Let A be an $N \times d$ matrix of iid Rademacher variables (equivalently, A is uniform random in the Nd-dimensional discrete cube $\{-1,1\}^{N\times d}$) and for each $i \in [m]$ set*

$$y_i := \frac{1}{\sqrt{d}} A^{\mathsf{T}} x_i. \tag{1.9}$$

For any $\varepsilon \in (0,1)$, if

$$d \geq C\varepsilon^{-2} \log m \tag{1.10}$$

then

$$1 - \varepsilon \leq \frac{\|y_i - y_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \varepsilon \qquad \forall 1 \leq i < j \leq m \tag{1.11}$$

except with probability at most $\exp(-c\varepsilon^2 d)$.

Remark 1.6. Recall our convention from Section 0.2 – which will tend to go unmentioned in the sequel – that $C, c > 0$ denote universal constants that are sufficiently large and small, respectively.

Remark 1.7. Perhaps the most surprising and useful features of Theorem 1.5 are that

(1) N makes no appearance in (5.4), (5.5) and the probability bound;
(2) the reduced dimension d can be as small as logarithmic in m;
(3) one can achieve these with a simple linear transformation, and moreover with almost any sign matrix![4]

The key fact for the proof of Theorem 1.5 is the following concentration of measure bound for the norm of the image of a fixed vector under A^{T}.

Lemma 1.8. *For any fixed $u \in \mathbb{S}^{N-1}$ and $\varepsilon \in (0,1)$,*

$$\mathbb{P}\Big(\Big|\frac{1}{d}\|A^{\mathsf{T}}u\|_2^2 - 1\Big| \geq \varepsilon\Big) \leq 2\exp(-c\varepsilon^2 d). \tag{1.12}$$

[4]We'll later see that the same holds for a much wider class of random matrices.

We'll prove Lemma 1.8 next time. For now we just note that the mapping $d^{-1/2}A^{\mathsf{T}}$ does preserve squared ℓ_2-norms in expectation: denoting the columns of A by $A_1, \dots, A_d \in \{-1,1\}^N$, we have

$$\mathbb{E}\|A^{\mathsf{T}}u\|_2^2 = \mathbb{E}\sum_{j=1}^{d}\langle A_j, u\rangle^2 = \sum_{j=1}^{d}\sum_{k,\ell=1}^{N}\mathbb{E}A_{kj}A_{\ell j}u_k u_j = \sum_{j=1}^{d}\|u\|_2^2 = d \tag{1.13}$$

so (1.12) indeed provides concentration of $\|A^{\mathsf{T}}u\|_2^2$ about its expectation.

2. Jan 18: Exponential moments and sub-Gaussian distributions

- We conclude our first proof of Theorem 1.5.
- Along the way we state and proof Hoeffding's inequality, as well as an extension to sums of sub-exponential random variables.
- We define the classes of sub-Gaussian and sub-exponential random variables and state some equivalent characterizations.

2.1. Dimension reduction (continued).

We begin by using Lemma 1.8 to conclude the

Proof of Theorem 1.5. Without loss of generality we may assume all of the points x_i are distinct. Denote the $\binom{m}{2}$ "bad" events

$$\mathcal{B}_{ij} := \left\{ \left| \frac{\|y_i - y_j\|_2^2}{\|x_i - x_j\|_2^2} - 1 \right| > \varepsilon \right\}, \qquad 1 \le i < j \le m.$$

For fixed $i < j$, applying Lemma 1.8 with $u = (x_i - x_j)/\|x_i - x_j\|_2$ gives

$$\mathbb{P}(\mathcal{B}_{ij}) = \mathbb{P}\Big(\Big| \frac{1}{d}\|A^{\mathsf{T}}u\|_2^2 - 1 \Big| > \varepsilon \Big) \le 2\exp(-c\varepsilon^2 d).$$

We can then apply the union bound to conclude

$$\begin{aligned}\mathbb{P}((5.5) \text{ fails}) &= \mathbb{P}\Big(\bigcup_{1\le i<j\le m} \mathcal{B}_{ij} \Big) \\ &\le \sum_{1\le i<j\le m} \mathbb{P}(\mathcal{B}_{ij}) \le m^2 \exp(-c\varepsilon^2 d) \le \exp(-\tfrac{1}{2}c\varepsilon^2 d)\end{aligned}$$

where in the final bound we took the constant C in (5.4) sufficiently large. □

2.2. Hoeffding's inequality.

We have the following generalization of Theorem 1.4.

Theorem 2.1 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in [a_i, b_i]$ a.s. for each i, and set $S_n := X_1 + \cdots + X_n$. Then*

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \ge t) \le 2\exp(-ct^2/B^2) \qquad \forall t \ge 0 \tag{2.1}$$

where $B^2 := \sum_{i=1}^n (b_i - a_i)^2$.

Proof. Since $S_n - \mathbb{E}S_n = \sum_{i=1}^n X_i - \mathbb{E}X_i$, by replacing X_i with $X_i - \mathbb{E}X_i$ we may assume without loss of generality that $\mathbb{E}X_i = 0$ for each i, and in particular $\mathbb{E}S_n = 0$. By symmetry, it suffices to show

$$\mathbb{P}(S_n \ge t) \le \exp(-ct^2/B^2) \qquad \forall t \ge 0. \tag{2.2}$$

From the pointwise bound $1_{x\geq 0} \leq e^x$ for $x \in \mathbb{R}$, we have

$$\mathbb{P}(S_n \geq t) = \mathbb{E}1_{S_n - t \geq 0} \leq \mathbb{E}\exp(\lambda(S_n - t)) = \exp(-\lambda t)\mathbb{E}\exp(\lambda S_n) \tag{2.3}$$

for any $\lambda > 0$. Letting $\lambda > 0$ to be chosen later, we thus seek a bound on the moment generating function $\mathbb{E}\exp(\lambda S_n)$. By independence,

$$\mathbb{E}\exp(\lambda S_n) = \prod_{i=1}^n \mathbb{E}\exp(\lambda X_i).$$

Now for arbitrary $i \in [n]$, we claim

$$\mathbb{E}\exp(\lambda X_i) = \exp(O(\lambda^2 (b_i - a_i)^2)). \tag{2.4}$$

Indeed,

$$\begin{aligned}
\mathbb{E}\exp(\lambda X_i) &= \mathbb{E}[1 + \lambda X_i + \frac{\lambda^2 X_i^2}{2!} + \cdots] \\
&= 1 + \mathbb{E}\sum_{k\geq 2} \frac{\lambda^k X_i^k}{k!} \\
&\leq 1 + \sum_{k\geq 2} \frac{\lambda^k (b_i - a_i)^k}{k!} \\
&= \exp(\lambda(b_i - a_i)) - \lambda(b_i - a_i)
\end{aligned}$$

where in the second line we used the assumption that $\mathbb{E}X_i = 0$. The bound (2.4) now follows from the pointwise bound $e^x \leq x + e^{Cx^2}$ for $x \in \mathbb{R}$. Substituting (2.4) into the product we obtain

$$\mathbb{E}\exp(\lambda S_n) = \exp(O(\lambda^2 B^2)). \tag{2.5}$$

combining with (2.3) we get

$$\mathbb{P}(S_n \geq t) \leq \exp(-\lambda t + C\lambda^2 B^2).$$

Finally, choosing $\lambda = t/(2CB^2)$ yields (2.2) to complete the proof. □

2.3. **Dimension reduction (part 3).** We turn to the proof of Lemma 1.8. We'll actually see two different proofs (one may wait until next time). Our first proof makes use of Theorem 1.4.

Setting $Z_j := \langle A_j, u\rangle$, we want to show that

$$\|A^{\mathsf{T}} u\|_2^2 = \sum_{j=1}^d Z_j^2 \tag{2.6}$$

concentrates around its expectation of d (as shown in (1.13)). We note that $\|A^{\mathsf{T}} u\|_2^2$ is a sum of iid random variables, so it is reasonable to expect something like Hoeffding's inequality to give this. However, even the general version of Theorem 2.1 would only be effective if the variables Z_j^2 were bounded uniformly in n.

Theorem 1.4 does tell us that each Z_j has a light tail. Indeed, A_j is a length-N Rademacher vector, so

$$\mathbb{P}(|Z_j| \geq t) \leq 2\exp(-ct^2) \quad \forall t \geq 0.$$

This means that Z_j^2 has a sub-exponential tail:

$$\mathbb{P}(Z_j^2 \geq t) \leq 2\exp(-ct) \quad \forall t \geq 0. \tag{2.7}$$

Lemma 2.2. *Let $Y_1, \dots, Y_n$ be independent, centered variables with sub-exponential tails, in the sense that*

$$\mathbb{P}(|Y_i| \geq t) \leq 2\exp(-t/K) \qquad \forall t \geq 0,\ \forall i \in [n] \tag{2.8}$$

for some finite K. Setting $S_n := \sum_{i=1}^n Y_i$, we have

$$\mathbb{P}(|S_n| \geq Ksn) \leq 2\exp(-\frac{cs^2}{1+s}n) \qquad \forall s \geq 0. \tag{2.9}$$

Before proving the lemma we conclude the

Proof of Lemma 1.8. With $Z_j = \langle A_j, u\rangle$ as above, denote the centered variables $Y_j := Z_j^2 - 1$. From (2.7) and lowering c we have

$$\mathbb{P}(|Y_j| \geq t) \leq 2\exp(-ct) \qquad \forall t \geq 0,\ \forall j \in [d]. \tag{2.10}$$

We may hence apply Lemma 2.2, with $n = d$, $S_d = \sum_{i=1}^d Y_i = \|A^{\mathsf{T}}u\|_2^2 - d$, $K = O(1)$, and $s = c\varepsilon$ for a sufficiently small constant $c > 0$ to obtain

$$\mathbb{P}(|\|A^{\mathsf{T}}u\|_2^2 - d| \geq \varepsilon d) \leq 2\exp(-c\varepsilon^2 d)$$

as desired. □

Proof of Lemma 2.2. By replacing Y_i with Y_i/K we may assume $K = 1$.

As in the proof of Theorem 2.1 we proceed by bounding the moment generating function of S_n. We claim

$$\mathbb{E}\exp(\lambda S_n) = \exp(O(\lambda^2 n)) \tag{2.11}$$

if $\lambda \in (0, c_0)$ for a sufficiently small universal constant $c_0 > 0$. From this the lemma follows by similar lines as in the proof of Theorem 2.1 (exercise!).

By independence, to prove (2.11) it suffices to show

$$\mathbb{E}\exp(\lambda Y_i) = \exp(O(\lambda^2)) \tag{2.12}$$

for each $i \in [n]$ and all $\lambda \in (0, c_0)$. To that end, we expand

$$\begin{aligned}\mathbb{E}\exp(\lambda Y_i) &= \mathbb{E}\left(1 + \lambda Y_i + \sum_{k\geq 2}\frac{\lambda^k Y_i^k}{k!}\right)\\ &= 1 + \mathbb{E}\sum_{k\geq 2}\frac{\lambda^k Y_i^k}{k!}\\ &\leq 1 + \sum_{k\geq 2}\frac{\lambda^k \mathbb{E}|Y_i|^k}{k!}\end{aligned}$$

where in the second line we used the assumption that the Y_i are centered. Using the tail assumption (2.8) we can bound the absolute moments as follows:

$$\mathbb{E}|Y_i|^k = O(k!) \tag{2.13}$$

for all $k \in \mathbb{N}$ (exercise!). Substituting this in the previous line gives

$$\mathbb{E}\exp(\lambda Y_i) \leq 1 + \sum_{k\geq 2} O(\lambda)^k = 1 + O(\lambda^2) = \exp(O(\lambda^2)) \tag{2.14}$$

taking c_0 sufficiently small so that the geometric series converges. This gives (2.12) to conclude the proof. □

Exercise 2.1. Fill in the steps marked (exercise!) in the above proof. Note we could just as well write $\min\{s^2, s\}$ in place of $s^2/(1+s)$ in the exponential in (2.9) (recall (0.5)).

In (2.9) we see a tail of a different shape – often called a "Bernstein-type tail" – than the sub-Gaussian tail $\exp(-cs^2n)$ from Hoeffding's inequality. In particular, for smaller deviations with $s \lesssim 1$ (2.9) gives

$$\mathbb{P}(|S_n| \geq Ksn) \leq 2\exp(-cs^2n) \tag{2.15}$$

while for larger deviations with $s \gtrsim 1$,

$$\mathbb{P}(|S_n| \geq Ksn) \leq 2\exp(-csn) \tag{2.16}$$

(up to modification of the constant c). If the Y_i were bounded rather than just sub-exponential then Theorem 2.1 would give the Gaussian tail (2.15) for all $s \geq 0$. However, the exponential tail of (2.16) is necessary for $s \gtrsim 1$, as shown in the following:

Exercise 2.2. Show that the bound (2.16) is sharp for $s \geq 1$ (up to modifying the universal constant $c > 0$) for any n, by considering the case that the Y_i are iid centered exponential random variables (with density proportional to $\exp(-|y|)$, say). *(Hint: note for instance that the left hand side in (2.16) is bounded below by the probability of the event that $Y_1 \approx 2Ksn$ and $\sum_{i=2}^n Y_i = O(K\sqrt{n})$.)*

Bernstein-type tails also commonly arise for sums of bounded random variables for which the variance is much smaller than the L^∞-norm, such as Bernoulli(p) variables with p very small.

Theorem 2.3. *Let $X_1, \dots, X_n$ be independent real random variables with $\mathrm{Var}(X_i) \leq \sigma_i^2$ and $|X_i| \leq b$ a.s. for each i. Then with $S_n = X_1 + \cdots + X_n$,*

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq t) \leq 2\exp(-\frac{ct^2}{\sigma^2 + bt}) \qquad \forall t \geq 0 \tag{2.17}$$

where $\sigma^2 := \sum_{i=1}^n \sigma_i^2$.

Exercise 2.3. Prove Theorem 2.3. *(Hint: after centering the variables X_i and normalizing b to be 1, follow the general approach of the proofs of Theorem 2.1 and Lemma 2.2, but in place of (2.4) and (2.12) prove the bound*

$$\mathbb{E}\exp(\lambda X_i) = \exp(O(\lambda^2\sigma_i^2)) \qquad \forall \lambda \in [0,1]\,.) \tag{2.18}$$

2.4. **Sub-Gaussian and sub-exponential distributions.** Since many of our objectives in this first part of the course will be to show that various random variables have sub-Gaussian tails, it will be useful to formalize this property in a definition and establish some equivalent properties.

Definition 2.4 (Sub-Gaussian variable)**.** For a constant $K > 0$, we say that a real-valued random variable X is *K-sub-Gaussian* if

$$\mathbb{E}\exp(X^2/K^2) \leq 2. \tag{2.19}$$

The best constant K is called the ψ_2-norm of X:

$$\|X\|_{\psi_2} := \inf\{K > 0 : \mathbb{E}\exp(X^2/K^2) \leq 2\}. \tag{2.20}$$

More generally, we say that a random vector $X \in \mathbb{R}^d$ is K-sub-Gaussian if $\langle X, u\rangle$ is K-sub-Gaussian for every fixed unit vector $u \in \mathbb{S}^{d-1}$. If X is K-sub-Gaussian for some finite K then we may simply say that X is sub-Gaussian.

Exercise 2.4. Show that $\|\cdot\|_{\psi_2}$ indeed defines a norm on the space of sub-Gaussian random variables.

An immediate consequence of (2.28) and Markov's inequality is that X has sub-Gaussian tails:

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K^2) \qquad t \geq 0. \tag{2.21}$$

In fact the reverse implication holds (up to modification of K by a universal constant factor). These as well as a couple of other useful equivalent properties are summarized in the following:

Proposition 2.5 (Equivalent characterizations of sub-Gaussian variables)**.** *Let X be a real-valued random variable. The following are equivalent, in the sense that if property (i) holds, then property (j) also holds with $K_j = O(K_i)$ (all constants K_i are assumed to be positive and finite).*

(1) X is K_1-sub-Gaussian:

$$\mathbb{E}\exp(X^2/K_1^2) \leq 2. \tag{2.22}$$

(2) X has sub-Gaussian tails:

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_2^2) \qquad \forall t \geq 0. \tag{2.23}$$

(3) X has sub-Gaussian L^p-norms:

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq K_3\sqrt{p} \qquad \forall p \geq 1. \tag{2.24}$$

Moreover, if $\mathbb{E}X = 0$, then the above properties are equivalent to the following (with $K_i \asymp K_4$ for $i = 1, 2, 3$):

(4) X has sub-Gaussian moment generating function:

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_4^2\lambda^2) \qquad \forall \lambda \in \mathbb{R}. \tag{2.25}$$

Proof. See [Ver18, §2.5.1]. (Many of the arguments there were already used in the proofs of Theorem 2.1 and Lemma 2.2.) □

Of course, Gaussians are examples of sub-Gaussian random variables, as are bounded random variables – indeed, you can check that

$$\|X\|_{\psi_2} \lesssim \|X\|_{L^\infty}. \tag{2.26}$$

Theorem 1.4 states that for uniform random $X \in \{-1, 1\}^n$ and fixed $u \in \mathbb{S}^{n-1}$, $\langle X, u\rangle$ is $O(1)$-sub-Gaussian. From this one sees that if $Y \sim \mathrm{Bin}(n, \frac{1}{2})$ then $Y - \mathbb{E}Y$ is $O(\sqrt{n})$-sub-Gaussian.

Using (2.5) we can prove the following generalization of Theorem 2.1 with a short argument.

Theorem 2.6 (Hoeffding's inequality for sub-Gaussian variables)**.** *Let $X_1, \dots, X_n$ be independent random variables such that for each $i \in [n]$, $X_i - \mathbb{E}X_i$ is K_i-sub-Gaussian, and let $S_n = X_1 + \cdots + X_n$. Then $S_n - \mathbb{E}S_n$ is $O(K)$-sub-Gaussian, where $K := (\sum_{i=1}^n K_i^2)^{1/2}$.*

Indeed, Theorem 2.1 follows from Theorem 2.6 and (2.26).

We can equivalently state Hoeffding's inequality in terms of the sub-Gaussian norm as

$$X_1, \dots, X_n \text{ independent} \implies \Big\| \sum_{i=1}^n X_i \Big\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|X_i\|_{\psi_2}^2 . \tag{2.27}$$

Proof of Theorem 2.6. Since $S_n - \mathbb{E}S_n = \sum_{i=1}^n X_i - \mathbb{E}X_i$ it suffices to establish the claim under the assumption that $\mathbb{E}X_i = 0$ for each i. For arbitrary $\lambda \in \mathbb{R}$, by the characterization (4) of sub-Gaussian variables in Proposition 2.5 we have $\mathbb{E}\exp(\lambda X_i) = \exp(O(K_i^2\lambda^2))$ for each i, and by independence,

$$\mathbb{E}\exp(\lambda S_n) = \prod_{i=1}^n \mathbb{E}\exp(\lambda X_i) = \prod_{i=1}^n \exp(O(K_i^2\lambda^2)) = \exp(O(K^2\lambda^2)).$$

The claim now follows by another application of Proposition 2.5(4). □

In the proof of Theorem 1.5 we encountered variables satisfying the following weaker tail hypothesis.

Definition 2.7 (Sub-exponential variable)**.** For a constant $K > 0$, we say that a real-valued random variable X is *K-sub-exponential* if

$$\mathbb{E}\exp(|X|/K) \le 2. \tag{2.28}$$

The best constant K is called the ψ_1-norm of X:

$$\|X\|_{\psi_1} := \inf\{K > 0 : \mathbb{E}\exp(|X|/K) \le 2\}. \tag{2.29}$$

More generally, we say that a random vector $X \in \mathbb{R}^d$ is K-sub-exponential if $\langle X, u\rangle$ is K-sub-exponential for every fixed unit vector $u \in \mathbb{S}^{d-1}$. If X is K-sub-exponential for some finite K then we may simply say that X is sub-exponential.

Proposition 2.8 (Equivalent characterizations of sub-exponential variables)**.** *Let X be a real-valued random variable. The following are equivalent, in the sense that if property (i) holds, then property (j) also holds with $K_j = O(K_i)$.*

(1) X has a finite absolute exponential moment: for some $K_1 \in (0, \infty)$,

$$\mathbb{E}\exp(|X|/K_1) \le 2 \tag{2.30}$$

(2) X has sub-exponential tails: for some $K_2 \in (0, \infty)$,

$$\mathbb{P}(|X| \ge t) \le 2\exp(-t/K_2) \qquad \forall t \ge 0. \tag{2.31}$$

(3) X has sub-exponential L^p-norms: for some $K_3 \in (0, \infty)$,

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \le K_3 p \qquad \forall p \ge 1. \tag{2.32}$$

Moreover, if $\mathbb{E}X = 0$, then the above properties are equivalent to the following (with $K_i \asymp K_4$ for $i = 1, 2, 3$):

(4) X has sub-exponential moment generating function: for some $K_4 \in (0, \infty)$,

$$\mathbb{E}\exp(\lambda X) \le \exp(K_4^2\lambda^2) \qquad \forall \lambda \in [-K_4^{-1}, K_4^{-1}]. \tag{2.33}$$

Proof. Exercise. □

3. Jan 23: Concentration from the martingale method

- We state and prove the Azuma–Hoeffding inequality
- We state and prove an important consequence – McDiarmid's inequality – showing concentration for general functions on product probability spaces that are Lipschitz with respect to the Hamming metric.
- We see a few applications.

3.1. The Azuma–Hoeffding inequality.

Many random variables of interest cannot be decomposed as sums of independent random variables as in Theorem 2.1. However, a wide range of applications is opened up with the realization that the proof of Theorem 2.1 applies with minor modification to martingale sequences.

The basic idea is as follows: Suppose we have a random variable $Z = F(X_1, \dots, X_n)$ that is a function of a large number n of random variables (not necessarily independent). *A priori* our best guess of the value of Z is $\mathbb{E}Z$. We then consider "revealing" (or "exposing") the values of each X_i in turn. At stage i our best guess of Z is given by the conditional expectation $\mathbb{E}(F(X_1, \dots, X_n)|X_1, \dots, X_i))$. If we can show that at each step, the outcome for X_i cannot affect the conditional expectation very much, it will follow from Theorem 3.1 below that Z is concentrated. One should keep in mind the special case $Z = S_n = X_1 + \cdots X_n$ of the sum of independent centered bounded variables, which was already covered by Theorem 2.1 – in this case, X_i can only affect the full sum by $\|X_i\|_{L^\infty} \leq |b_i - a_i|$.

We formalize the general idea of "revealing" bits of information one at a time using a (finite) filtration of the probability space.[5] Let $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}$ be a finite filtration of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Recall that a sequence $Y_0, Y_1, \dots, Y_n \in L^1(\Omega, \mathcal{F})$ is a *martingale* with respect to the filtration if $Y_i \in L^1(\Omega, \mathcal{F}_i)$ and $\mathbb{E}(Y_i|\mathcal{F}_{i-1}) = Y_{i-1}$ for each $1 \leq i \leq n$. (Thus $Y_0 = \mathbb{E}Y_i$ for all $1 \leq i \leq n$.) The sequences of differences $X_i := Y_i - Y_{i-1}$ is called a *martingale difference sequence*. Note that $\mathbb{E}(X_i|\mathcal{F}_{i-1}) = 0$ (in particular $\mathbb{E}X_i = 0$) for all $1 \leq i \leq n$.

A special case is if $Y_n - \mathbb{E}Y_n$ is the sum S_n of independent centered random variables $X_1, \dots, X_n$, where the filtration is the one generated by the sequence: $\mathcal{F}_i := \sigma(X_1, \dots, X_i)$ (if you haven't seen martingales before then it would be instructive to verify this).

Theorem 3.1 (Azuma–Hoeffding inequality)**.** *Let $X_1, \dots, X_n$ be a martingale difference sequence on a filtered probability space as above, and assume $|X_i| \leq b_i$ a.s. for all $i \in [n]$. Then with $B := (\sum_{i=1}^n b_i^2)^{1/2}$, we have that $Y_n - \mathbb{E}Y_n$ is $O(B)$-sub-Gaussian.*

Proof. The proof largely follows that of Theorem 2.1. We may assume $\mathbb{E}S_n = 0$. To bound the moment generating function we now write

$$\mathbb{E}\exp(\lambda S_n) = \mathbb{E}\exp(\lambda S_{n-1} + \lambda X_n) = \mathbb{E}\exp(\lambda S_{n-1})\mathbb{E}(\exp(\lambda X_n)|\mathcal{F}_{n-1}).$$

Now we can bound the inner expectation

$$\mathbb{E}(\exp(\lambda X_n)|\mathcal{F}_{n-1}) = \exp(O(\lambda^2 b_n^2)) \tag{3.1}$$

in the same way we established (2.4), and hence

$$\mathbb{E}\exp(\lambda S_n) = \exp(O(\lambda^2 b_n^2))\mathbb{E}\exp(\lambda S_{n-1}).$$

[5] A standard graduate course on probability mostly focuses on the case of infinite filtrations $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$ and convergence properties of martingales, but such questions are largely irrelevant for our aims of getting quantiative, non-asymptotic bounds.

Iterating, we obtain

$$\mathbb{E}\exp(\lambda S_n) = \exp(O(\lambda^2 B^2))$$

and the proof concludes as in the proof of Theorem 2.1. □

3.2. **Concentration for Hamming-Lipschitz functions.** The general concentration of measure (CoM) phenomenon is often stated informally as follows:

> *Lipschitz functions of a large number of independent (or approximately independent) random variables are typically very close to their expectation.* (CoM)

This phenomenon holds for a wide variety of metric spaces that are "high-dimensional" in some sense, such as high-dimensional spheres $\mathbb{S}^{d-1}$ with the geodesic distance (or more generally, positively curved manifolds – see the appendix of [MS86]), and high-dimensional product spaces. For the case of product spaces, the choice of metric is important. In this section we use Azuma's inequality to give a version of (CoM) for product spaces equipped with the *Hamming metric* (or more generally a weighted Hamming metric) known as McDiarmid's inequality. Next lecture we'll see another powerful version of (CoM) on product spaces with the Euclidean metric.

We'll deduce McDiarmid's inequality from Theorem 3.1. For random X in a metric space $(\mathcal{X}, d)$ and a function $F : \mathcal{X} \to \mathbb{R}$, in order to show $F(X)$ is close to its expectation with high probability, one aims to find a filtration under which the *Doob martingale* $Y_i := \mathbb{E}(F(X)|\mathcal{F}_i)$ has bounded differences, and then from Theorem 3.1 it follows that $Y_n - \mathbb{E}Y_n = F(X) - \mathbb{E}F(X)$ is $O(B)$-sub-Gaussian.

We put this in a general framework. Given metric spaces $(\mathcal{X}_1, d_1), \dots, (\mathcal{X}_n, d_n)$, we endow the product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with the *Hamming metric*

$$d_H(x, y) := \sum_{i=1}^{n} 1_{x_i \neq y_i} \tag{3.2}$$

or more generally, a *weighted Hamming metric*

$$d_{H,b}(x, y) := \sum_{i=1}^{n} b_i 1_{x_i \neq y_i} \tag{3.3}$$

for given positive weights $b_1, \dots, b_n$. (So far we are ignoring the metrics d_i on each factor, but we include these so that we can talk about random elements $X_i \in \mathcal{X}_i$, i.e. measurable functions from $(\Omega, \mathcal{F})$ to $\mathcal{X}_i$ equipped with the Borel σ-algebra induced by the metric d_i. In many of the cases we'll consider $\mathcal{X}_i$ will be finite sets and the metrics d_i are unimportant.)

Note that a function $F : \mathcal{X} \to \mathbb{R}$ is 1-Lipschitz with respect to $d_{H,b}$ if $F(x)$ can change by at most b_i when all but the ith coordinate of x is held fixed and x_i is allowed to vary. As a consequence of Theorem 3.1 we have the following:

Theorem 3.2 (McDiarmid's inequality)**.** *In the above setup, let $F : \mathcal{X} \to \mathbb{R}$ be 1-Lipschitz with respect to $d_{H,b}$ for some weight vector $b = (b_1, \dots, b_n)$, and let $X = (X_1, \dots, X_n) \in \mathcal{X}$ be a random element with independent components. Then with $B := (\sum_{i=1}^n b_i^2)^{1/2}$, we have that $F(X) - \mathbb{E}F(X)$ is $O(B)$-sub-Gaussian.*

Proof. Taking the filtration $\mathcal{F}_i := \sigma(X_1, \dots, X_i)$ given by successively exposing each coordinate of X, the claim will follow from Theorem 3.1 once we show that the sequence $Y_i := \mathbb{E}(F(X)|\mathcal{F}_i)$ satisfies the bounded differences assumption $|Y_i - Y_{i-1}| \leq b_i$ a.s. To

show this, we introduce an independent copy $X' = (X'_1, \dots, X'_n)$ of the vector X, and note (using the shorthand $X_{<i} := (X_1, \dots, X_{i-1})$, etc.)

$$\begin{aligned} |Y_i - Y_{i-1}| &= |\mathbb{E}(F(X)|X_{\leq i}) - \mathbb{E}(F(X)|X_{<i})| \\ &= |\mathbb{E}[F(X_{<i}, X_i, X_{>i}) - F(X_{<i}, X'_i, X_{>i})|X_{\leq i}]| \\ &\leq \mathbb{E}[|F(X_{<i}, X_i, X_{>i}) - F(X_{<i}, X'_i, X_{>i})||X_{\leq i}] \\ &\leq b_i \end{aligned}$$

as desired. □

Example 3.3 (Random walk in a Banach space)**.** Let $f_1, \dots, f_n$ be elements of a normed space $(V, \|\cdot\|)$, let $X_1, \dots, X_n$ be iid Rademacher variables, and set $Z := \|X_1 f_1 + \cdots + X_n f_n\|$. We claim

$$\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2\exp(-ct^2/B^2) \qquad \forall t \geq 0 \tag{3.4}$$

with $B := (\sum_{i=1}^n \|f_i\|^2)^{1/2}$. Indeed, with $F : \{-1, 1\}^n \to \mathbb{R}$ given by $F(x_1, \dots, x_n) = \|x_1 f_1 + \cdots + x_n f_n\|$, we have $Z = F(X)$ with $X = (X_1, \dots, X_n)$. From the triangle inequality, if x, y differ only on coordinate i, then

$$|F(x) - F(y)| \leq \|(x_i - y_i)f_i\| = 2\|f_i\|$$

so F is 1-Lipschitz under the weighted Hamming metric $d_{H,b}$ with $b_i = 2\|f_i\|$. The claim now follows from Theorem 3.2. ◊

Example 3.4 (Longest common subsequence)**.** For two elements $x, y \in \{0, 1\}^n$ let $F(x, y)$ be the length of the longest common subsequence of the two vectors. That is, $F(x, y)$ is the largest k for which there are increasing sequences $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$ such that

$$x_{i_1} = y_{i_1}, \dots, x_{i_k} = y_{i_k}.$$

(The problem of finding long common subsequences of long sequences of bits, or more generally of letters from other finite alphabets such as $\{A, C, G, T\}$, is of relevance for the problem of matching samples of DNA sequences.)

Let $X_1, \dots, X_n, Y_1, \dots, Y_n$ be independent Bernoulli variables (possibly with different expectations. We claim that $F(X, Y) - \mathbb{E}F(X, Y)$ is $O(\sqrt{n})$-sub-Gaussian. Indeed, suppose $F(x, y) = \ell$, and note that modifying a single coordinate of x or y can only decrease the length of the longest common subsequence by at most 1 – if the modified coordinate was part of the optimizing subsequence, simply deleting it gives a common subsequence of length $\ell - 1$, which is a lower bound for the longest common subsequence of the new sequences. Thus, given two pairs $(x, y), (x', y') \in \{0, 1\}^{2n}$ that differ on a single coordinate, we have $F(x', y') \geq F(x, y) - 1$. Applying this bound with (x, y) and (x', y') reversed shows that F is 1-Lipschitz under the Hamming metric d_H on the hypercube $\{0, 1\}^{2n}$, and the claim follows from Theorem 3.2. ◊

Remark 3.5. We showed that the longest common subsequence for length-n random binary inputs is concentrates at scale at most $O(\sqrt{n})$ around its expectation $\mathbb{E}F(X, Y)$. However, we haven't said anything about the order of magnitude of the expectation. A sub-additivity argument shows that for iid Bernoulli($\frac{1}{2}$) inputs, $\frac{1}{n}\mathbb{E}F(X, Y)$ converges to a constant γ_2 (called the Chvátal–Sankoff constant) as $n \to \infty$, but the value of this constant is unknown as of this writing. (One similarly has existence of a constant γ_k for inputs drawn uniformly from an alphabet of size k.) The current best results due to G. S. Lueker put γ_2 in the range $(0.788071, 0.826280)$ [**?**].

3.3. **The chromatic number of Erdős–Rényi graphs.** Recall that the *chromatic number* $\chi(G)$ of a graph G is the minimal number of colors needed to color the vertices of G so that no edge has the same color at both endpoints (such a coloring is called a *proper* coloring). Alternatively, $\chi(G)$ is the minimal number of parts in a partition of the vertices of G into independent sets.

Let $n \in \mathbb{N}$ and $p \in (0,1)$, and let $G_{n,p}$ be an Erdős–Rényi random graph on labeled vertices $1, \ldots, n$ – that is, each pair $\{i,j\} \subset [n]$ is included as an edge of $G_{n,p}$ independently with probability p. Let $\{A_{ij}\}_{1 \le i < j \le n}$ be the iid Bernoulli(p) indicator variables for the events $\mathcal{E}_{ij} = \{\{i,j\} \text{ is an edge in } G_{n,p}\}$. Let $Z := \chi(G_{n,p})$.

Theorem 3.6 (Shamir and Spencer '87)**.** *$Z - \mathbb{E}Z$ is $O(\sqrt{n})$-sub-Gaussian (specifically, it is $\sqrt{2(n-1)}$-sub-Gaussian).*

(The sharp bound $\sqrt{2(n-1)}$ follows from the same argument as below and applying Theorem 3.2 with the sharp value for c there.)

Let us first describe the proof at an informal level. We consider the filtration $(\mathcal{F}_i)_{i=0}^n$ that reveals the induced subgraph on vertices $1, \ldots, i$ for each $1 \le i \le n$. Thus, at step i we are told which vertices in $[i-1]$ are neighbors in i. Since we can always introduce a new color for vertex i, this new information at step i can only affect the conditional expectation of Z by at most 1. Thus, the result will follow from Theorem 3.1, taking all weights b_i equal to 1.

We can argue more formally using Theorem 3.2 as follows.

Proof. Since the $\binom{n}{2}$ potential edges are included independently, $G_{n,p}$ is a random sample from a product probability space. That is, we associate the set of graphs over $[n]$ with the discrete cube $\mathcal{G}_n = \{0,1\}^{\binom{[n]}{2}}$, where $g = (g_{ij})_{1 \le i < j \le n} \in \mathcal{G}_n$ is associated to the graph that includes edge $\{i,j\}$ whenever $g_{ij} = 1$. For each $2 \le k \le n$ let $E_k = \{\{i,k\} : 1 \le i \le k-1\}$ be the set of potential edges joining k to an earlier vertex in the ordering. $E_2, \ldots, E_n$ is a partition of the set $\binom{[n]}{2}$ of potential edges. This gives a product space decomposition $\mathcal{G}_n = \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$, where $\mathcal{X}_k = \{0,1\}^{E_k}$. For each $1 \le k \le n$ let $X_k = (A_{ik})_{i=1}^{k-1} \in \mathcal{X}_k$ be the random Bernoulli vector determining the neighbors of vertex k in $[k-1]$ in the random graph $G_{n,p}$. An element $(x_2, \ldots, x_n) \in \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ determines a graph $g \in \mathcal{G}_n$, and we define $F(x_2, \ldots, x_n)$ to be $\chi(g)$. We have thus represented $\chi(G_{n,p})$ as a function $F(X_2, \cdots, X_n)$ of independent random elements of the respective cubes $\mathcal{X}_2, \ldots, \mathcal{X}_n$.

The claim that $Z - \mathbb{E}Z$ is $O(\sqrt{n})$-sub-Gaussian will follow from Theorem 3.2 as soon as we can show that F is $O(1)$-Lipschitz under the Hamming metric d_H on $\mathcal{X}_2 \times \cdots \times \mathcal{X}_n$. Consider arbitrary $x, y \in \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ such that x and y differ on a single factor $\mathcal{X}_k$ for some $2 \le k \le n$. Thus, the associated graphs only differ on the set of edges joining k to $[k-1]$. We claim that

$$F(y) \le F(x) + 1. \tag{3.5}$$

By applying the same bound with x and y reversed it will then follow that F is 1-Lipschitz. To see why (3.5) is true, we consider a fixed coloring of the vertices that is proper for the graph associated to x and uses the minimal number $F(x)$ of colors. From this coloring, we can get a proper coloring for the graph associated to y by simply assigning a new color to vertex k, thus using $F(x)+1$ colors. The minimal number of colors needed is thus at most $F(x)+1$, and (3.5) follows, completing the proof of Theorem 3.6. □

4. Jan 25: Concentration from isoperimetry

Last time we saw how the martingale method can be used to formalize the general principle (CoM) for functions on product spaces that are Lipschitz under the Hamming metric. Now we consider functions that are Lipschitz under the Euclidean metric (or a geodesic distance derived from an ambient Euclidean metric). The basic approach exploits isoperimetric theorems (or weak versions of isoperimetric theorems). Later in the course we'll see an alternative approach through log-Sobolev inequalities.

Whereas the martingale method led naturally to bounds for the deviation of random variables from their *expectation*, we'll see that the isoperimetric approach instead leads naturally to bounds on deviations from the *median*. The difference will be unimportant for applications we'll consider, as shown in the following.

Exercise 4.1 (Equivalence of concentration about the mean and median)**.** Recall that $m \in \mathbb{R}$ is a median for a real-valued random variable if

$$\mathbb{P}(X \leq m) \geq 1/2 \quad \text{and} \quad \mathbb{P}(X \geq m) \geq 1/2.$$

(a) Show that any real random variable X has at least one median, and that the set of all medians of X is a closed interval. Give an example of a random variable having more than one median value.

(b) Let m be any median of X, and suppose there are $a \in \mathbb{R}$ and $K > 0$ such that

$$\mathbb{P}(|X - a| \geq t) \leq 2\exp(-t^2/K^2) \quad \forall t \geq 0. \tag{4.1}$$

Show that $|a - m| = O(K)$, and deduce that

$$\mathbb{P}(|X - m| \geq t) \leq 2\exp(-ct^2/K^2) \quad \forall t \geq 0 \tag{4.2}$$

for some universal constant $c > 0$. Deduce from this that $|m - \mathbb{E}X| = O(K)$, and that (4.2) holds with m replaced by $\mathbb{E}X$, for a possibly smaller universal constant $c > 0$.

(Hint: Note that the bound (4.2) holds trivially for $t \leq K\sqrt{(\log 2)/c}$, so by shrinking c we may assume without loss of generality that $t \geq CK$ for any fixed constant $C > 0$ as large as we please.)

(c) Show that if (4.1) holds then $\mathrm{Var}(X) = O(K^2)$. Deduce that if $X \geq 0$ almost surely, then (4.2) holds with m replaced by $(\mathbb{E}X^2)^{1/2}$, for a possibly smaller universal constant $c > 0$.

4.1. From isoperimetry to concentration.

Let $(\mathcal{X}, d, \mu)$ be a metric measure space, that is, a metric space $(\mathcal{X}, d)$ equipped with a Borel σ-algebra $\mathcal{B}$ generated by the topology induced by d, and a (not necessarily probability) measure $\mu : \mathcal{B} \to [0, +\infty]$. For $A \in \mathcal{B}$, we can define the t-blowups

$$A_t := \{x \in \mathcal{X} : d(x, A) \leq t\} \tag{4.3}$$

and the boundary measure

$$\mu^+(A) := \liminf_{t \to 0} \frac{1}{t}\mu(A_t \setminus A). \tag{4.4}$$

A basic problem is to determine, for give $a > 0$, the minimizers of the *isoperimetric ratio*

$$\mu^+(A)/\mu(A) \tag{4.5}$$

over all A of measure $\mu(A) = a$.

For the case of $\mathbb{R}^d$ equipped with the Euclidean metric and Lebesgue measure $\mu = \mathrm{vol}_d$, we denote $\mathrm{vol}_{d-1}(\partial A) := \mu^+(A)$. We have the following classical result:

Theorem 4.1 (Isoperimetric theorem)**.** *Let $A \subset \mathbb{R}^d$ be nonempty and Lebesgue measurable and let B be a Euclidean ball of the same volume. Then for any $t \geq 0$,*

$$\mathrm{vol}_d(A_t) \geq \mathrm{vol}_d(B_t) \tag{4.6}$$

and hence (if A has smooth boundary, say),

$$\mathrm{vol}_{d-1}(\partial A) \geq \mathrm{vol}_{d-1}(\partial B). \tag{4.7}$$

We give the proof of Theorem 4.1 in Section 4.2, using the Brunn–Minkowski inequality.

Another setting in which the isoperimetric problem has been solved is the sphere $\mathbb{S}^{d-1}$ (with the Euclidean geodesic distance), where the extremizing sets are *spherical caps*, i.e. sets of the form

$$C_{v,b} := \{u \in \mathbb{S}^{d-1} : \langle u, v\rangle \geq b\} \tag{4.8}$$

for some $v \in \mathbb{S}^{d-1}$ and $b \in [-1, 1]$.

Theorem 4.2 (Lévy's isoperimetric theorem)**.** *On $\mathbb{S}^{d-1}$, let d_g be the Euclidean geodesic distance and let σ be the uniform surface measure (normalized to be a probability measure). For any $a > 0$ and $A \subset \mathbb{S}^{d-1}$ with $\sigma(A) = a$, we have*

$$\sigma(A_t) \geq \sigma(C_t) \qquad \forall t \geq 0 \tag{4.9}$$

where C is a spherical cap of measure $\sigma(C) = a$.

(An inequality of the form (4.7) follows from this but will not be needed.)

Exercise 4.2. Use Theorem 4.2 to prove Theorem 4.1.

We do not give the proof of Theorem 4.2 in these notes, but show how it combines with the computation of the measure of spherical caps from Proposition 1.2 to imply a general concentration of measure phenomenon.

Corollary 4.3 (Concentration on the sphere, blowup form)**.** *For any $A \subset \mathbb{S}^{d-1}$ with $\sigma(A) \geq \frac{1}{2}$, we have*

$$\sigma(A_t^c) \leq 2\exp(-ct^2 d) \qquad \forall t \geq 0 \tag{4.10}$$

(writing $A^c := \mathbb{S}^{d-1} \setminus A$ for $A \subset \mathbb{S}^{d-1}$).

(One can remove the factor 2 in (4.10) using the proof of Proposition 1.2, but this is not important for us.)

Proof. Fixing A and $t \geq 0$, by Theorem 4.2,

$$\sigma(A_t^c) = 1 - \sigma(A_t) \leq 1 - \sigma(C_t) = \sigma(C_t^c)$$

where C is a spherical cap of measure $\sigma(C) = \sigma(A) \geq \frac{1}{2}$. From Proposition 1.2, $\sigma(C_t^c) \leq 2\exp(-ct^2 d)$ for all $t \geq 0$, and the claim follows. □

From this we also deduce a general concentration of measure result of the type (CoM).

Corollary 4.4 (Concentration on the sphere, functional form)**.** *Let $f : \mathbb{S}^{d-1} \to \mathbb{R}$ be 1-Lipschitz under the Euclidean geodesic distance d_g, and let $U \in \mathbb{S}^{d-1}$ have distribution σ. Then*

$$\mathbb{P}(|f(U) - m_f| \geq t) \leq 2\exp(-ct^2 d) \qquad \forall t \geq 0$$

for a universal constant $c > 0$, where m_f is the (unique) median of $f(U)$.

Proof. Uniqueness of the median in this case is left as an exercise. Taking $A = \{f \leq m_f\}$, we have by the Lipschitz property that $\{f \geq m_f + t\} \subseteq A_t^c$, so by Corollary 4.3,

$$\mathbb{P}(f(U) \geq m_f + t) \leq \mathbb{P}(U \in A_t^c) \leq \exp(-ct^2 d).$$

We similarly obtain the same bound for the lower tail by taking $A = \{f \geq m_f\}$, and the claim follows. □

In fact the blowup and functional formulations of the concentration of measure phenomenon are equivalent.

Exercise 4.3. Show that Corollary 4.4 implies Corollary 4.3.

One notes that the arguments to deduce Corollaries 4.3 and 4.4 from Theorem 4.2 apply quite generally, showing how to deduce concentration of measure from an isoperimetric theorem together with a bound for the measure of extremizing sets.

Another important case where the isoperimetric theorem has been solved is for $\mathbb{R}^n$ equipped with the standard Gaussian measure γ_n rather than the Lebesgue measure.

Exercise 4.4 (Gaussian concentration from isoperimetry)**.** The isoperimetric theorem for n-dimensional Gauss space (i.e. $\mathbb{R}^n$ equipped with the Euclidean distance $d_2(x, y) = \|x - y\|_2$ and the standard Gaussian measure γ_n) states that for any $m \in (0, 1)$ and $r > 0$, among all Borel sets $A \subset \mathbb{R}^n$ of measure $\gamma_n(A) = m$, the ones that minimize $\gamma_n(A_r)$ are half-spaces, i.e. sets of the form

$$H_{u,a} = \{x \in \mathbb{R}^n : \langle x, u\rangle \leq a\}$$

for some $u \in S^{n-1}$ and $a \in \mathbb{R}$. Use this fact to show that for any 1-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, $f(G) - \mathbb{E}f(G)$ is K-sub-Gaussian with $K = O(1)$, where $G \sim N(0, I_n)$ is a standard Gaussian vector. (You may find a result from Exercise 4.1 helpful for this.) *Bonus: find the optimal value of K.*

Corollary 4.4 and Exercise 4.4 give two important examples where concentration of measure for Lipschitz functions can be deduced from an isoperimetric theorem. In fact, isoperimetric theorems are only known in a few cases (another is the discrete hypercube with the uniform measure and Hamming metric, leading to another proof of Theorem 3.2). Concentration of measure is generally easier to establish than an isoperimetric principle (particularly if you don't care about the optimal constant c). In particular, in Section 4.2 we give a proof of Corollary 4.4 using the Brunn–Minkowsky inequality for Lebesgue measure that bypasses the more difficult Theorem 4.2.

4.2. **(Optional) Proof of the classical isoperimetric theorem.** Here we write $\lambda = \lambda_d$ for Lebesgue measure on $\mathbb{R}^d$. The Brunn–Minkowski theorem (BMI = BMI(d)) says

$$\lambda(A + B)^{1/d} \geq \lambda(A)^{1/d} + \lambda(B)^{1/d} \qquad \forall\, A, B \tag{4.11}$$

(where A, B are assumed to be measurable and non-empty here and in the sequel). This is often stated in the equivalent form

$$\lambda(\theta A + (1 - \theta)B) \geq \lambda(A)^\theta \lambda(B)^{1-\theta} \qquad \forall \theta \in [0, 1] \quad \forall\, A, B. \tag{4.12}$$

Claim 4.5. (4.11) ⇔ (4.12).

Proof. ($\Rightarrow$) Let $\theta \in [0,1]$. From (4.11) and the weighted AM-GM inequality (or Jensen after taking logs) we have

$$\begin{aligned}\lambda(\theta A + (1-\theta)B)^{1/d} &\geq \lambda(\theta A)^{1/d} + \lambda((1-\theta)B)^{1/d} \\ &= \theta\lambda(A)^{1/d} + (1-\theta)\lambda(B)^{1/d} \\ &\geq \lambda(A)^{\theta/d}\lambda(B)^{(1-\theta)/d}\end{aligned}$$

□

($\Leftarrow$) Fix sets A, B. Applying (4.12) with dilates A/θ, $B/(1-\theta)$ in place of A, B, we have

$$\lambda(A+B) \geq \lambda(A/\theta)^{\theta}\lambda(B/(1-\theta))^{1-\theta}$$

for all $\theta \in [0,1]$. Taking dth roots on both sides,

$$\lambda(A+B)^{1/d} \geq (\lambda(A)^{1/d}/\theta)^{\theta}(\lambda(B)^{1/d}/(1-\theta))^{1-\theta}.$$

Optimizing θ on the right hand side yields the claim. □

...

5. Jan 30: Talagrand's inequality

References:

- Talagrand's original papers [Tal96, Tal95] (the shorter review [Tal96] is a nice introduction).
- These notes
- Alon & Spencer [AS16]
- [Tao12, §2.1]

5.1. **Dimension-free concentration for product measures?** From Exercise 4.4 we see that 1-Lipschitz functions (under the Euclidean metric) of standard Gaussian vectors in $\mathbb{R}^n$ enjoy sub-Gaussian concentration about their means (or, equivalently, their medians) of width $O(1)$. Thus, Gaussian vectors enjoy *dimension-free* concentration under the Euclidean metric.

It's worth noting how this can improve on McDiarmid's inequality, which is only sensitive to the coordinate-wise Lipschitz behavior of functions. Let $f : \mathbb{R}^n \to \mathbb{R}$ be 1-Lipschitz under $\|\cdot\|_2$, and let $h : \mathbb{R}^n \to \mathbb{R}^n$ be the mapping that applies the arctangent function entrywise, thus $h(x) = (\arctan(x_i))_{i=1}^n$. Then $F = f \circ h$ is $\|\cdot\|_2$-Lipschitz. Moreover, because of the composition with h we have that F is $O(1)$-Hamming Lipschitz, so McDiarmid's inequality implies that for $G \sim N(0, I_n)$, $F(G) - \mathbb{E}F(G)$ is $O(\sqrt{n})$-sub-Gaussian, which is far worse that than the scale $O(1)$ implied by Exercise 4.4. (We only applied h in order to get a vector with almost-surely bounded entries, a minor technical point since Gaussians have very light tails.) In fact, for the example $f(x) = \|x\|_2$ the result of McDiarmid's inequality is *trivial*, as $F(G) = O(\sqrt{n})$ a.s. in this case. Thus, the Euclidean 1-Lipschitz property is an important strengthening of the Hamming-Lipschitz property.

It is then natural to ask the following:

Question 5.1. *Does dimension-free concentration for Euclidean Lipschitz functions extend to general product measures (apart form the Gaussian)?*

The answer to this turns out to be "no", as shown by the following:

Example 5.2. Let $A = \{x \in \{0,1\}^n : \sum x_i \leq n/2\}$ and let $F : \mathbb{R}^n \to \mathbb{R}$ be the function $F(x) = d_2(x, A)$ (where we write $d_2(x, y) = \|x - y\|_2$ for the Euclidean metric). Then F is 1-Lipschitz under d_2. Letting $X \in \{0,1\}^n$ be uniform random, we claim that $F(X)$ does not concentrate at scale $O(1)$, or in fact at any scale $o(n^{1/4})$. First, note that 0 is a median of $F(X)$ since A contains at least half of the discrete cube. On the other hand, for small $\delta > 0$ let

$$B_\delta = \{x \in \{0,1\}^n : \sum_i x_i \geq n/2 + \delta\sqrt{n}\}.$$

We claim $B_\delta \subseteq A^c_{\delta^{1/2} n^{1/4}}$, where $A_t = \{x \in \mathbb{R}^n : d_2(x, A) \leq t\}$ is the t-blowup of A under d_2. Indeed, for any $x \in B_\delta$ and $y \in A$, we have

$$\delta\sqrt{n} \leq \sum_i x_i - n/2 \leq \sum_i x_i - y_i \leq \sum_i (x_i - y_i)^2 \tag{5.1}$$

where in the last bound we crucially used that both x and y lie in the discrete cube. Thus $\|x - y\|_2 \geq \delta^{1/2} n^{1/4}$ as claimed. We hence have

$$\mathbb{P}(F(X) \geq \delta^{1/2} n^{1/4}) \geq \mathbb{P}(X \in B_\delta).$$

But since $S_n = \sum_{i=1}^n X_i$ has mean $n/2$ and standard deviation $\gtrsim \sqrt{n}$, one can easily show that $\mathbb{P}(X \in B_\delta) \geq \frac{1}{10}$, say, if δ is a sufficiently small universal constant. Thus we have

$$\mathbb{P}(F(X) \leq 0) \geq \frac{1}{2} \qquad \text{and} \qquad \mathbb{P}(F(X) \geq c n^{1/4}) \geq \frac{1}{10}$$

for a sufficiently small constant $c > 0$, so $F(X)$ does not concentrate at a scale smaller than $cn^{1/4}$. ◊

A key feature of the function F in the above example is the integrality gap between elements of the subset A of the discrete cube, which was applied in the last inequality in (5.1). As the next result shows, if instead of $d_2(x, A)$ we had taken $F(x)$ to be the distance from x to the *convex hull* $\widetilde{A}$ of A, then $F(X)$ would enjoy dimension-free concentration (and one checks that the last bound in (5.1) would fail for this example, for general $y \in \widetilde{A}$).

Theorem 5.3 (Talagrand's inequality, convex functional form)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}$ be 1-Lipschitz under $\|\cdot\|_2$ and convex, and let $X \in [-1,1]^n$ have independent components. Then for any median m_F of $F(X)$, we have that $F(X) - m_F$ is $O(1)$-sub-Gaussian. In fact,*

$$\mathbb{P}(|F(X) - m_F| \geq t) \leq 4\exp(-t^2/16) \qquad \forall t \geq 0. \tag{5.2}$$

Remark 5.4. In fact, as the proof shows, we may relax the convexity assumption to assume only that the sub-level sets $\{F \leq a\}$ are convex for all $a \in \mathbb{R}$ (this is sometimes called *quasi-convexity*).

Remark 5.5. From Exercise 4.1 we get that $F(X) - \mathbb{E}F(X)$ is $O(1)$-sub-Gaussian, and if F is non-negative then $F(X) - (\mathbb{E}F(X)^2)^{1/2}$ is $O(1)$-sub-Gaussian.

Exercise 5.1. Deduce from Theorem 5.3 the more general statement that if F is convex and L-Lipschitz, and $X \in \mathbb{R}^n$ has independent entries bounded in absolute value by K a.s., then $F(X)$ is $O(KL)$-sub-Gaussian.

Here are some examples of functions to which Theorem 5.3 applies.

Example 5.6 (Distance to a subspace)**.** If $A \subset \mathbb{R}^n$ is convex, then $F(x) := d_2(x, A)$ is convex and 1-Lipschitz under d_2, and Theorem 5.3 implies that $d_2(X, A)$ enjoys $O(1)$-sub-Gaussian concentration for any random vector X with independent components of size $O(1)$ a.s.

An important special case in the study of random matrices is that A is a fixed d-dimensional subspace $V \subset \mathbb{R}^n$. A computation shows $\mathbb{E}d_2(X, V)^2 = n - d$ if X has standardized entries (centered and unit variance), and Theorem 5.3, together with a result from Exercise 4.1, shows that $d_2(X, V) - \sqrt{n-d}$ is $O(1)$-sub-Gaussian. Here again is a situation where McDiarmid's inequality only provides concentration at the trivial scale $O(\sqrt{n})$. ♢

Example 5.7 (Norm of a random matrix)**.** Identifying the space of $n \times n$ matrices with real entries with $\mathbb{R}^{n\times n}$, the Euclidean norm is given by $\|A\|_2 = (\sum_{i,j=1}^n A_{ij}^2)^{1/2} = \|A\|_F$, the Frobenius (or Hilbert–Schmidt) norm. Let $F(A) := \|A\|_{\mathrm{op}} = \sup_{u\in\mathbb{S}^{n-1}} \|Au\|_2$ be the $\ell^2 \to \ell^2$ operator norm of A. Since

$$\|A\|_F^2 = \sum_{i=1}^n \sigma_i(A)^2 \geq \sigma_1(A)^2 = F(A)^2$$

where $\sigma_1(A) \geq \cdots \geq \sigma_n(A) \geq 0$ are the singular values of A, it follows that F is 1-Lipschitz under the Euclidean metric. Since F is a norm, we have by the triangle inequality and homogeneity that

$$F(\theta A + (1-\theta)B) \leq \theta F(A) + (1-\theta)F(B) \qquad \forall A, B \in \mathbb{R}^{n\times n}, \theta \in [0,1]$$

so F is convex. From Theorem 5.3 it follows that if $X = (X_{ij})$ is an $n \times n$ random matrix with independent entries $X_{ij} \in [-1, 1]$, then $\|X\|_{\mathrm{op}} - m_F$ is $O(1)$-sub-Gaussian for any median m_F of $\|X\|_{\mathrm{op}}$.

We can compare this with the typical order of $\|X\|_{\mathrm{op}}$. So far everything we've said applies to the matrix of all zeros, but if the entries have variances uniformly bounded below then it's not hard to show that any median of $\|X\|_{\mathrm{op}}$ is of size $\gtrsim \sqrt{n}$ (and in fact with a bit more work one has a matching upper bound $O(\sqrt{n})$). This is particularly easy for the random sign matrix with iid Rademacher entries (thus X is uniform in $\{-1, 1\}^{n\times n}$). Then $\|X\|_{\mathrm{op}} \geq \|Xe_1\|_2 = (\sum_{i=1}^n X_{i1}^2)^{1/2} = \sqrt{n}$ a.s., where $e_1 = (1, 0, \ldots, 0)$ is the first standard basis vector. ♢

Remark 5.8. While this dimension-free concentration for the norm of random matrices is a surprising and useful fact, it turns out that in fact $\|X\|_{\mathrm{op}}$ has fluctuations of order $n^{-1/6}$! Moreover, $n^{1/6}(\|X\|_{\mathrm{op}} - 2\sqrt{n})$ converges in distribution to a *Tracy–Widom distribution*, a measure which, like the Gaussian, arises for mysterious reasons in diverse contexts. Tracy–Widom universality aside, the concentration at scale smaller than what is implied by the "off-the-shelf" concentration result of Theorem 5.3 is an instance of what is known as the *superconcentration phenomenon* – see [Cha14] for more on this. (We may have time to explore this a bit later in the course.) Below we'll consider concentration for another example of a random variable enjoying superconcentration (as well as Tracy–Widom universality) – the longest increasing subsequence of iid samples $X_i \in [0, 1]$ – using a related concentration inequality of Talagrand of a more combinatorial nature.

As we'll see, Talagrand's inequalities (Theorem 5.3 and the combinatorial version we state below) are particularly effective for showing concentration for functions F involving an optimization problem over linear/sub-linear functionals. (Note that both of the preceding examples are of this type.)

5.2. **Another proof of the Johnson–Lindenstrauss Lemma.** As another quick application of Theorem 5.3, we give another proof of Theorem 1.5. (The following more general statement could have been proved by the same lines as our first proof of Theorem 1.5, just using the general Hoeffding inequality of Theorem 2.1 in place of Theorem 1.4.)

Theorem 5.9 (Johnson–Lindenstrauss lemma)**.** *Let $x_1, \dots, x_m$ be fixed (deterministic) points in $\mathbb{R}^N$. Let X be an $N \times d$ matrix with independent standardized real entries X_{ij} (that is, $\mathbb{E}X_{ij} = 0$ and $\mathbb{E}X_{ij}^2 = 1$ for all i, j) with $|X_{ij}| \leq B$ a.s. for all i, j and some finite B. For each $i \in [m]$ set*

$$y_i := \frac{1}{\sqrt{d}} X^{\mathsf{T}} x_i. \tag{5.3}$$

For any $\varepsilon \in (0,1)$, if

$$d \geq C\varepsilon^{-2} \log m \tag{5.4}$$

then

$$1 - \varepsilon \leq \frac{\|y_i - y_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \varepsilon \qquad \forall 1 \leq i < j \leq m \tag{5.5}$$

except with probability at most $\exp(-c\varepsilon^2 d)$.

Proof. We may assume without loss of generality that $\varepsilon \leq \frac{1}{2}$. For fixed $u \in \mathbb{S}^{N-1}$ consider the function $F_u : \mathbb{R}^{N \times d} \to \mathbb{R}$ given by

$$F_u(A) := \|A^{\mathsf{T}} u\|_2.$$

Then for any $A, B \in \mathbb{R}^{N \times d}$,

$$F_u(A) - F_u(B) \leq \|(A - B)^{\mathsf{T}} u\|_2 \leq \|A - B\|_{\mathrm{op}} \leq \|A - B\|_F$$

so F_u is 1-Lipschitz under the Euclidean Frobenius norm on $\mathbb{R}^{N \times d}$. One similarly sees from the triangle inequality and homogeneity of norms that F_u is convex. From Theorem 5.3 it follows that $F_u(X) - m_u$ is $O(B)$-sub-Gaussian for any fixed $u \in \mathbb{S}^{N-1}$, where m_u is any median of $F_u(X)$. From Exercise 4.1 we get that $F_u(X) - (\mathbb{E}F_u(X)^2)^{1/2}$ is $O(B)$-sub-Gaussian. We already computed in the proof of Theorem 1.5 that $\mathbb{E}F_u(X)^2 = d$, so $\|X^{\mathsf{T}} u\|_2 - \sqrt{d}$ is $O(B)$-sub-Gaussian. Applying this with $u = (x_i - x_j)/\|x_i - x_j\|_2$ for any $1 \leq i < j \leq m$ (we may assume without loss of generality that all m points are distinct) gives

$$\begin{aligned}
\mathbb{P}\left(\left| \frac{\|y_i - y_j\|_2^2}{\|x_i - x_j\|_2^2} - 1 \right| > \varepsilon \right) &\leq \mathbb{P}\left(\left| \frac{\|y_i - y_j\|_2}{\|x_i - x_j\|_2} - 1 \right| > c\varepsilon \right) \\
&= \mathbb{P}(|\|X^{\mathsf{T}} u\|_2 - \sqrt{d}| > c\varepsilon\sqrt{d}) \\
&\leq 2\exp(-c\varepsilon^2 d)
\end{aligned}$$

for a sufficiently small constant $c > 0$, where in the first bound we used that the mapping $t \mapsto t^2$ is $O(1)$-Lipschitz on $[\frac{1}{2}, \frac{3}{2}]$. The proof concludes by a union bound over all pairs $\{i, j\}$ just as in the proof of Theorem 1.5. □

6. Feb 01: Talagrand's inequality – proof and further applications

6.1. **Talagrand's inequality on the discrete cube.** The general form of Talagrand's inequality involves an interesting way of quantifying the distance between a point and a set that is somewhat hard to absorb at first. To motivate the idea we consider the special case

of product measures on the discrete cube $Q_n = \{0,1\}^n$, where the ideas are more transparent. (Recall that the counterexample of Example 5.2 was in this special setting.) We'll then explain how to generalize the definition and argument in the next subsection.

Theorem 6.1 (Talagrand's inequality – discrete cube case)**.** *Let $X = (X_1, \dots, X_n) \in Q_n$ have independent components. For any convex set $D \subset \mathbb{R}^n$ with $\mathbb{P}(X \in D) > 0$, we have*

$$\mathbb{E} \exp(d_2(X, D)^2/4) \leq \frac{1}{\mathbb{P}(X \in D)}. \tag{6.1}$$

Exercise 6.1. Use Theorem 6.1 to deduce the functional form Theorem 5.3 for the case that X is supported in $\{-1, 1\}^n$.

Proof of Theorem 6.1. First note that we can replace D with the convex hull of its intersection with the discrete cube. Indeed, with $A := D \cap Q_n$ and $\mathrm{conv}(A) \subset \mathbb{R}^n$ the convex hull of A, we have $\mathrm{conv}(A) \subseteq D$, so the left hand side in (6.1) can only increase when we replace D with $\mathrm{conv}(A)$, while the right hand side is unchanged. It thus suffices to show

$$\mathbb{E} \exp(d_2(X, \mathrm{conv}(A))^2/4) \leq \frac{1}{\mathbb{P}(X \in A)} \qquad \forall A \subseteq Q_n. \tag{6.2}$$

We proceed by induction on the dimension n. For the base case $n = 1$, the desired bound (6.2) reads

$$p + e^{1/4}(1 - p) \leq 1/p \tag{6.3}$$

where $p = \mathbb{P}(X \in A)$, and one easily verifies this inequality holds for any $p \in [0, 1]$.

Now letting $n \geq 2$, we aim to establish (6.2) assuming the statement holds with $n - 1$ in place of n. Fix $A \subseteq Q_n$. For a general point $x \in Q_n$ we'll write $x = (x', x_n) \in Q_{n-1} \times \{0, 1\}$. We define three subsets of Q_{n-1}:

$$A_b := \{x' \in Q_{n-1} : (x', b) \in A\}\,, \; b = 0, 1, \qquad B := A_0 \cup A_1. \tag{6.4}$$

Thus, A_0, A_1 are the two slices of A according to the value of the last coordinate, and B is the projection of A to Q_{n-1}. The key to the induction is the following claim controlling the distance from x to $\mathrm{conv}(A)$ in terms of the distance from x' to the convex hulls of A_0, A_1 and B.

Claim 6.2. For any $x = (x', x_n) \in Q_n$ and $\lambda \in [0, 1]$,

$$d_2(x, \mathrm{conv}(A))^2 \leq \lambda d_2(x', \mathrm{conv}(A_{x_n}))^2 + (1 - \lambda) d_2(x', \mathrm{conv}(B))^2 + (1 - \lambda)^2. \tag{6.5}$$

Assuming the claim for now, let $X = (X', X_n) \in Q_n$ have independent components. We write $\mathbb{E}' := \mathbb{E}_{X'}$ for expectation under the randomness of X' only (i.e. conditional on X_n). Talagrand notes that the key to the proof is to resist the temptation to optimize the bound (6.5) in λ at this point! Instead, we first exponentiate the inequality and average over X', to bound

$$\begin{aligned}
\mathbb{E}' \exp(d_2(X, \mathrm{conv}(A))^2/4) &\leq e^{(1-\lambda)^2/4} \mathbb{E}' \left[\left(e^{d_2(X', \mathrm{conv}(A_{X_n}))^2/4} \right)^{\lambda} \left(e^{d_2(X', \mathrm{conv}(B))^2/4} \right)^{1-\lambda} \right] \\
&\leq e^{(1-\lambda)^2/4} \left(\mathbb{E}' e^{d_2(X', \mathrm{conv}(A_{X_n}))^2/4} \right)^{\lambda} \left(\mathbb{E}' e^{d_2(X', \mathrm{conv}(B))^2/4} \right)^{1-\lambda} \\
&\leq e^{(1-\lambda)^2/4} \mathbb{P}'(X' \in A_{X_n})^{-\lambda} \mathbb{P}(X' \in B)^{\lambda - 1}
\end{aligned}$$

where in the second line we applied Hölder's inequality and in the third line we used the induction hypothesis. We can express the right hand side in the last line as

$$\mathbb{P}(X' \in B)^{-1} e^{(1-\lambda)^2/4} r^{-\lambda}$$

where $r := \mathbb{P}'(X' \in A_{X_n})/\mathbb{P}(X' \in B) \in [0,1]$. Now we optimize λ depending on r. With the choice

$$\lambda(r) := (1 + 2\log r) 1_{r \in [e^{-1/4}, 1]}$$

one can show (exercise) that

$$e^{(1-\lambda(r))^2/4} r^{-\lambda(r)} \leq 2 - r \qquad \forall r \in [0,1].$$

Substituting into the previous bound, we've shown

$$\mathbb{E}' \exp(d_2(X, \mathrm{conv}(A))^2/4) \leq \mathbb{P}(X' \in B)^{-1} (2 - \frac{\mathbb{P}'(X' \in A_{X_n})}{\mathbb{P}(X' \in B)}).$$

Writing $u := \mathbb{P}(X \in A)/\mathbb{P}(X' \in B) \in [0,1]$, upon averaging the above inequality over X_n we obtain

$$\mathbb{E} \exp(d_2(X, \mathrm{conv}(A))^2/4) \leq \mathbb{P}(X' \in B)^{-1}(2-u) = \mathbb{P}(X \in A)^{-1} u(2-u) \leq \mathbb{P}(X \in A)^{-1}$$

giving (6.2) to complete the proof of Theorem 6.1 given Claim 6.2. □

For the proof of Claim 6.2 we need an elementary lemma. For $A \subset Q_n$ we write $A_\uparrow$ for the upwards closure of A, i.e. the set of all vectors $\mathbf{1}_J = (1_{j \in J})_{j=1}^n$ such that J contains the support of some vector $y \in A$. Thus, $A_\uparrow$ is a monotone subset of Q_n, in the sense that

$$y \in A_\uparrow,\ z \in Q_n,\ y_i \leq z_i\ \forall i \qquad \Longrightarrow \qquad z \in A_\uparrow.$$

Lemma 6.3. *For any $A \subset Q_n$, we have $d_2(0, \mathrm{conv}(A)) = d_2(0, \mathrm{conv}(A_\uparrow))$.*

(The reason for passing to $A_\uparrow$ will become clear in the proof of Claim 6.2.)

Intuitively, the reason for Lemma 6.3 is that the extra points we've included in $A_\uparrow$ are on the "opposite side" of $\mathrm{conv}(A)$ from the origin.

Proof. It suffices to consider adding one point of the upward closure at a time: we claim that for any $E \subset Q_n$ and $z \in Q_n \setminus E$ such that $\mathrm{supp}(y) \subset \mathrm{supp}(z)$ for some $y \in E$, writing $E' := E \cup \{z\}$, we have

$$d_2(0, \mathrm{conv}(E)) = d_2(0, \mathrm{conv}(E')). \tag{6.6}$$

That the left hand side is at least as large as the right is obvious. For the reverse inequality, note we can express any $w \in \mathrm{conv}(E') \setminus \mathrm{conv}(E)$ as $w = \alpha x + (1-\alpha) z$ for $x \in E$ and $\alpha \in [0,1]$. Then note that the point $w_0 = \alpha x + (1-\alpha) y \in \mathrm{conv}(E)$ is closer to the origin, since all components of $w - w_0 = (1-\lambda)(z-y)$ are non-negative. □

Proof of Claim 6.2. We can apply an isometry of Q_n to assume $x = 0$. Specifically, viewing Q_n as the abelian group $(\mathbb{Z}/2\mathbb{Z})^n$, the map $y \mapsto \phi(y) = y - x$ is an isometry and takes x to 0. (Note that $\pm x \pm y$ are all equal to $(1_{x_i \neq y_i})_{i=1}^n$ in $(\mathbb{Z}/2\mathbb{Z})^n$!) So replacing A with $\phi^{-1}(A)$ we may assume $x = 0$.

With this reduction and from Lemma 6.3, to establish the claim it now suffices to show

$$d_2(0, \mathrm{conv}(A_\uparrow))^2 \leq \lambda d_2(0, \mathrm{conv}(A_0))^2 + (1-\lambda) d_2(0, \mathrm{conv}(B))^2 + (1-\lambda)^2. \tag{6.7}$$

To see this, first note that

$$v \in A_0 \ \Rightarrow (v, 0) \in A \subseteq A_\uparrow \tag{6.8}$$

and

$$w \in B \;\Rightarrow\; (w,1) \in A_\uparrow \tag{6.9}$$

(here is where we needed to pass to $A_\uparrow$, as we can't guarantee $(w,1)$ lies in A). Taking convex hulls, we deduce that

$$v \in \mathrm{conv}(A_0)\,, w \in \mathrm{conv}(B) \quad \Longrightarrow \quad (v,0),(w,1) \in \mathrm{conv}(A_\uparrow) \tag{6.10}$$

and hence

$$\lambda(v,0) + (1-\lambda)(w,1) \in \mathrm{conv}(A_\uparrow) \qquad \forall v \in \mathrm{conv}(A_0), w \in \mathrm{conv}(B), \lambda \in [0,1]. \tag{6.11}$$

Applying this with v, w of minimal ℓ^2-norm, we conclude

$$\begin{aligned} d_2(0,\mathrm{conv}(A_\uparrow))^2 &\le \|(\lambda v + (1-\lambda)w, 1-\lambda)\|_2^2 \\ &= \|\lambda v + (1-\lambda)w\|_2^2 + (1-\lambda)^2 \\ &\le \lambda\|v\|_2^2 + (1-\lambda)\|w\|_2^2 + (1-\lambda)^2 \\ &= \lambda d_2(0,\mathrm{conv}(A_0))^2 + (1-\lambda) d_2(0,\mathrm{conv}(B))^2 + (1-\lambda)^2 \end{aligned}$$

giving (6.7), where we used Pythagoras's theorem in the second line and convexity in the third line. □

6.2. **Generalizing to arbitrary product measures.** Now we generalize Theorem 6.1 to arbitrary product probability spaces. That is let, $(\mathcal{X}_i, \mu_i)$, $i \in [n]$ be probability spaces, and form the product space $(\mathcal{X}, \mu)$ with $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$, $\mu = \mu_1 \otimes \cdots \otimes \mu_n$.

Recall that Theorem 6.1 is equivalent to the statement (see (6.2))

$$\mathbb{E}\exp(d_2(X,\mathrm{conv}(A))^2/4) \le \frac{1}{\mathbb{P}(X \in A)} \qquad \forall A \subseteq Q_n \tag{6.12}$$

for any subset A of the discrete cube $Q_n = \{0,1\}^n$, where $X \in Q_n$ is any random vector with independent components. In the general product space setting the notions of convex hull and Euclidean distance don't make sense, so the expression $d_2(X,\mathrm{conv}(A))$ has no meaning. The following result replaces $d_2(X,\mathrm{conv}(A))$ with a quantity $d_C(X,A)$ called the *convex* (or sometimes *combinatorial*) distance from X to A.

Let $x \in \mathcal{X}$ and $A \subseteq \mathcal{X}$ measurable. The basic idea for defining $d_C(x,A)$ is to treat $\mathcal{X}$ just like the discrete cube – on each coordinate we'll only keep track of the Boolean variable of whether or not we can go from x to A by varying i. First, we let

$$U'_A(x) = \{(1_{x_i \ne y_i})_{i=1}^n : y \in A\}. \tag{6.13}$$

In words, this is the set of all vectors $\mathbf{1}_J \in Q_n$ for $J \subseteq [n]$ for which there exists $y \in A$ such that x and y differ on exactly the indices in J. Thus, for any $\mathbf{1}_J \in U'_A(x)$, one can get from x to A by changing only those coordinates $j \in J$. (Note that for the case $\mathcal{X} = Q_n$, the set $U'_A(x)$ is exactly the recentered set $A - x$ that we used in the proof of Theorem 6.1, with subtraction taken in $(\mathbb{Z}/2\mathbb{Z})^n$.) We define

$$d_C(x,A) := d_2(0,\mathrm{conv}(U'_A(x))). \tag{6.14}$$

From Lemma 6.3, the above is equivalent to

$$d_C(x,A) := d_2(0,V_A(x)) \tag{6.15}$$

where $V_A(x) := \mathrm{conv}(U_A(x))$ and $U_A(x) := U'_A(x)_\uparrow$ is the set of all vectors $\mathbf{1}_J \in Q_n$ such that one can get from x to A by varying coordinates in J (but possibly not needing to change all

of them). This passage to the monotone set $U_A(x)$ is needed for the same reason as in the proof of Theorem 6.1.

One checks that for the case of the Boolean cube $\mathcal{X} = Q_n$, $d_C(x, A) = d_2(x, \mathrm{conv}(A))$.

With the convex distance thus defined, we can state the general result.

Theorem 6.4 (Talagrand's inequality)**.** *Let X be a random element of a product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with independent components $X_1, \ldots, X_n$. For any $A \subseteq \mathcal{X}$ with $\mathbb{P}(X \in A) > 0$, we have*

$$\mathbb{E} \exp(d_C(X, A)^2/4) \leq \frac{1}{\mathbb{P}(X \in A)}. \tag{6.16}$$

Exercise 6.2. Prove Theorem 6.4 by adapting the proof of Theorem 6.1.

In Exercise 6.1 we used Theorem 6.1 to establish Theorem 5.3 for the case of Bernoulli vectors. To deduce Theorem 5.3 in general we have the following:

Lemma 6.5 (Convex distance controls Euclidean distance)**.** *For any convex $A \subset [0,1]^n$ and $x \in [0,1]^n$, we have $d_2(x, A) \leq d_C(x, A)$.*

Proof. Let $w \in V_A(x)$ be such that $\|w\|_2 = d_C(x, A)$. We can express $w = \sum_{s \in U_A(x)} \lambda_s s$ for weights $\lambda_s \geq 0$ with $\sum_{s \in U_A(x)} \lambda_s = 1$. By definition, for each $s \in U_A(x)$ there exists $z^s = (z_i^s)_{i=1}^n \in A - x$ such that $|z_i^s| \leq s_i$ for all $i \in [n]$ (note that the entries of z^s lie in $[-1, 1]$ since A and x are both contained in $[0,1]^n$). Letting $z := \sum_{s \in U_A(x)} \lambda_s z_s$, we have $|z_i| \leq |w_i|$ for all $i \in [n]$, so $d_2(x, A) = \inf_{y \in A - x} \|y\|_2 \leq \|z\|_2 \leq \|w\|_2 = d_C(x, A)$. □

Exercise 6.3. Use Theorem 6.4 and Lemma 6.5 to prove Theorem 5.3.

Exercise 6.4. Generalize Theorem 5.3 to allow $F : \mathbb{R}^{nd} \to \mathbb{R}$ 1-Lipschitz and convex, and $X \in (\mathbb{B}^d)^n$ with independent components, for arbitrary $d \in \mathbb{N}$, where $\mathbb{B}^d$ is the closed Euclidean unit ball in $\mathbb{R}^d$.

6.3. **A combinatorial perspective on the convex distance.** The convex distance can alternatively be expressed as a supremum over weighted Hamming distances d_H^α, which is useful towards applications to combinatorial optimization, as well as for clarifying the way in which Theorem 6.4 improves over McDiarmid's inequality. Recall the weighted Hamming distances on a product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$:

$$d_H^\alpha(x, y) := \sum_{i=1}^n \alpha_i 1_{x_i \neq y_i} \tag{6.17}$$

for a vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}_+^n$ of positive weights. As usual we take $d_H^\alpha(x, A) = \inf_{y \in A} d_H^\alpha(x, y)$.

Lemma 6.6. *We have*

$$d_C(x, A) = D_H(x, A) := \sup_{\|\alpha\|_2 = 1} d_H^\alpha(x, A). \tag{6.18}$$

Proof. This is perhaps easiest to see with Sion's minimax theorem, but we spell out the argument for both inequalities.

($\leq$): Let $w \in V_A(x) = \mathrm{conv}(U_A(x))$ attain the infimum in the definition of $d_C(x, A)$, thus $\|w\|_2 = d_C(x, A)$. Then the hyperplane through w and perpendicular to w separates the origin from $V_A(x)$, so for every $v \in V_A(x)$ we have $v \cdot w/\|w\|_2 \geq w \cdot w/\|w\|_2 = \|w\|_2$. With

$\alpha = w/\|w\|_2$ we then have $\mathbf{1}_J \cdot \alpha \geq \|w\|_2$ for every $\mathbf{1}_J \in U_A(x)$, and hence $D_H(x,A) \geq d_H^\alpha(x,A) \geq \|w\|_2 = d_C(x,A)$.

($\geq$): Fixing arbitrary $\alpha \in \mathbb{S}^{n-1} \cap \mathbb{R}^n_+$ and $w \in V_A(x)$, it suffices to show $\|w\|_2 \geq d_H^\alpha(x,A)$. We can express $w = \sum_{s \in U_A(x)} \lambda_s s$ for some weights $\lambda_s \geq 0$ such that $\sum_{s \in U_A(x)} \lambda_s = 1$. Then $\|w\|_2 \geq \alpha \cdot w = \sum_s \lambda_s \alpha \cdot s \geq \min_{s \in U_A(x)} \alpha \cdot s = d_H^\alpha(x,A)$ as desired. □

The relation to weighted Hamming distances gives us the following.

Corollary 6.7. *Let $X = (X_1, \dots, X_n)$ be a random element of a product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. Suppose $f : \mathcal{X} \to \mathbb{R}$ has the property that for every $x \in \mathcal{X}$ there exists $\alpha(x) \in \mathbb{S}^{n-1} \cap \mathbb{R}^n_+$ such that*

$$f(x) \leq f(y) + L d_H^{\alpha(x)}(x,y) \qquad \forall y \in \mathcal{X} \tag{6.19}$$

for some $L > 0$. Then for any $a \in \mathbb{R}$,

$$\mathbb{P}(f(X) \leq a)\mathbb{P}(f(X) \geq a + t) \leq \exp(-\frac{t^2}{4L^2}). \tag{6.20}$$

As a consequence by taking a and $a + t$, respectively, to be medians of $f(X)$ in (6.20), we have

$$\mathbb{P}(|f(X) - m_f| \geq t) \leq 4\exp(-\frac{t^2}{4L^2}) \qquad \forall t \geq 0 \tag{6.21}$$

for any median m_f of $f(X)$. From Exercise 4.1 we deduce that $f(X) - \mathbb{E}f(X)$ is $O(L)$-sub-Gaussian.

The freedom in the Lipschitz condition (6.19) to choose weights $\alpha(x)$ depending on the point x can give considerable power over McDiarmid's inequality, particularly when f is defined as a supremum (or infimum) over simpler random variables. We illustrate this with two examples – for further examples we refer to [Ste97, AS16].

Example 6.8 (Largest eigenvalue of a random matrix [AKV02])**.** Let W be a random $n \times n$ symmetric matrix with independent entries W_{ij} on and above the diagonal ranging in $[0,1]$. (This includes, for instance, the adjacency matrix for an Erdős–Rényi graph, with $W_{ii} \equiv 0$ and W_{ij} iid Bernoulli(p) for $i < j$.) We identify $n \times n$ symmetric matrices A with entries $A_{ij} \in [0,1]$ with elements of the $\binom{n+1}{2}$-dimensional product space $[0,1]^S$ where $S = \{(i,j) : 1 \leq i \leq j \leq n\}$. For $A \in [0,1]^S$ let $\lambda_1(A)$ be the right-most eigenvalue of the associated symmetric matrix (a.k.a. the Perron–Frobenius eigenvalue). We claim that for any $A, B \in [0,1]^S$, we have

$$\lambda_1(B) \leq \lambda_1(A) + C d_H^{\alpha(B)}(A,B) \qquad \forall A, B \in [0,1]^S \tag{6.22}$$

for an appropriate unit vector $(\alpha(B)_{ij})_{1 \leq i \leq j \leq n}$ of non-negative weights depending on B. Indeed, fixing such A, B, let v be a unit eigenvector of B with associated eigenvalue $\lambda_1(B)$. By the Courant–Fischer minimax formula we have

$$\begin{aligned}\lambda_1(B) - \lambda_1(A) \leq v^\mathsf{T}(B - A)v &= \sum_{i,j=1}^n v_i v_j (B_{ij} - A_{ij}) \\ &\leq \sum_{i,j=1}^n |v_i v_j| 1_{A_{ij} \neq B_{ij}} = \sum_{i \leq j} (1 + 1_{i \neq j}) |v_i v_j| 1_{A_{ij} \neq B_{ij}}.\end{aligned}$$

Since

$$\sum_{i \leq j} (1 + 1_{i \neq j})^2 |v_i v_j|^2 \leq 2 \sum_{i,j=1}^n v_i^2 v_j^2 = 2$$

we see that (6.22) holds with the unit vector $\alpha(B)$ having entries proportional to $|v_i v_j|$, where the constant of proportionality is of size $\Theta(1)$. From Corollary 6.7 and Exercise 4.1 we conclude that $\lambda_1(W) - \mathbb{E}\lambda_1(W)$ is $O(1)$-sub-Gaussian.

On the other hand, if W_{ij}, $1 \leq i < j \leq n$ are iid Bernoulli(p), it is not hard to show that $\lambda_1(W)$ is typically of order np, the average degree in the associated random graph, so the concentration scale $O(1)$ is drastically smaller than the typical scale of $\lambda_1(W)$. We also note that the same argument as above applies when the entries are centered taking values in a bounded range, in which case $\lambda_1(W)$ is typically of size $\Theta(\sqrt{n})$, which is still much larger than the scale of fluctuations.

Finally, we note that, as with the operator norm in Example 5.7, the true scale of fluctuations in the centered case is $n^{-1/6}$. (One way to see that Talagrand's inequality cannot capture this is that the argument applies equally to centered and uncentered random matrices, and the fluctuation scale $O(1)$ is optimal for the uncentered case considered above.) $\Diamond$

We note that $O(1)$-sub-Gaussian concentration for $\lambda_1(W)$ can also be deduced from Theorem 5.3 by a similar argument as in Example 5.7.

Example 6.9 (Longest increasing subsequence (see also [Tal96, Theorem 6.5])**.** For a finite sequence $x = (x_1, \ldots, x_n) \in [0,1]^n$, let $f(x)$ be the length of the longest increasing subsequence, that is, the largest ℓ such that there exist $i_1 < \cdots < i_\ell$ such that $x_{i_1} < \cdots < x_{i_\ell}$. By a similar argument as in Example 3.4 one can show that f is 1-Hamming Lipschitz (changing one coordinate can only decrease the longest increasing subsequence in length by at most 1), and from Theorem 3.2 we deduce that for $X \in [0,1]^n$ is random with independent components, $f(X)$ has sub-Gaussian concentration around its expectation at scale $O(\sqrt{n})$. However, it turns out that $\mathbb{E}f(X)$ is itself of order $\sqrt{n}$ (see [Ste97]) so this concentration is not even enough to deduce a law of large numbers for $f(X)$.

We can prove stronger concentration by verifying a stronger Lipschitz property of the form (6.19) (in fact we will get a slight variant of a "self-bounding" type). The problem is to come up with the right weight vector $\alpha(x)$ depending on x. For each $x \in [0,1]^n$ let $J(x) \subset [n]$ be a set of indices of size $|J(x)| = f(x)$ that realizes the longest increasing subsequence, thus $x_i < x_j$ for all $i, j \in J(x)$ with $i < j$. (We can select $J(x)$ in some measurable fashion depending on x.) Then the same argument to show $f(x)$ is 1-Hamming Lipschitz actually shows

$$f(x) \leq f(y) + \sum_{i \in J(x)} 1_{x_i \neq y_i} \qquad \forall x, y \in [0,1]^n. \tag{6.23}$$

Indeed, given a longest increasing subsequence of x on indices $J(x)$, then deleting the elements where $x_i \neq y_i$ yields an increasing subsequence for y, whose length is at most $f(y)$. Taking the weight vector $\alpha(x) = f(x)^{-1/2} \mathbf{1}_{J(x)} \in \mathbb{S}^{n-1}$, (6.23) can be reexpressed

$$f(x) \leq f(y) + f(x)^{1/2} d_H^{\alpha(x)}(x, y) \qquad \forall x, y \in [0,1]^n. \tag{6.24}$$

Here we can't directly apply Corollary 6.7 since the coefficient $f(x)^{1/2}$ of the weighted Hamming distance depends on x. Instead we go back to Theorem 6.4 together with the identity of Lemma 6.6. Let $a \in \mathbb{R}, t > 0$ be arbitrary and set $A := \{f \leq a\}, B := \{f \geq a + t\}$. From Equation (6.24) it follows that for any $x \in [0,1]^n$ and $y \in A$,

$$\frac{f(x) - a}{\sqrt{f(x)}} \leq \frac{f(x) - f(y)}{\sqrt{f(x)}} \leq d_H^{\alpha(x)}(x, y).$$

Taking the infimum over $y \in A$, we get

$$\frac{f(x)-a}{\sqrt{f(x)}} \le d_H^{\alpha(x)}(x,A) \le d_C(x,A)$$

by Lemma 6.6. Now since $s \mapsto (s-a)/\sqrt{s}$ is monotone increasing for $s \ge a$, we have for any $t > 0$ that

$$\begin{aligned}\mathbb{P}(f(X) \ge a+t) &\le \mathbb{P}(\frac{f(X)-a}{\sqrt{f(X)}} \ge \frac{t}{a+t}) \\ &\le \mathbb{P}(d_C(X,A) \ge \frac{t}{a+t}) \\ &\le \frac{1}{\mathbb{P}(f(X) \le a)} \exp(-\frac{t^2}{4(a+t)})\end{aligned}$$

where the last bound follows from Theorem 6.4 and Markov's inequality. Taking a to be a median m_f for $f(X)$ and $m_f - t$, respectively, yields the Bernstein-type upper and lower tail bounds

$$\mathbb{P}(f(X) \ge m_f + t) \le 2\exp(-\frac{t^2}{4(m_f+t)}), \qquad \mathbb{P}(f(X) \le m_f - t) \le 2\exp(-\frac{t^2}{4m_f}). \quad (6.25)$$

◊

Remark 6.10. Like the largest eigenvalue/singular value of a centered random matrix, the longest increasing sub-sequence is another example of a random variable that "superconcentrates" at a scale not captured by Talagrand's inequality, and moreover has asymptotically Tracy–Widom fluctuations. For a heuristic explanation of the superconcentration property see the discussion after Theorem 6.5 in [Tal96].

7. Feb 06: Random matrices – bounds on singular values

References:

- These notes
- [Ver18, Chapter 4].

7.1. **Singular values of rectangular matrices.** For an $n \times m$ matrix M with complex entries, the singular values are the eigenvalues of M^*M (or MM^*, up to inclusion of $\max(m,n) - \min(m,n)$ singular values that are trivially zero). We label the first $\min(m,n)$ singular values in nonincreasing order

$$\sigma_1(M) \ge \cdots \ge \sigma_{\min(m,n)}(M) \ge 0$$

with the remaining singular values $\sigma_{\min(m,n)+1}(M) = \cdots = \sigma_{\max(m,n)}(M) = 0$.

Consider the case the case $m \le n$ and the entries of M are real. Then M defines a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$ (that we abusively denote by M). The transformation is injective if and only if $\sigma_m(M) > 0$. In fact, $\sigma_m(M)$ quantifies "how injective" M is. Indeed, from the Courant–Fischer minimax theorem we have

$$\sigma_m(M) = \inf_{u \in \mathbb{S}^{m-1}} \|Mu\|_2, \qquad \sigma_1(M) = \|M\|_{\mathrm{op}} = \sup_{u \in \mathbb{S}^{m-1}} \|Mu\|_2 \quad (7.1)$$

so $\sigma_m(M), \sigma_1(M)$ are the smallest and largest factors, respectively, by which a vector is stretched by M. In order for an additive perturbation $M' = M + A$ to be non-injective, A must have norm at least $\sigma_m(M)$, since

$$\sigma_m(M') \geq \inf_{u \in \mathbb{S}^{m-1}} \big\{ \|Mu\|_2 - \|Au\|_2 \big\} \geq \inf_{u \in \mathbb{S}^{m-1}} \|Mu\|_2 - \|A\|_{\mathrm{op}} = \sigma_m(M) - \|A\|_{\mathrm{op}}.$$

Geometrically, $\sigma_1(M), \dots, \sigma_m(M)$ are the principal radii of the ellipsoid $M\mathbb{B}^n$, i.e. the image under M of the closed Euclidean unit ball. The ellipsoid lies in an m-dimensional subspace of $\mathbb{R}^n$, and the ellipsoid is of maximal dimension m if and only if $\sigma_m(M) > 0$.

7.2. **Singular values of random rectangular matrices: the Marchenko–Pastur law.** Now consider an $n \times m$ matrix X with independent real random entries ξ_{ij} that are centered and of unit variance. (We will later relax some of these distributional assumptions.) Our focus in the next couple of lectures will be to get upper and lower bounds, correct up to constant factors, for the largest and smallest singular values of X.

First we review what asymptotic random matrix theory says. Suppose $n \geq m$ and $m = m_n$ is such that m_n/n convertges to a constant $\alpha \in (0, 1]$ as $n \to \infty$, and for each n we have a random matrix $X^{(n)}$ as above (we will often suppress the dependence on n from the notation). The Marchenko–Pastur law states that for any fixed interval $J \subset \mathbb{R}$,

$$\frac{1}{m} |\{k \in [m] : \sigma_k(\tfrac{1}{\sqrt{n}} X^{(n)})^2 \in J\}| \longrightarrow \nu_\alpha(J) \tag{7.2}$$

in probability, where ν_α is the compactly supported continuous distribution on $\mathbb{R}$ with density

$$\nu_\alpha(dx) = \frac{1}{2\pi\alpha x} \sqrt{(\beta_+ - x)(x - \beta_-)}\, 1_{x \in [\beta_-, \beta_+]} dx \tag{7.3}$$

with respect to Lebesgue measure, where $\beta_\pm := (1 \pm \sqrt{\alpha})^2$. We only focus on the fact that the left and right ends of the limiting support are $\beta_\pm$.

The Marchenko–Pastur law *suggests* that

$$\frac{1}{\sqrt{n}} \sigma_m(X) \to \beta_-^{1/2}, \qquad \frac{1}{\sqrt{n}} \sigma_1(X) \to \beta_+^{1/2} \tag{7.4}$$

in probability as $n \to \infty$. However, since the limiting law ν_α only controls linear proportions of singular values, the Machenko–Pastur law does not rule out the possibility that, say, $n^{0.9}$ singular values escape to 0 or $+\infty$. (It does, however, show that $\limsup \frac{1}{\sqrt{n}} \sigma_m(X) \leq \beta_- + \varepsilon$ and $\liminf \frac{1}{\sqrt{n}} \sigma_1(X) \geq \beta_+ - \varepsilon$ with probability $1 - o_\varepsilon(1)$ for any $\varepsilon > 0$. (Exercise!).)

In particular, we expect that when m, n are larger and $n/m \geq 1 + \delta$ for fixed $\delta > 0$, then all singular values of X should be of size $\asymp \sqrt{n}$ with high probability. This turns out to indeed be the case, at least under some additional tail assumptions on the entries ξ_{ij}.

The Bai–Yin law states that if the entries of $X^{(n)}$ have uniformly bounded fourth moment, i.e.

$$\sup_{n \in \mathbb{N}, i \leq n, j \leq m_n} \mathbb{E} (\xi_{ij}^{(n)})^4 < \infty \tag{7.5}$$

then (7.4) holds in probability. (One can get almost-sure convergence under some slightly stronger assumptions, but as we are eventually concerned with quantitative bounds at finite n we do not comment on this further.) Thus, in the parlance of random matrix theory, we have that the extreme singular values of $X^{(n)}$ "stick to the bulk" (i.e. the edges of the limiting support of the empirical singular value distribution).

The Bai–Yin law is established by the *moment method*, which controls the operator norm via control on spectral moments: noting that for any $\ell \in \mathbb{N}$,

$$\|X\|_{\mathrm{op}}^{2\ell} = \lambda_1(X^\mathsf{T})^{2\ell} \le \sum_{j=1}^{n} \lambda_j(X^\mathsf{T}X)^{2\ell} = \mathrm{Tr}[(X^\mathsf{T}X)^{2\ell}] \tag{7.6}$$

one aims to estimate $\mathbb{E}\,\mathrm{Tr}[(X^\mathsf{T}X)^{2\ell}]$ to leading order for large powers ℓ. Expanding out $(X^\mathsf{T}X)^{2\ell}$, we get a sum over products of 4ℓ entries of X. Many of these terms disappear after taking expectation (for instance any product of entries where some entry appears exactly once in the product, by the independence and centering assumption). It turns out that the leading order contribution comes from "walks" of length 4ℓ with each participating entry appearing exactly twice, and we reduce to a counting problem. We refer to the books [AGZ10, Tao12] for detailed arguments.

We will instead take an easier geometric route to showing a softer bound $O(\sqrt{n})$ of the correct order for the operator norm. The advantage of the geometric approach is that it is much shorter, and also easier to generalize to matrices with structure or dependence among entries.

7.3. **The square case.** Note that when $m = n$, the Bai–Yin law already follows quickly from the Marchenko–Pastur law. Indeed, we have $\sigma_n(X/\sqrt{n}) \ge 0$ for all n, whereas if $\sigma_n(X/\sqrt{n}) \ge \varepsilon$ infinitely often then we get a contradiction to (7.2) since ν_1 has positive density in a neighborhood of 0.

The question is then: what is the order of vanishing of $\sigma_n(X/\sqrt{n})$? In particular we have the very basic question: Is X invertible with high probability?

(7.2) suggests that $\sigma_n(X/\sqrt{n})$ may be of order $1/n$, assuming the singular values are roughly evenly spaced within the limiting support $[0, 4]$, and this turns out to be the case, but this fact was not obtained in any level of generality until the past couple of decades (whereas the problem was brought up by von Neuymann in the 1940s, motivated by his work in numerical analysis for the Los Alamos project).

The invertibility question is trivially "yes, with probability 1" for matrices with densities that are continuous with respect to Lebesgue measure, since the set of singular matrices is a variety of Lebesgue measure zero in the space $\mathbb{R}^{n\times n}$ (the zero set of the determinant polynomial). But it turned out to be a surprisingly subtle question in the discrete case, of which the most basic example is random Bernoulli matrices. The first positive answer in this case came from Komlós in 1967 [Kom67], making an ingenious connection with anticoncentration properties for scalar random walks, which later motivated a long line of refinements of his bound on the singularity probability using methods from additive combinatorics. An optimal bound at exponential scale was only obtained in the last few years by Tikhomirov [Tik20]; we refer to his work and references therein for more history on this problem.

8. Feb 08: Random matrices – bounds on singular values (cont.)

8.1. **Easy arguments.** To get a feel for the problem of bounding the typical size of singular values, we see first see how far we can get from very basic observations. We'll be a bit loose with language (saying "with high probability") but these arguments can be made precise for X having iid entries and suitable tail hypotheses.

First it's not hard to see

$$\sigma_1(X) \gtrsim \sqrt{n} \tag{8.1}$$

with high probability. Indeed, from the variational formula, we see the norm is bounded below by the norm of the first column of X:

$$\sigma_1(X) = \|X\|_{\text{op}} = \sup_{u \in \mathbb{S}^{m-1}} \|Xu\|_2 \geq \|Xe_1\|_2$$

where e_1 is the first canonical basis vector. Now $\|Xe_1\|_2^2 = \sum_{j=1}^n \xi_{j1}^2$ is a sum of independent variables with mean n, so $\|Xe_1\|_2 \gtrsim \sqrt{n}$ with high probability.

On the other hand, it's also not hard to see that most singular values are of size $O(\sqrt{n})$ with high probability, i.e. for any $\varepsilon \in (0,1)$, with high probability

$$\sigma_{\lfloor \varepsilon m \rfloor}(X) \lesssim_\varepsilon \sqrt{n}. \tag{8.2}$$

Indeed, we can express the Frobenius norm of X in two different ways:

$$\sum_{i,j} \xi_{ij}^2 = \|X\|_F^2 = \sum_{k=1}^m \sigma_k(X)^2. \tag{8.3}$$

The left hand side is a sum of independent variables of expectation 1, so from Markov's inequality,

$$\mathbb{P}(\|X\|_F^2 \geq Knm) \leq 1/K$$

for all $K > 0$. Hence, with probability $1 - O(K^{-1})$,

$$\frac{1}{m} \sum_{k=1}^m \sigma_k(X)^2 \leq Kn.$$

On the event that the above bound holds, an application of Markov's inequality to the sum over k shows that all but at most εn of the singular values have size $O_{\varepsilon.K}(\sqrt{n})$, as claimed. (One can improve the probability bound under higher moment assumptions on the entries – for instance if they are sub-Gaussian than we have a Bernstein-type exponential tail for $\|X\|_F^2 - nm$ (exercise!).)

The bounds (8.1) and (8.3) are weak but already capture the correct scale $\sqrt{n}$ for typical singular values.

8.2. **Upper tail for the norm.** Recall that a random vector $X \in \mathbb{R}^n$ is K-sub-Gaussian if $\langle X, u \rangle$ is K-sub-Gaussian for every deterministic $u \in \mathbb{S}^{n-1}$. The following gives a wide class of such vectors:

Lemma 8.1. *Let $X = (\xi_1, \dots, \xi_n)$ have independent K-sub-Gaussian components. Then X is K-sub-Gaussian.*

Proof. Fix an arbitrary $u \in \mathbb{S}^{n-1}$. From (2.6) and homogeneity of the ψ_2-norm,

$$\|\langle X, u \rangle\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|u_i \xi_i\|_{\psi_2}^2 = \sum_{i=1}^n u_i^2 \|\xi_i\|_{\psi_2}^2 \leq K^2 \|u\|_2^2 = K^2$$

as desired. □

Further examples of K-sub-Gaussian vectors not covered by Lemma 8.1 include uniform random points in the scaled sphere $\sqrt{n}\mathbb{S}^{n-1}$, and Gaussian vectors $X \sim N(0, \Sigma)$ (with $K = O(\|\Sigma\|_{\mathrm{op}})$).

In this subsection we prove the following:

Theorem 8.2 (Upper tail for the operator norm)**.** *Let X be $n \times n$ with independent K-sub-Gaussian rows $R_1, \dots, R_n$. Then*

$$\mathbb{P}(\|X\|_{\mathrm{op}} \geq t\sqrt{n}) \leq \exp(-ct^2 n/K^2) \qquad \forall t \geq C_0 K \tag{8.4}$$

for some absolute constant C_0.

Remark 8.3. Note this implies the more general statement assuming X is $n \times m$ with $m \leq n$, since such a matrix can be extended to an $n \times n$ matrix as in Theorem 8.2 by adding $n - m$ columns of zeros.

The proof will be broadly similar to the proof of the Johnson–Lindenstrauss lemma (Theorem 1.5), in that we will first get an upper tail for the norm $\|Xu\|_2$ of the image of a fixed vector u in the sphere, and then get uniform control over all points by taking a union bound. However, in Theorem 1.5 the collection of points is finite from the start. Here we need to control the random continuous function $u \mapsto \|Xu\|_2$ over the *uncountable* set $\mathbb{S}^{n-1}$. This will require a discretization step.

Lemma 8.4. *With X as in Theorem 8.2, let $u \in \mathbb{S}^{n-1}$ be arbitrary and deterministic (or independent of X). Then*

$$\mathbb{P}(\|Xu\|_2 \geq t\sqrt{n}) \leq \exp(-ct^2 n/K^2) \qquad \forall t \geq CK. \tag{8.5}$$

Proof. Fix u. By definition, for each $i \in [n]$ we have

$$\mathbb{E} \exp(\langle X, u\rangle^2/K^2) \leq 2.$$

By independence,

$$\mathbb{E} \exp(\|Xu\|_2^2/K^2) = \prod_{i=1}^{n} \mathbb{E} \exp(\langle X, u\rangle^2/K^2) \leq 2^n.$$

The claim then follows from Markov's inequality. □

Definition 8.5 (ε-net)**.** Let T be a subset of a metric space. A subset $\mathcal{N} \subset T$ is an ε-net for T if for every $x \in T$ there exists $y \in \mathcal{N}$ that is within distance at most ε of x.

Any compact subset of $\mathbb{R}^n$ has an ε-net. However, in high-dimensional probability we tend to need ε-nets of size that is quantitatively controlled in terms of the dimension n. Thus, the size of an ε-net for a set T quantifies "how compact" T is. The logarithm of the minimal size of an ε-net is sometimes called the "metric entropy" of T. The following then says that the metric entropy of the sphere in $\mathbb{R}^n$ is on the order of its dimension n.

Lemma 8.6 (Metric entropy of the sphere)**.** *For any $T \subset \mathbb{S}^{n-1}$ and $\varepsilon \in (0, 1)$, T has an ε-net (under the Euclidean metric on $\mathbb{R}^n$) of size at most $(3/\varepsilon)^n$.*

Proof. We consider the special case $T = \mathbb{S}^{n-1}$ (which is all we need to prove Theorem 8.2), leaving the general case as an exercise. Let $\mathcal{N} \subset \mathbb{S}^{n-1}$ be an ε-separated set that is maximal under the partial order $\subseteq$ of set inclusion. Thus, $\|x - y\|_2 \geq \varepsilon$ for all distinct $x, y \in \mathcal{N}$ and any set $\mathcal{N}' = \mathcal{N} \cup \{z\}$ formed by adjoining a single new element of $\mathbb{S}^{n-1}$ is not ε-separated.

One can obtain such $\mathcal{N}$ by starting with a set of a single point $\mathcal{N}_1 = \{x_1\}$ and for each $k \geq 2$ finding a point x_k that is distance at least ε from $\mathcal{N}_{k-1}$, and setting $\mathcal{N}_k := \mathcal{N}_{k-1} \cup \{x_k\}$. This procedure is guaranteed to stop within a finite number of steps depending only on n and ε (why?), ending with a maximal ε-separated set.

We claim $\mathcal{N}$ is an ε-net of the claimed cardinality. To see that $\mathcal{N}$ is an ε-net, assume toward a contradiction that there exists $z \in \mathbb{S}^{n-1}$ of distance at least ε from every point of $\mathcal{N}$. Then $\mathcal{N} \cup \{z\}$ would be ε-separated, contradicting the maximality assumption.

To see the cardinality bound, note that the set $E = \mathcal{N} + \frac{\varepsilon}{2}\mathbb{B}^n$ (the union of balls of radius $\frac{\varepsilon}{2}$ with centers at the points of $\mathcal{N}$) is a union of $|\mathcal{N}|$ *pairwise disjoint* such balls. Assume for convenience that the balls are open. Indeed, if two of the balls had nonempty overlap, by the triangle inequality we would contradict the assumption that $\mathcal{N}$ is ε-separated. Thus, the Lebesgue measure of E is

$$Leb(E) = |\mathcal{N}| Leb(\tfrac{\varepsilon}{2}\mathbb{B}^n) = |\mathcal{N}|(\varepsilon/2)^n Leb(\mathbb{B}^n).$$

On the other hand, we certainly have $E \subset \frac{3}{2}\mathbb{B}^n$, so

$$Leb(E) \leq Leb(\tfrac{3}{2}\mathbb{B}^n) = (3/2)^n Leb(\mathbb{B}^n).$$

Combining with the previous bound yields the claim. □

(Note how we didn't need to know the volume of $\mathbb{B}^n$ for the above argument.)

Exercise 8.1. Prove the general case of Lemma 8.6.

Exercise 8.2. Formulate and prove a generalization of Lemma 8.6 for $\mathbb{R}^n$ equipped with a general norm $\|\cdot\|$ in place of $\|\cdot\|_2$.

9. Feb 13: Random matrices – singular values and restricted isometry property

9.1. **Concluding the proof of Theorem 8.2.** Now that we have nets of reasonable size, we need a continuity argument to pass from the supremum over $\mathbb{S}^{n-1}$ to a maximum over a finite net.

Lemma 9.1 (Passing to a net)**.** *Let $\varepsilon \in (0,1)$ and let $\mathcal{N}$ be an ε-net for $\mathbb{S}^{n-1}$. For any $m \times n$ matrix M,*

$$\|M\|_{\text{op}} = \sup_{u \in \mathbb{S}^{n-1}} \|Mu\|_2 \leq \frac{1}{1-\varepsilon} \sup_{u \in \mathcal{N}} \|Mu\|_2. \tag{9.1}$$

Proof. Let $v \in \mathbb{S}^{n-1}$ such that $\|Mv\|_2 = \|M\|_{\text{op}}$. There exists $u \in \mathcal{N}$ such that $\|v - u\|_2 \leq \varepsilon$. Then by the triangle inequality and the definition of the operator norm,

$$\|M\|_{\text{op}} = \|Mv\|_2 = \|Mu + M(v-u)\|_2 \leq \|Mu\|_2 + \|M(v-u)\|_2 \leq \|Mu\|_2 + \|M\|_{\text{op}}\|v-u\|_2 \leq \|Mu\|_2 + \varepsilon\|M\|_{\text{op}}.$$

Rearranging and taking the supremum over u completes the proof. □

Proof of Theorem 8.2. From Lemma 8.6 we may fix a $\frac{1}{2}$-net $\mathcal{N}$ for $\mathbb{S}^{n-1}$ of cardinality $|\mathcal{N}| \leq 6^n$. Then from Lemma 9.1 and the union bound,

$$\mathbb{P}(\|X\|_{\text{op}} \geq t\sqrt{n}) \leq \mathbb{P}(\exists u \in \mathcal{N} : \|X\|_{\text{op}} \geq \tfrac{1}{2}t\sqrt{n}) \leq \sum_{u \in \mathcal{N}} \mathbb{P}(\|Xu\|_2 \geq \tfrac{1}{2}t\sqrt{n}).$$

From Lemma 8.4 and taking C_0 sufficiently large, each term in the sum is bounded by $\exp(-ct^2n/K^2)$. Thus,

$$\mathbb{P}(\|X\|_{\mathrm{op}} \geq t\sqrt{n}) \leq |\mathcal{N}| \exp(-ct^2n/K^2) \leq \exp(n(\log 6 - ct^2/K^2)) \leq \exp(-\tfrac{1}{2}ct^2n/K^2)$$

taking C_0 larger if necessary. □

Remark 9.2. The proof of Theorem 8.2 illustrates a common thread to many arguments in high-dimensional probability to get uniform control on an extreme values of a stochastic process $(X_t)_{t\in T}$, where the index set T is a general metric space. By passing to a net and taking a union bound, the upper tail for the supremum $\sup_{t\in T} X_t$ comes down to a competition between the metric entropy of T and the exponential tail for X_t at arbitrary fixed t provided by concentration of measure. A similar but slightly more delicate approach is needed to control the smallest singular value of rectangular matrices, i.e. $N \times n$ with $N \geq (1+\delta)n$ for an arbitrary constant $\delta > 0$ – then the competition is between the metric entropy of $\mathbb{S}^{n-1}$ and *small ball probabilities* for the image of a fixed vector. Later in the course we'll see more advanced arguments based on *chaining*, where one uses a sequence of approximations to the maximizing point t at multiple scales, which is sometimes necessary in order to capture the correct order of the upper tail.

9.2. **Tall isotropic random matrices are almost isometries.** For rectangular matrices of sufficiently large aspect ratio N/n we can control both ends of the singular value distribution using a similar argument as for the proof of Theorem 8.2. Of course, to control the smallest singular value from below we need some additional assumption on the distribution of the rows (recall that Theorem 8.2 covers the matrix of all zeros).

Theorem 9.3 (Very tall sub-Gaussian matrices are almost isometries)**.** *Let X be an $N \times n$ matrix with independent K-sub-Gaussian rows $R_1, \dots, R_N \in \mathbb{R}^n$ that are centered and isotropic, i.e. $\mathbb{E}R_i = 0$ and $\mathbb{E}R_i^{\mathsf{T}}R_i = I_n$. For every $\varepsilon \in (0,1)$, if $N \geq C_0K^4\varepsilon^{-2}n$, then*

$$\mathbb{P}\left(\sup_{u\in\mathbb{S}^{n-1}} \left| \frac{1}{\sqrt{N}}\|Xu\|_2 - 1 \right| \geq \varepsilon \right) \leq \exp(-c\varepsilon^2 N/K^4). \tag{9.2}$$

We may equivalently express the event in (9.2) as

$$1 - \varepsilon \leq \sigma_n(\frac{1}{\sqrt{N}}X) \leq \sigma_1(\frac{1}{\sqrt{N}}X) \leq 1 + \varepsilon. \tag{9.3}$$

In particular we have as an immediate corollary the following non-asymptotic result recovering the correct scaling $1 + O(\sqrt{\alpha})$ of the edges of the support with the aspect ratio $\alpha = n/N$ as in the Bai–Yin theorem.

Corollary 9.4. *If $n/N \leq \alpha$ then with probability at least $1 - e^{-n}$, all of the singular values of $\frac{1}{\sqrt{N}}X$ lie in $[1 - O(K\sqrt{\alpha}), 1 + O(K\sqrt{\alpha})]$.*

We can deduce Theorem 9.3 from the following result of independent interest.

Theorem 9.5 (Quantitative Law of Large Numbers for sample covariance matrices)**.** *With assumptions as in Theorem 9.3, we have*

$$\mathbb{P}(\|\frac{1}{N}X^{\mathsf{T}}X - I_n\|_{\mathrm{op}} \geq \varepsilon) \leq \exp(-c\varepsilon^2 N/K^4)\,. \tag{9.4}$$

In the language of statistics, we consider a collection $R_1, \dots, R_N$ of iid samples from a distribution on $\mathbb{R}^n$ of mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma = \mathbb{E}R_i^{\mathsf{T}}R_i$. The sample mean

$\overline{R} = \frac{1}{N}\sum_{i=1}^N R_i$ provides an unbiased estimator of the *population mean* μ, and the *sample covariance matrix* $\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^N (R_i - \overline{R})^{\mathsf{T}}(R_i - \overline{R})$ is a nearly unbiased estimator of the *population covariance matrix* Σ (a computation shows $\mathbb{E}\widehat{\Sigma} = (1 - \frac{1}{N})\Sigma$, so that this estimator has slight bias). (With slight abuse of terminology, in random matrix theory we often refer to a matrix of the form $\frac{1}{N}X^{\mathsf{T}}X = \frac{1}{N}\sum_{i=1}^N R_i^{\mathsf{T}}R_i$ for centered random row vectors R_i as a sample covariance matrix, removing the small shifts by the sample mean.)

In the centered isotropic setting of Theorem 9.5, the law of large numbers says in the large sample limit $N \to \infty$ with n fixed, the sample covariance matrix $\frac{1}{N}X^{\mathsf{T}}N$ converges to the population covariance matrix I_n (in any norm). Theorem 9.5 refines this to a non-asymptotic result, showing that $\frac{1}{N}X^{\mathsf{T}}X$ is a good approximation for the population mean as soon as N is at least a sufficiently large constant times the dimension n of the data (for fixed K and ε). This result allows both n and N to be large, which is important for modern applications to high-dimensional data.

To see how Theorem 9.5 implies Theorem 9.3, note that for arbitrary fixed $u \in \mathbb{S}^{n-1}$ we have

$$\left|\frac{1}{\sqrt{N}}\|Xu\|_2 - 1\right| \le \left|\frac{1}{N}\|Xu\|_2^2 - 1\right| = \left|\left\langle u, (\tfrac{1}{N}X^{\mathsf{T}}Xu - I_n)u\right\rangle\right| \le \left\|\tfrac{1}{N}X^{\mathsf{T}}X - I_n\right\|_{\mathrm{op}}.$$

Theorem 9.5 can be proved using a net and concentration of measure

Exercise 9.1. Prove Theorem 9.5. *(Hint: use a net to get uniform control on the quadratic form $\langle u, (\frac{1}{N}X^{\mathsf{T}}X - I_n)u\rangle$ for $u \in \mathbb{S}^{n-1}$. For pointwise tails you can use something like Lemma 1.8 from the proof of the Johnson–Lindenstrauss lemma.)*

9.3. **The restricted isometry property for random matrices.** In the field of *compressed sensing*, one is interested in solving an underdetermined linear system

$$y = Ax \tag{9.5}$$

for given $y \in \mathbb{R}^m$ and $m \times n$ matrix A with $n \gg m$. We call A a *measurement matrix* and think of the given data $y = (y_k)_{k=1}^m$ as a list of "measurements" $y_k = \langle r_k, x\rangle$ of the vector x, where r_i are the rows of A. For instance, we might measure a signal x by measuring a few of its Fourier coefficients, taking $r_k = (\exp(2\pi ijk/n))_{j=1}^n$ (or the real or imaginary part of this); in this case A is formed by taking m rows from the discrete Fourier transform matrix.

Of course, from linear algebra we know that if this system has a solution, it is not uniquely determined. However, in many applications we know more about x: that it is sparse in some basis. Then if the measurement vectors r_k are sufficiently "incoherent" in this basis, it turns out that x can be recovered *exactly* by a solving simple convex optimization problem.

There is a natural extension of this problem to incorporate possible noise in the measurements, taking instead

$$y = Ax + w \tag{9.6}$$

where w has, say, independent Gaussian entries of some variance σ^2. However, we don't consider the noisy recover problem in these notes and refer to [Ver18, Chapter 10].

For $v \in \mathbb{R}^n$ we write $\|v\|_0 := |\operatorname{supp}(v)|$ for the cardinality of its support $\operatorname{supp}(v) = \{j \in [n] : v_j \neq 0\}$. A vector v is said to be *r-sparse* if $\|v\|_0 \le r$.

Definition 9.6 (Restricted isometry property)**.** For integers m, n, r and $\varepsilon \in (0,1)$, an $m \times n$ matrix A is said to have the *restricted isometry property* with parameters r, ε (or, "A is

RIP(r,ε)" for short) if

$$(1-\varepsilon)\|v\|_2 \le \|Av\|_2 \le (1+\varepsilon)\|v\|_2 \qquad \text{for all } r\text{-sparse } v \in \mathbb{R}^n. \tag{9.7}$$

Equivalently, A is RIP(r,ε) if

$$1-\varepsilon \le \sigma_r(A_J) \le \sigma_1(A_J) \le 1+\varepsilon \tag{9.8}$$

for all $J \in \binom{[n]}{r}$ (the set of subsets of $[n]$ of size r), where A_J is the $m \times r$ matrix formed by the columns of A lying in J.

The usefulness of the RIP for compressed sensing is encapsulated by the following:

Theorem 9.7. *Let $s, m, n \in \mathbb{N}$ and suppose an $m \times n$ matrix A is RIP(r,ε) for some $\varepsilon \in (0,1)$ and integer $r > \frac{4s}{(1-\varepsilon)^2}$. For any $y \in \mathbb{R}^m$, the unique s-sparse solution to the system (9.5) is given by the (unique) solution $\hat{x}$ to the optimization problem*

$$\text{minimize} \quad \|x\|_1 \quad \text{s.t.} \quad Ax = y. \tag{9.9}$$

Proof. See [Ver18, §10.5.2]. □

It is thus of interest to have measurement matrices with the restricted isometry property, ideally with very few rows compared to the dimension n of the data. Note that an RIP(r,ε) matrix necessarily as at least r rows. As an easy corollary of Theorem 9.3 we see that matrices with independent isotropic sub-Gaussian rows have this property with high probability, as soon as m is at least a log factor larger than the sparsity of the data.

Theorem 9.8 (RIP for random matrices)**.** *Let X be $m \times n$ with independent centered isotropic K-sub-Gaussian rows. Then for every $r \le n$ and $\varepsilon \in (0,1)$, if*

$$m \ge C_0 K^4 \varepsilon^{-2} r \log(\tfrac{en}{r}) \tag{9.10}$$

then $\frac{1}{\sqrt{m}} X$ is RIP(r,ε) with probability at least $1 - \exp(-c\varepsilon^2 m/K^4)$.

Informally, to recover r-sparse signals in $\mathbb{R}^n$, we can use a random matrix as in Theorem 9.8 with $\gg r \log n$ rows.

Proof. From (9.8), we want to show

$$\mathbb{P}(\mathcal{G}) \ge 1 - \exp(-c_0 \varepsilon^2 m/K^4) \tag{9.11}$$

for some constant $c_0 > 0$, where

$$\mathcal{G} := \bigcap_{J \in \binom{[n]}{r}} \mathcal{G}_J\,, \qquad \mathcal{G}_J := \{1-\varepsilon \le \sigma_r(\tfrac{1}{\sqrt{m}} X_J) \le \sigma_1(\tfrac{1}{\sqrt{m}} X_J) \le 1+\varepsilon\}.$$

From Theorem 9.3 and (9.3), for each $J \in \binom{[n]}{r}$ we have

$$\mathbb{P}(\mathcal{G}_J) \ge 1 - \exp(-c_1 \varepsilon^2 m/K^4)$$

for some constant $c_1 > 0$, as long as $m \ge C_0 K^4 \varepsilon^{-2} r$, which holds by our assumption (9.10). Applying the union bound,

$$\mathbb{P}(\mathcal{G}^c) \le \sum_{J \in \binom{[n]}{r}} \mathbb{P}(\mathcal{G}_J^c) \le \binom{n}{r} \exp(-c_1 \varepsilon^2 m/K^4) \le \exp\left(r \log(\tfrac{en}{r}) - c_1 \varepsilon^2 m/K^4\right)$$

where we applied the elementary estimate $\binom{n}{r} \le (\frac{en}{r})^r$ for Binomial coefficients. The claim (9.11) now follows from the assumption (9.10), taking $c_0 = c_1/2$ and C_0 sufficiently large. □

The RIP was introduced in the work [CT06] of Candès and Tao (where it was called the Uniform Uncertainty Principle). That work established Theorem 9.8 for the case that X has iid Gaussian or Rademacher entries, as well as the much more challenging case that the m rows are sampled uniformly and independently (without replacement) from the $n \times n$ discrete Fourier transform matrix, showing that $m \gg r \log^C n$ Fourier coefficients are sufficient for some constant C. This latter result has been sharpened (lowering the order C of the poly-log factor); see [RV08, Bou14, **?**].

The difficulty for establishing RIP for sub-sampled Fourier matrices is that there is *less randomness* (in particular less concentration of measure) to compete with the union bounds taken over all subsets of columns and all points in nets for the respective spheres. These results rely on a more subtle net construction based on an idea going back to an argument of Maurey, and is summarized in the following exercise. The basic idea for efficiently covering a set T with balls is to contain T in a convex set U with a small number of extreme points; an efficient covering of U is then obtained by a probabilistic argument.

Exercise 9.2 (Maurey's empirical method for constructing nets)**.** In this exercise the set of r-sparse unit vectors in $\mathbb{R}^n$ is denoted

$$S_{n,r} = \{u \in \mathbb{S}^{n-1} : |\operatorname{supp}(u)| \leq r\}. \tag{9.12}$$

(a) Let $w_1, \dots, w_m \in \mathbb{R}^n$ be m points in the cube $\mathbb{B}_\infty^n$, i.e. $\|w_i\|_\infty \leq 1$ for each i, and let T be their convex hull. For a given $y = \sum_{k=1}^m \alpha_k w_k \in T$, let $Y_1, \dots, Y_N$ be iid vectors in $\{w_1, \dots, w_m\}$ with distribution $\sum_{k=1}^m \alpha_k \delta_{w_k}$ (so $\mathbb{P}(Y_i = w_k) = \alpha_k$ for each i, k). With $\overline{Y}_N = \frac{1}{N}\sum_{i=1}^N Y_i$ the sample mean, show that for any $\varepsilon > 0$,

$$\mathbb{P}(\|y - \overline{Y}_N\|_\infty > \varepsilon) \leq 2n \exp(-c\varepsilon^2 N).$$

(Hoeffding's inequality will be useful for this.)

(b) Deduce that T can be covered by $\exp(O(\varepsilon^{-2}(\log n)(\log m)))$ translates of $\varepsilon \cdot \mathbb{B}_\infty^n$ with centers in T (i.e. T has an ε-net under the ℓ_∞ metric of size $\exp(O(\varepsilon^{-2}(\log n)(\log m)))$).

(c) Let H be an $n \times n$ matrix with entries bounded by 1. With $S_{n,r}$ as in (9.12), show that $S_{n,r} \subset \sqrt{r}\mathbb{B}_1^n$, and use this to construct an ε-net for $HS_{n,r} = \{Hu : u \in S_{n,r}\}$ under the ℓ_∞ metric of size $\exp(O(\varepsilon^{-2} r (\log n)^2))$. *(Hint: Note that $\mathbb{B}_1^n$ is the convex hull of the $2n$ signed standard basis vectors $\pm e_1, \dots, \pm e_n$.)*

But how about constructing RIP matrices with no randomness at all? Currently the best result direction is by Bourgain et al. [BDF+11], using techniques from additive combinatorics and number theory. This was the first work to break the "square-root barrier" for deterministic constructions, achieving RIP(r, ε) with $m = \Theta(r^{2-c})$ for a small universal constant $c > 0$.

10. Feb 15: Anticoncentration and the smallest singular value

10.1. **The smallest singular value for rectangular matrices.** Theorem 9.3 shows the smallest singular value of an $N \times n$ matrix X with independent isotropic sub-Gaussian rows is of size $\gtrsim \sqrt{N}$ with high probability, provided N is a sufficiently large constant multiple of n. Recall that a key element of the proof was concentration properties of inner products $\langle R_i, u\rangle$ with a row of X and a fixed unit vector $u \in \mathbb{S}^{n-1}$.

We turn now to lower bounds on the smallest singular value $\sigma_n(X)$ for matrices with $N \sim (1+\gamma)n$ for an arbitrary fixed $\gamma \geq 0$, the square case $\gamma = 0$ being the most delicate. As we'll see, the key to lower bounding the smallest singular value boils down to

anti-concentration properties of projections of rows. For this the strong sub-Gaussian tail property is not important – we will hence make much lighter tail assumptions on the entries, only assuming a bounded third moment (which could be relaxed to bounded moments of order $2+\varepsilon$ without much more work). We do however assume the entries of the rows are independent.

We first consider the case that $N \geq (1+\gamma)n$ for arbitrary fixed $\gamma > 0$, which is covered by the following result from [LPRTJ05].

Theorem 10.1. *Let $N, n \in \mathbb{N}$ with $N \geq (1+\gamma)n$ for some $\gamma > 0$. Let X be an $N \times n$ matrix with independent real entries ξ_{ij} satisfying*

$$\mathbb{E}\xi_{ij} = 0\,, \qquad \mathbb{E}\xi_{ij}^2 = 1\,, \qquad \mathbb{E}|\xi_{ij}|^3 \leq A$$

for some finite A. For any $L > 0$, there exist $a = a(\gamma, A, L) > 0$ and $b = b(A) > 0$ such that

$$\mathbb{P}(\sigma_n(X) \leq a\sqrt{N}\,,\ \sigma_1(X) \leq L\sqrt{N}) \leq 2e^{-bN}. \tag{10.1}$$

Remark 10.2. While we only assume a uniformly bounded third moment for the entries, it's worth noting that one needs to assume at least a uniformly bounded *fourth* moment for the event that $\|X\| \leq L\sqrt{N}$ to hold with high probability. In any case, these moment hypotheses are much weaker than the sub-Gaussian hypothesis from Theorem 9.3.

As in the proof of Theorem 9.3, we will use nets to get simultaneous control on the size of $\|Xu\|_2$ for all $u \in \mathbb{S}^{n-1}$. After a union bound over an appropriate net our task is then to bound the probability that $\|Xu\|_2 \leq \varepsilon\sqrt{N}$ for some fixed $u \in \mathbb{S}^{n-1}$ and sufficiently small fixed $\varepsilon > 0$. As in the proof of Theorem 9.3, we note that the squared norm of Xu is a sum of independent random variables:

$$\|Xu\|_2^2 = \sum_{i=1}^{N} \langle R_i, u\rangle^2$$

where R_i are the independent rows of X. For R_i isotropic and sub-Gaussian, $\langle R_i, u\rangle^2$ has sub-exponential concentration around 1. However, here we only assume the entries of R_i have finite third moment. Instead we'll use *anticoncentration* (or *small ball*) estimates to control the event that many of the dot products $\langle R_i, u\rangle$ are small.

We quantify anticoncentration as follows:

Definition 10.3. For a random vector $Y \in \mathbb{R}^d$, the *Lévy concentration function* is

$$\mathcal{L}(Y, t) := \sup_{y \in \mathbb{R}^d} \mathbb{P}(\|Y - y\| \leq t)\,, \qquad t \geq 0. \tag{10.2}$$

That is, $\mathcal{L}(Y, t)$ is the largest measure the distribution of Y assigns to a Euclidean ball of radius t.

To get some intuition, let's first consider the Lévy concentration function for dot products

$$W = \langle R, u\rangle = \sum_{j=1}^{n} \xi_j u_j$$

for the case that $R = (\xi_1, \ldots, \xi_n) \in \{-1, 1\}^n$ is a vector of independent Rademacher variables ξ_i, and let $u \in \mathbb{S}^{n-1}$. Here we write W for "walk", as we view the sum as a random walk, where the vector of step lengths $|u_j|$ is fixed. (In some sense the Rademacher case already captures the essential challenges for sharp anticoncentration estimates for dot products.)

Consider $t \in (0, \frac{1}{10})$, say (we are ultimately interested in arbitrarily small $t > 0$). What is the best bound we can hope to achieve for general $u \in \mathbb{S}^{n-1}$? For the case $u = e_1$ we have

$$\mathcal{L}(\langle R, e_1 \rangle, t) = \mathcal{L}(\xi_1, t) = 1/2,$$

a bound that does not improve as t gets smaller, but at least we are bounded away from 1, which will be useful in some cases. It turns out such a crude bound holds for general u, as shown in the following:

Exercise 10.1 (Crude anticoncentration for scalar random variables)**.** Let ξ be a standardized (centered and unit variance) real random variable.

(a) Show that

$$\mathcal{L}(\xi, \tfrac{1}{4}\mathbb{E}|\xi|) \leq 1 - c_0(\mathbb{E}|\xi|)^2 \tag{10.3}$$

for some absolute constant $c_0 > 0$. Show this bound is sharp in the sense that for arbitrarily small $\varepsilon > 0$ there is a standardized random variable ξ with $\mathbb{E}|\xi| \leq \varepsilon$ and for which the reverse of the above inequality holds (for some possibly modified value of c_0 – you don't need to find the sharp constant).

(b) Show that if we further assume $\mathbb{E}|\xi|^q \leq A$ for some $q > 2$ and $A < \infty$ then

$$\mathcal{L}(\xi, 0.99) \leq 1 - c_1 \tag{10.4}$$

for some $c_1 > 0$ depending only on q and A. (Thus, a mild concentration assumption – namely, the moment bound $\mathbb{E}|\xi|^q \leq A$ – is enough to guarantee some amount of anticoncentration for a standardized variable ξ.)

(Hint: use (or adapt the proof of) the Paley–Zygmund inequality.)

On the other hand, for the case $u = n^{-1/2}(1, \dots, 1)$ we have

$$W = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \xi_j \,. \tag{10.5}$$

The CLT says that this random variable converges in distribution to a standard Gaussian $G \sim N(0,1)$ in the large-n limit. The Gaussian G enjoys a strong anticoncentration bound $\mathcal{L}(G, t) = O(t)$ for all $t \geq 0$. Indeed, since G has a density that is bounded by $1/\sqrt{2\pi}$,

$$\mathcal{L}(G, t) \leq t/\sqrt{2\pi} \quad \forall t \geq 0.$$

This is a strong small ball estimate in the sense that it becomes arbitrarily small as the scale t shrinks. Of course, this doesn't carry over to the discrete random variable (10.5), which assigns measure at least 2^{-n} to all points in its support, and measure as large as $\Theta(n^{-1/2})$ to points in a $O(n^{-1/2})$-neighborhood of the origin (in particular, when n is even we have $\mathbb{P}(W = 0) = \binom{n}{n/2} 2^{-n} \asymp n^{-1/2}$). However, we can deduce a nontrivial anticoncentration bound for W from the following quantitative version of the CLT:

Theorem 10.4 (Berry–Esseen theorem for non-identically distributed summands)**.** *Let $\zeta_1, \dots, \zeta_n$ be independent centered random variables with $\mathbb{E}|\zeta_i|^3 < \infty$ for each $i \in [n]$, set*

$$S = \Big(\sum_{i=1}^{n} \mathbb{E}\zeta_i^2 \Big)^{-1/2} \sum_{i=1}^{n} \zeta_i,$$

and let G be a standard Gaussian variable. For any $t \in \mathbb{R}$ we have

$$|\mathbb{P}(S < t) - \mathbb{P}(G < t)| \lesssim \frac{\sum_{i=1}^n \mathbb{E}|\zeta_i|^3}{\left(\sum_{i=1}^n \mathbb{E}\zeta_i^2\right)^{3/2}}.$$

From Theorem 10.4 we get

$$\mathcal{L}(W,t) = \mathcal{L}(G,t) + O(n^{-1/2}) \lesssim t + n^{-1/2} \quad \forall t \geq 0 \tag{10.6}$$

for W as in (10.5). Informally, the distribution of W behaves like a continuous random variable with bounded density on intervals above scale $n^{-1/2}$.

Exercise 10.2 (Anti-concentration from Berry–Esseen)**.** Using Theorem 10.4, show that if $R = (\xi_1, \dots, \xi_n)$ is uniform in $\{-1,1\}^n$ and $u \in \mathbb{S}^{n-1}$ is a fixed unit vector satisfying

$$\sum_{i=1}^n u_i^2 1_{|u_i| \leq b/\sqrt{n}} \geq a^2$$

then

$$\mathcal{L}(\langle R, u\rangle, t) \lesssim t/a \qquad \forall t \geq \frac{b}{\sqrt{n}}\,.$$

Thus, although $\langle R, u\rangle$ is a discrete random variable, it effectively has bounded density at scales $\gg n^{-1/2}$ if u has a constant proportion of its ℓ_2 mass on coordinates of size $O(1/\sqrt{n})$ (a property that holds for generic $u \in \mathbb{S}^{n-1}$).

(Hint: condition on variables ξ_i for which u_i is large.)

11. Feb 20&22: Metric entropy vs. anticoncentration

11.1. Anticoncentration for the image of a general unit vector.

Lemma 11.1 (Crude anticoncentration for random walks)**.** *Let $R = (\xi_1, \dots, \xi_n)$ be a vector of independent centered random variables with $\mathbb{E}\xi_i^2 = 1$ and $\mathbb{E}|\xi_i|^q \leq A$ for some $q > 2$ and $A < \infty$. Then for any fixed $u \in \mathbb{S}^{n-1}$, we have*

$$\mathcal{L}(\langle R, u\rangle, 0.99) \leq 1 - c_1'$$

for some constant $c_1' > 0$ depending only on q and A.

Proof. We follow an argument from [LPRTJ05]. From part (b) of Exercise 10.1 if suffices to show

$$\mathbb{E}|\langle R, u\rangle|^q \lesssim_{q,A} 1. \tag{11.1}$$

Let $R' = (\xi_i')$ be an independent copy of R, and let (ε_i) be an independent sequence of iid Rademacher variables. From Jensen's inequality and the assumption that the ξ_i are centered, we have

$$\mathbb{E}|\sum_i \xi_i|^q = \mathbb{E}|\sum_i \xi_i - \mathbb{E}\xi_i'|^q \leq \mathbb{E}|\sum_i \xi_i - \xi_i'|^q = \mathbb{E}|\sum_i \varepsilon_i(\xi_i - \xi_i')|^q$$

where in the final equality we used symmetry. Applying Khinchine's inequality, we have

$$\mathbb{E}|\sum_i \xi_i|^q \lesssim \mathbb{E}|\sum_i (\xi_i - \xi_i')^2 u_i^2|^{q/2} \lesssim_q \mathbb{E}|\sum_i \xi_i^2 u_i^2|^{q/2}$$

where in the last bound we used the triangle inequality. Now we note that the function

$$\phi : \Delta_n \to \mathbb{R}_{\geq 0}\,, \qquad \phi(s) := \mathbb{E}|\sum_i \xi_i^2 s_i|^{q/2}$$

is convex on the simplex Δ_n (since $q \geq 2$), and thus it is bounded by its maximum over the extreme points $e_1, \dots, e_n$. This gives

$$\phi(s) \leq \max_i \mathbb{E}|\xi_i|^q \leq A$$

to complete the proof. □

Now we leverage the scalar anticoncentration of Lemma 11.1 to get anticoncentration for the image of a fixed unit vector under X.

Lemma 11.2 (Image of fixed vector)**.** *Let X be an $N \times n$ matrix with independent rows $R_1, \dots, R_N$ satisfying the distribution assumptions of Lemma 11.1. For any fixed $u \in \mathbb{S}^{n-1}$ we have*

$$\mathbb{P}(\|Xu\|_2 \leq c\sqrt{N}) \leq e^{-cN}$$

for some $c = c(q, A) > 0$ depending only on q and A.

Remark 11.3. Note that the lemma makes no assumption on the size of n relative to N, in particular we allow $n > N$. We will later reuse this lemma with $N = n - 1$.

Proof. Let $c > 0$ to be taken sufficiently small over the course of the proof, and fix an arbitrary $u \in \mathbb{S}^{n-1}$. Letting $\beta \in (0, \frac{1}{2}]$ to be chosen later, on the event that

$$\|Xu\|_2^2 = \sum_{i=1}^N \langle R_i, u\rangle^2 \leq c^2 N \tag{11.2}$$

we have that $|\langle R_i, u\rangle| \leq c/\beta$ for at least $(1-\beta^2)N$ values of $i \in [N]$. Applying the union bound to fix these rows and using independence, we have

$$\mathbb{P}(\|Xu\|_2 \leq c\sqrt{N}) \leq \sum_{I \in \binom{[N]}{\lfloor (1-\beta^2)N \rfloor}} \prod_{i \in I} \mathbb{P}\big(|\langle R_i, u\rangle| \leq c/\beta\big). \tag{11.3}$$

Assuming

$$c \leq 0.99\beta \tag{11.4}$$

we can apply Lemma 11.1 to the terms in (11.3) to get

$$\mathbb{P}(\|Xu\|_2 \leq c\sqrt{N}) \leq \binom{N}{\lfloor (1-\beta^2)N \rfloor} (1 - c_1')^{\lfloor (1-\beta^2)N \rfloor} \tag{11.5}$$

Using the bounds $\binom{k}{k-\ell} = \binom{k}{\ell} \leq (\frac{ek}{\ell})^\ell$ and $1 - c_1' \leq e^{-c_1'}$, we thus have

$$\mathbb{P}(\|Xu\|_2 \leq c\sqrt{N}) \leq \exp\Big(N\big(\beta^2 \log \frac{1}{\beta^2} - c_1'(1-\beta^2)\big)\Big) \leq \exp(-\frac{1}{2}c_1' N)$$

where in the last bound we fixed $\beta = \beta(q, A)$ sufficiently small depending on c_1'. Taking $c = \min(0.99\beta, \frac{1}{2}c_1')$ completes the proof. □

As in the proof of Theorem 8.2 will combine Lemma 11.2 with a union bound over a net, via the following:

Lemma 11.4 (Passing to a net)**.** *Let M be an $m \times n$ random matrix, let $\varepsilon > 0$, and let $\mathcal{N}$ be an ε-net for a subset T of $\mathbb{S}^{n-1}$. For any $L > 0$, we have the containment of events*

$$\{\|M\|_{\mathrm{op}} \leq L\sqrt{n}\} \cap \big\{ \inf_{v \in T} \|Mv\|_2 \leq \varepsilon L\sqrt{n} \big\} \subset \big\{ \inf_{u \in \mathcal{N}} \|Mu\|_2 \leq 2\varepsilon L\sqrt{n}\big\}.$$

Proof. On the event on the left hand side, fix $v \in T$ such that $\|Mv\|_2 \le \varepsilon L\sqrt{n}$, and let $u \in \mathcal{N}$ be such that $\|u - v\|_2 \le \varepsilon$. Then

$$\|Mu\|_2 \le \|Mv\|_2 + \|M(u-v)\|_2 \le \|Mv\|_2 + \|M\|_{\mathrm{op}}\|u-v\|_2 \le 2\varepsilon L\sqrt{n}. \qquad \square$$

A straightforward combination of Lemma 11.2 with Lemma 11.4 fails to establish Theorem 10.1. Indeed, letting $\mathcal{N}$ be an ε-net for $\mathbb{S}^{n-1}$ of size $O(1/\varepsilon)^n$, we obtain

$$\mathbb{P}(\sigma_n(X) \le \varepsilon\sqrt{N}, \|X\|_{\mathrm{op}} \le L\sqrt{N}) \le \sum_{u\in\mathcal{N}} \mathbb{P}(\|Xu\|_2 \le 2\varepsilon L\sqrt{N}) \le O(1/\varepsilon)^n e^{-cN}$$

if ε is at most a sufficiently small constant multiple of $1/L$. However, the right hand side above is only a nontrivial bound for $N \ge Cn\log(1/\varepsilon)$, whereas we aim to allow $N \ge (1+\gamma)n$ for arbitrary fixed $\gamma > 0$.

The key will be to split the sphere into two parts and control the infimum over each part by different arguments. The first part is the so-called "compressible" vectors, which are well approximated by δn-sparse vectors for a sufficiently small constant δ. For this set we can follow the above line together with a union bound to fix the support, similarly to how we argued for Theorem 9.8. On the complementary set of "incompressible" vectors we will be able to establish a stronger anticoncentration bound than Lemma 11.2 using Exercise 10.2.

Definition 11.5. Recall the set $S_{n,r}$ of r-sparse unit vectors defined in (9.12). For $\delta, \varepsilon \in (0,1)$, we define the set of (δ,ε)-*compressible* unit vectors to be the ε-neighborhood in $\mathbb{S}^{n-1}$ of $S_{n,\delta n}$, that is

$$\begin{aligned}\mathrm{Comp}(\delta,\varepsilon) &:= \mathbb{S}^{n-1} \cap (S_{n,\delta n} + \varepsilon\cdot\mathbb{B}^n) \\ &= \{u \in \mathbb{S}^{n-1} : \exists v \in \mathbb{S}^{n-1}, \|v\|_0 \le \delta n, \|u-v\|_2 \le \varepsilon\}\end{aligned}$$

and the complementary set of (δ,ε)-*incompressible* unit vectors

$$\mathrm{Incomp}(\delta,\varepsilon) := \mathbb{S}^{n-1} \setminus \mathrm{Comp}(\delta,\varepsilon). \tag{11.6}$$

Denote the boundedness event

$$\mathcal{B}_L := \{\|X\|_{\mathrm{op}} \le L\sqrt{N}\}. \tag{11.7}$$

Proposition 11.6 (Invertibility over compressible vectors)**.** *Let X be as in Lemma 11.2, and assume $n \le 100N$. There exists $c_0 = c_0(q,A) > 0$ such that*

$$\mathbb{P}(\mathcal{B}_L \cap \{\inf_{u\in\mathrm{Comp}(c_0,c_0/L)} \|Xu\|_2 \le c_0\sqrt{N}\}) \le \exp(-c_0 N). \tag{11.8}$$

Exercise 11.1. Use Lemmas 11.2 and 11.4 to prove Proposition 11.6.

Proposition 11.7 (Invertibility over incompressible vectors)**.** *Let X be as in Theorem 10.1. There exists $\beta_0 > 0$ depending only on $\delta, \varepsilon, L, \gamma$ such that*

$$\mathbb{P}(\mathcal{B}_L \cap \{\inf_{u\in\mathrm{Incomp}(\delta,\varepsilon)} \|Xu\|_2 \le \beta_0\sqrt{N}\}) \le e^{-N}$$

for all N sufficiently large.

Since $\mathbb{S}^{n-1} = \mathrm{Comp}(\delta,\varepsilon) \cup \mathrm{Incomp}(\delta,\varepsilon)$, Theorem 10.1 follows from an application of the union bound followed by Propositions 11.6 and 11.7.

Lemma 11.8 (Incompressible vectors are spread)**.** *Let $u \in \mathrm{Incomp}(\delta,\varepsilon)$.*

(1) There exists $J_0 \subset [n]$ with $|J_0| \ge \delta n$ such that $|u_j| \ge \varepsilon/\sqrt{n}$ for all $j \in J_0$.

(2) There exists $J \subset [n]$ with $|J| \geq \delta n/2$ such that $|u_j| \in [\frac{\varepsilon}{\sqrt{n}}, \frac{2}{\sqrt{\delta n}}]$ for all $j \in J$.

Proof. For (1) we take J_0 to be the set of the largest δn coordinates of u. Then u_{J_0}, the projection of u to $\mathbb{R}^{J_0}$, is δn-sparse. If $|u_j| < \varepsilon/\sqrt{n}$ for some $j \in J_0$, then $|u_j| < \varepsilon/\sqrt{n}$ for all $j \in [n] \setminus J_0$ and so $\|u - u_{J_0}\| < \varepsilon$. This implies u is within distance ε of a δn-sparse vector, a contradiction.

Now for (2), since $u \in \mathbb{S}^{n-1}$, $|u_j| > 2/\sqrt{\delta n}$ for at most $\delta n/4$ values of $j \in [n]$ (by Markov's inequality). We can thus obtain the desired set J by removing at most $\delta n/4$ bad elements from J_0. □

Lemma 11.9 (Image of a fixed incompressible vector)**.** *For $u \in \mathrm{Incomp}(\delta, \varepsilon)$,*

$$\mathcal{L}(Xu, t\sqrt{N}) = O_{\delta,\varepsilon}(t)^m \qquad \forall t \geq n^{-1/2}.$$

To prove Lemma 11.9 we combine the Berry–Esseen anticoncentration bound of Exercise 10.2 with the following:

Exercise 11.2 (Tensorization of Anticoncentration))**.** The Lévy concentration function for a random vector $X \in \mathbb{R}^d$ is defined

$$\mathcal{L}(X, t) := \sup_{x_0 \in \mathbb{R}^d} \mathbb{P}(\|X - x_0\|_2 \leq t)\,, \qquad t \geq 0 \tag{11.9}$$

generalizing (10.3) for the case $d = 1$. Suppose $X = (\xi_1, \dots, \xi_d)$ has independent components.

(a) Show that if $\mathcal{L}(\xi_i, a) \leq b$ for some $a > 0$ and $b \in (0, 1)$ and all $i \in [n]$, then $\mathcal{L}(X, c\sqrt{d}) \leq \exp(-cd)$ for some $c > 0$ depending only on a, b.
(b) Show that if $\mathcal{L}(\xi_i, \varepsilon) \leq L\varepsilon$ for all $\varepsilon \geq \varepsilon_0$ and $i \in [n]$, then $\mathcal{L}(X, \varepsilon\sqrt{d}) \leq O(L\varepsilon)^d$ for all $\varepsilon \geq \varepsilon_0$.

(Hint: after fixing x_0, you can control the event that a sum S of independent random variables is small by bounding an inverse exponential moment $\mathbb{E}\exp(-\lambda S)$ for some $\lambda > 0$.)

Proof of Lemma 11.9. This follows by combining the results of Exercise 10.2, Exercise 11.2(b), and Lemma 11.8. □

Proof of Proposition 11.7. By removing rows from X (which can only decrease the smallest singular value and the norm – exercise!) we may assume $N = \lfloor (1+\gamma)n \rfloor$. Let $\beta > 0$ to be chosen sufficiently small, and let $\mathcal{N} \subset \mathbb{S}^{n-1}$ be a β-net for $\mathrm{Incomp}(\delta, \varepsilon)$ of size $O(1/\beta)^n$ (the existence of which is guaranteed by Lemma 8.6). From Lemma 11.4 we have

$$\begin{aligned}\mathbb{P}(\exists u \in \mathrm{Incomp}(\delta, \varepsilon) : \|Xu\|_2 \leq \beta L\sqrt{N}, \mathcal{B}_L) &\leq \mathbb{P}(\exists u \in \mathcal{N} : \|Xu\|_2 \leq 2\beta L\sqrt{N}) \\ &\leq O(1/\beta)^n O_{\delta,\varepsilon}(\beta L)^N = O_{\delta,\varepsilon,L}(\beta^{\gamma/(1+\gamma)})^N\end{aligned}$$

where for the last line we apply Lemma 11.9 and assume $\beta L \geq 1/\sqrt{n}$. The claim follows by taking β sufficiently small depending on δ, ε, L. □

12. Feb 27&29: Square random matrices

See these notes.

13. Feb 29: Suprema of Gaussian processes

Sources: [Ver18, Chapter 7], [vH], [Zei]

We covered:

- General random processes: definitions and examples (random walk, Branching random walk, norm of random matrix)
- The Borell–TIS inequality (see [Zei, Section 2], [Led01, Section 7.1])
- Slepian's inequality (statement and interpretation)

14. Mar 05: Suprema of Gaussian processes: comparison inequalities

Sources: [Ver18, Section 7.2], [vH], [Zei]

We covered:

- Application of Slepian's inequality: speed of the right-most particle in a Gaussian binary branching random walk
- Proof of Slepian by Gaussian interpolation (following [Ver18])

14.1. **The right-most particle in a branching random walk.** We set up some notation to index the edges/vertices of a binary tree of depth n. We index the leaves by $T_n = \{0,1\}^n$. Let $T_{\le n} = \bigcup_{m=1}^n \{m\} \times T_m$, which indexes the edges of the tree (the point is to treat T_m as a disjoint set from T_n rather than as a subset). We associate the elements of $T_{\le n}$ with binary strings of length at most n. For $s \in T_m, t \in T_n$ with $m \le n$, write $s \le t$ if s is a prefix of t.

Let $(g_s)_{s \in T_{\le n}}$ be iid standard Gaussian variables. For each $t \in T_n$ set

$$X_t := \sum_{s \le t} g_s. \tag{14.1}$$

This is a process on the leaves of the tree, with value at leaf t given by the sum of the n Gaussian weights on the edges of the path leading from t back to the root. Thus, $(X_t)_{t \in T_n}$ is a collection of 2^n centered Gaussians of variance n. The correlation structure is determined by the (ultra)metric structure of the tree. For $s, t \in T_n$ let $s \wedge t$ be the generation of their most recent common ancestor, that is

$$s \wedge t = \max\{m : (s_1, \dots, s_m) = (t_1, \dots, t_m)\}. \tag{14.2}$$

Then

$$\mathbb{E} X_s X_t = s \wedge t \tag{14.3}$$

so the canonical metric is

$$d(s,t) = \|X_s - X_t\|_{L^2} = \sqrt{2(n - s \wedge t)}. \tag{14.4}$$

We are interested in the supremum of this process:

$$M_n := \max_{t \in T_n} X_t \tag{14.5}$$

which one can interpret as the position of the right-most particle in the nth generation of a binary Gaussian branching random walk (BRW). A lot is known about M_n – we refer to [Zei16]. Here we will just use Slepian's inequality to get an upper bound on $\mathbb{E} M_n$ that turns out to be sharp to leading order in n.

What is a simpler process we can compare with? Let's consider $(Y_t)_{t\in T_n}$ to be 2^n iid centered Gaussians of variance n. Thus, Y has pointwise the same means and variances as X, but has no correlations. We compute

$$\mathbb{E}(Y_s - Y_t)^2 = 2n \geq \mathbb{E}(X_s - X_t)^2. \tag{14.6}$$

From Slepian's inequality we conclude

$$\mathbb{E}M_n \leq \mathbb{E}\sup_{t\in T_n} Y_t. \tag{14.7}$$

An exercise shows the right hand side is $cn + O(1)$ for $c = \sqrt{2\log 2}$ (an upper bound for the asymptotic speed of the right-most particle that turns out to be sharp). In fact a more careful computation gives a refined asymptotic

$$\mathbb{E}\sup_{t\in T_n} Y_t = cn - \frac{1}{2c}\log n + O(1).$$

An important result going back to Bramson in the setting of branching Brownian motion is that

$$\mathbb{E}M_n = cn - \frac{3}{2c}\log n + O(1). \tag{14.8}$$

Thus, the correlation structure of the BRW shows itself in the sub-leading logarithmic term with a factor 3. This turns out to be a universal feature of extremes of *logarithmically correlated fields* – another example being the planar discrete Gaussian free field. Again we refer to [Zei16] for further background.

15. Mar 07: Suprema of Gaussian processes: comparison inequalities

Source: [Ver18, Sections 7.2–7.4]

- Sudakov–Fernique inequality
- Application: norm of a Gaussian random matrix
- Sudakov Minoration inequality
- Application: Gaussian width and covering numbers for polytopes

16. Mar 19: Chaining

16.1. **Motivation: Uniform laws of large numbers for empirical processes.** The basic Monte Carlo method for approximating an integral $\int f d\mu$ over a probability space $(\mathcal{X}, \Sigma, \mu)$ is to consider the empirical average

$$\mu_n(f) = \int f d\mu_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$$

for an iid sample $X_1, \dots, X_n \sim \mu$, where we define the empirical measure

$$\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}.$$

Note this is a random probability measure.

For a fixed function f that is bounded, say, the Law of Large Numbers tells us $\mu_n(f) \to \mu(f) = \mathbb{E}f(X_1)$ in probability. A *uniform law of large numbers* seeks uniform control over a class of test functions $\mathcal{F}$, i.e. to show

$$\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)| \to 0$$

in probability, say. Here is an example.

Theorem 16.1. *Let $\mathcal{F}$ be the class of 1-Lipschitz functions on $[0,1]$, and let $X_1, \dots, X_n$ be iid in $[0,1]$ with law μ. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)| \lesssim n^{-1/2}.$$

The 1-Wasserstein distance between probability measures μ, ν on a common metric space $(\mathcal{X}, d)$ is

$$W_1(\mu,\nu) = \sup_{f\in\mathcal{F}} |\mu(f) - \nu(f)|$$

where the supremum is taken over the class of 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$. Thus, Theorem 16.1 implies (via Markov's inequailty) that $\mu_n \to \mu$ in probability under the 1-Wasserstein distance.

To prove Theorem 16.1 we will develop a basic result of the chaining method known as Dudley's inequality. However, let's first see what we can get from a more basic approach that we've already applied to estimate the norm of random matrices, combining concentration of measure with a union bound over a net.

Proof of Theorem 16.1 – first attempt. By subtracting constants we may take $\mathcal{F} = \{$ 1-Lipschitz $f : [0,1] \to [-1,1]\}$. We view

$$X_f := \mu_n(f) - \mu(f) = \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i) \tag{16.1}$$

as a stochastic process indexed by $\mathcal{F}$. For any fixed $f \in \mathcal{F}$, from Hoeffding's inequality we have

$$\mathbb{P}(|X_f| \geq \lambda n^{-1/2}) \leq 2\exp(-c\lambda^2), \quad \forall \lambda \geq 0. \tag{16.2}$$

To pass to fixed f we'll apply a union bound over a net, and to pass to a net we need some continuity of $f \mapsto X_f$. For this we note that for any $f, g \in \mathcal{F}$,

$$|X_f - X_g| = |X_{f-g}| \leq 2\|f-g\|_{L^\infty} \quad a.s. \tag{16.3}$$

So let's construct an ε-net $\mathcal{N}_\varepsilon$ for $\mathcal{F}$ under the L^∞ metric. We can take $\mathcal{N}_\varepsilon$ to be the set of functions constant on the intervals $[\varepsilon k, \varepsilon(k+1))$ and taking values in an ε-mesh for $[-1,1]$, which has size $|\mathcal{N}_\varepsilon| = O(1/\varepsilon)^{1/\varepsilon}$. In fact, a more careful use of the Lipschitz property gives a net $\mathcal{N}_\varepsilon$ of size $O(1)^{1/\varepsilon}$ (exercise!).

Letting $\varepsilon \in (0, \frac{1}{2})$ to be chosen later, on the event that $f^* \in \mathcal{F}$ is such that $|X_{f^*}| \geq \sup_{f\in\mathcal{F}} |X_f| - \varepsilon$, there exists $g \in \mathcal{N}_\varepsilon$ such that $\|f^* - g\|_{L^\infty} \leq \varepsilon$, and hence $|X_g| \geq \sup_{f\in\mathcal{F}} |X_f| -$

3ε. We can then apply the union bound and Hoeffding's inequality (16.2): for any $\lambda \geq 6\varepsilon\sqrt{n}$,

$$\begin{aligned}\mathbb{P}(\sup_{f\in\mathcal{F}} |X_f| \geq \lambda n^{-1/2}) &\leq \mathbb{P}(\max_{g\in\mathcal{N}_\varepsilon} |X_g| \geq \tfrac{1}{2}\lambda n^{-1/2}) \\ &\leq |\mathcal{N}_\varepsilon| \exp(-c\lambda^2) \\ &\leq \exp(C\varepsilon^{-1} - c\lambda^2) \\ &\leq \exp(-\tfrac{1}{2}c\lambda^2)\end{aligned}$$

where in the last line we assumed $\lambda \geq C'/\sqrt{\varepsilon}$ for a sufficiently large constant C'. To meet the two constraints we've placed on ε in relation to λ, we cannot take λ any smaller than $Cn^{1/6}$ for a sufficiently large constant C (hence taking ε of order $n^{-1/3}$). We thus obtain the bound

$$\mathbb{E}\sup_{f\in\mathcal{F}} |X_f| \lesssim n^{-1/3} \tag{16.4}$$

which falls short of our goal by a factor $n^{1/6}$ (though for some applications this can be fine!). □

16.2. **Dudley's inequality.** Following [Ver18, Section 8.1], we proved the following:

Theorem 16.2 (Dudley's inequality)**.** *Let $(X_t)_{t\in T}$ be a centered random process on a metric space (T, d) with K-sub-Gaussian increments, i.e.*

$$\|X_s - X_t\|_{\psi_2} \leq Kd(s,t) \quad \forall s,t \in T. \tag{16.5}$$

Then

$$\mathbb{E}\sup_{t\in T} X_t \lesssim K\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N_d(T,\varepsilon)} \tag{16.6}$$

where we recall that $N_d(T,\varepsilon)$ is the minimal cardinality of an ε-net for T. Equivalently (up to modification of the implicit constant),

$$\mathbb{E}\sup_{t\in T} X_t \lesssim K\int_0^\infty \sqrt{\log N_d(T,\varepsilon)}d\varepsilon\,. \tag{16.7}$$

Moreover, the same bound holds for $\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}|$ for any fixed $t_0 \in t$ without the assumption that the process is centered, and consequently also for $\mathbb{E}\sup_{s,t\in T}|X_s - X_t|$.

Lemma 16.3 (1-step chaining, a.k.a. the union bound)**.** *For a finite collection of K-sub-Gaussian variables $(X_t)_{t\in T}$ (not assumed to be independent), we have*

$$\mathbb{P}\Big(\max_{t\in T}|X_k| > \lambda K\sqrt{1+\log|T|}\Big) \leq (e|T|)^{-\lambda^2/2} \qquad \forall\lambda \geq 2 \tag{16.8}$$

and

$$\mathbb{E}\max_{t\in T}|X_t| \lesssim K\sqrt{1+\log|T|}\,. \tag{16.9}$$

Proof. Applying the union bound, the left hand side in (16.8) is

$$\begin{aligned}\mathbb{P}(\exists t\in T : |X_k| > \lambda K\sqrt{\log(e|T|)}) &\leq \sum_{t\in T}\mathbb{P}(|X_k| > \lambda K\sqrt{\log(e|T|)}) \\ &\leq 2|T|\exp(-\lambda^2\log(e|T|)) \\ &\leq \exp(-\tfrac{1}{2}\lambda^2\log(e|T|))\end{aligned}$$

as claimed. Then from Fubini–Tonelli,

$$\begin{aligned}\mathbb{E}\max_{t\in T}|X_t| &= \int_0^\infty \mathbb{P}(\max_{t\in T}|X_t|\geq\lambda)d\lambda \\ &\leq 2\sqrt{K\log|T|}+\int_{2K\sqrt{\log|T|}}^\infty \mathbb{P}(\max_{t\in T}|X_t|\geq\lambda)d\lambda \\ &\lesssim \sqrt{K\log|T|}+\int_{2K\sqrt{\log|T|}}^\infty \exp(-\tfrac{1}{2}\lambda^2/K^2)d\lambda \\ &\lesssim K\sqrt{\log|T|}\,.\end{aligned}$$

□

The chaining argument to prove Theorem 16.2 applies (16.9) at dyadic scales $\varepsilon = 2^{-k}$, with T replaced by a 2^{-k}-net T_k for T. We refer to [Ver18, Section 8.1] for the details.

Exercise 16.1. Prove a matching lower bound for (16.9) for the case that the variables X_t are iid Gaussians.

Exercise 16.2. Let $X_1,\dots,X_n$ be K-sub-Gaussian variables. Show

$$\mathbb{E}\max_{1\leq k\leq n}\frac{X_k}{\sqrt{1+\log k}}\lesssim K. \tag{16.10}$$

16.3. **Proof of Theorem 16.1.** See [Ver18, Section 8.2.2].

17. Mar 21: Covering numbers and VC-dimension

(Based mainly on [Ver18, Section 8.3].)

Now we consider uniform laws of large numbers for classes of Boolean functions on a probability space $(\mathcal{X},\Sigma,\mu)$, that is, measurable functions $f:\mathcal{X}\to\{0,1\}$. (Such functions can be identified with their supports.) In Theorem 16.1 we controlled an empirical process indexed by a class of Lipschitz functions $f:\mathcal{X}\to\mathbb{R}$ using a net under the L^∞ metric. For Boolean functions the L^∞ metric is not very useful, as $\|f-g\|_{L^\infty}=1$ unless $f=g$ μ-a.e., so we don't get any reduction in cardinality by passing to nets. Instead we'll work with the L^2 metric. Toward an application of Dudley's inequality, our task is reduced to bounding the metric entropy numbers $\log N_{L^2(\mu)}(\mathcal{F},\varepsilon)$. These in turn can be controlled in terms of a combinatorial quantity called the VC dimension (the initials are for the originators Vapnik and Chervonenkis).

Definition 17.1 (VC dimension)**.** For a class $\mathcal{F}\subseteq\{0,1\}^{\mathcal{X}}$ of Boolean functions on a domain $\mathcal{X}$, we say a set $A\subseteq\mathcal{X}$ is *shattered by* $\mathcal{F}$ if any Boolean function $g:A\to\{0,1\}$ is the restriction $f|_A$ of some $f\in\mathcal{F}$. The *VC dimension* of $\mathcal{F}$, denoted $vc(\mathcal{F})$, is the cardinality of the largest set $A\subseteq\mathcal{X}$ that is shattered by $\mathcal{F}$ (if there is no largest such set then we set $vc(\mathcal{F}):=\infty$).

With slight abuse of terminology, we can also refer to the VC dimension of a family $\mathcal{A}$ of subsets $A\subseteq\mathcal{X}$ (also called a set system over $\mathcal{X}$), which we take to mean the VC dimension of the associated class of indicator functions $\mathcal{F}_{\mathcal{A}}=\{\chi_A : A\in\mathcal{A}\}$. A set system $\mathcal{A}$ shatters a set $B\subseteq\mathcal{X}$ if $\mathcal{F}_{\mathcal{A}}$ shatters B, i.e. if

$$\{A\cap B : A\in\mathcal{A}\}=2^B$$

(where the right hand side is the power set of B). The left hand side above is often called the *trace* of $\mathcal{A}$ on B, and denoted $\mathrm{Tr}_{\mathcal{A}}(B)$.

Exercise 17.1.

(a) For $\mathcal{X} = \mathbb{R}$, show that the family of closed intervals $\{[a, b] : a, b \in \mathbb{R}, a \leq b\}$ has VC dimension 2.
(b) For $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{H}_d$ be the family of halfspaces and $\mathcal{B}_d$ the family of balls of any radius and center. Show $vc(\mathcal{B}_d) \geq vc(\mathcal{H}_d) \geq d + 1$.
(c) Show $vc(\mathcal{B}_d) = d + 1$ (and hence also $vc(\mathcal{H}_d) = d + 1$).

The Sauer–Shelah lemma, Corollary 17.3 below, is a useful tool for controlling the size of finite set systems in terms of the VC dimension. It is a consequence of the following.

Lemma 17.2 (Pajor's lemma)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a finite set $\mathcal{X}$ with $|\mathcal{X}| = n$, and let*

$$\mathcal{S}_{\mathcal{F}} = \{B \subseteq \mathcal{X} : B \text{ is shattered by } \mathcal{F}\} \tag{17.1}$$

where we include $\emptyset \in \mathcal{S}(\mathcal{F})$. Then

$$|\mathcal{F}| \leq |\mathcal{S}_{\mathcal{F}}|. \tag{17.2}$$

Proof. We proceed by induction on n. For the case $n = 1$, if $\mathcal{F}$ shatters the singleton set $\mathcal{X} = \{x_1\}$ then $|\mathcal{F}| = 2$, while $\mathcal{S}_{\mathcal{F}}) = \{\emptyset, \mathcal{X}\}$ also has two elements. If $\mathcal{F}$ does not shatter $\mathcal{X}$ then $|\mathcal{F}| = 1$ and $\mathcal{S}_{\mathcal{F}} = \{\emptyset\}$ has one element. (So in both cases we actually have equality in (17.2).)

Now assume the claim holds for any set system over any set of size n, and let $\mathcal{X}$ have size $n + 1$. Fix an arbitrary $x_0 \in \mathcal{X}$ and let $\mathcal{X}_0 = \mathcal{X} \setminus \{x_0\}$. The class $\mathcal{F}$ is the disjoint union $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$, with

$$\mathcal{F}_i := \{f \in \mathcal{F} : f(x_0) = i\}, \qquad i = 0, 1.$$

We claim

$$|\mathcal{S}_{\mathcal{F}}| \geq |\mathcal{S}_{\mathcal{F}_0}| + |\mathcal{S}_{\mathcal{F}_1}|. \tag{17.3}$$

To see this, note that any element A of $\mathcal{S}_{\mathcal{F}_i}$, $i = 0, 1$ must be contained in $\mathcal{X}_0$. Also,

$$\mathcal{S}_{\mathcal{F}_i} \subseteq \mathcal{S}_{\mathcal{F}}, \qquad i = 0, 1$$

since $\mathcal{F}_i \subseteq \mathcal{F}$. Finally, we note that if A is shattered by both $\mathcal{F}_0$ and $\mathcal{F}_1$, then both A and $A \cup \{x_0\}$ are shattered by $\mathcal{F}$. From this we deduce (17.3).

Now with $\mathcal{F}_i' = \{f|_{\mathcal{X}_0} : f \in \mathcal{F}_i\}$, which has the same cardinality as $\mathcal{F}_i$, from the induction hypothesis we have

$$|\mathcal{F}_i| = |\mathcal{F}_i'| \leq |\mathcal{S}_{\mathcal{F}_i'}| = |\mathcal{S}_{\mathcal{F}_i}|$$

for $i = 0, 1$. The claim now follows from the above and (17.3). □

Corollary 17.3 (Sauer–Shelah lemma)**.** *Let $\mathcal{A}$ be a set system over a finite set $\mathcal{X}$ with $|\mathcal{X}| = n$, and let $d = vc(\mathcal{A})$ be its VC dimension. Then*

$$|\mathcal{A}| \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

Remark 17.4. The Sauer–Shelah lemma is often invoked in the contrapositive: if a set system $\mathcal{A}$ over a finite set $\mathcal{X}$ of size n has size $|\mathcal{A}| > \sum_{i=0}^{k-1} \binom{n}{i}$, then there is a set $B \subseteq \mathcal{X}$ of size k that is shattered by $\mathcal{A}$.

Proof. By definition any $B \in \mathcal{S}(\mathcal{A})$ has size at most d. The number of subsets of $\mathcal{X}$ of size at most d is $\sum_{i=0}^{d} \binom{n}{i}$, which together with Lemma 17.2 yields the claim (the final bound is left as an exercise). □

The Sauer–Shelah lemma controls the size of classes $\mathcal{F}$ of Boolean functions over finite sets $\mathcal{X}$. We can combine it with a discretization argument to control covering numbers of potentially infinite classes $\mathcal{F}$, giving the following:

Proposition 17.5 (VC dimension controls metric entropy)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a probability space $(\mathcal{X}, \Sigma, \mu)$. Assume $|\mathcal{X}| < \infty$. For any $\varepsilon \in (0,1)$,*

$$\log N_{L^2(\mu)}(\mathcal{F}, \varepsilon) \lesssim vc(\mathcal{F}) \log(2/\varepsilon). \tag{17.4}$$

Remark 17.6. Note this bound has the same shape as the volumetric bound

$$\log N_{\ell^2}(\mathbb{S}^{d-1}, \varepsilon) \leq d \log(3/\varepsilon)$$

for the metric entropy of the sphere from Lemma 8.6.

Proof. It is enough to show that for any set $\mathcal{G} \subset \mathcal{F}$ that is ε-separated in $L^2(\mu)$ we have

$$|\mathcal{G}| \leq O(1/\varepsilon)^{O(vc(\mathcal{F}))}. \tag{17.5}$$

(Note that any ε-separated set $\mathcal{G}$ must be finite. Indeed, the diameter of the set of all Boolean functions on $\mathcal{X}$ is bounded by 1 in $L^2(\mu)$, and $L^2(\mu)$ is finite-dimensional by the finiteness of $\mathcal{X}$, so $\mathcal{G}$ must be finite by volumetric considerations.) Taking $\mathcal{G}$ to be a maximal ε-separated set, we have that $\mathcal{G}$ is an ε-net, and hence $|\mathcal{G}| \geq N_{L^2(\mu)}(\mathcal{F}, \varepsilon)$, and the claim follows from (17.5). (The implicit constant in the base can be adjusted to 2 by adjusting the implicit constant in the exponent.)

Turning to prove (17.5), fix an arbitrary such $\mathcal{G}$. We will apply the Sauer–Shelah lemma not to $\mathcal{G}$, but to the restrictions of its elements to a small randomly sampled subset.

Claim 17.7. There exists a set $\mathcal{Y} = \{y_1, \dots, y_n\} \subset \mathcal{X}$ of size

$$n \lesssim \varepsilon^{-4} \log |\mathcal{G}| \tag{17.6}$$

such that the restrictions $\{g|_{\mathcal{Y}} : g \in \mathcal{G}\}$ are all distinct.

Indeed, we show a random set has this property with high probability. Let $Y_1, \dots Y_n \in \mathcal{X}$ be iid with distribution μ, and let $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$ denote their empirical measure. We will show that $\mathcal{G}$ is $\varepsilon/2$-separated in $L^2(\mu_n)$ with positive probability, from which the claim follows. Consider an arbitrary fixed pair $f, g \in \mathcal{G}$.

$$\|f - g\|_{L^2(\mu_n)}^2 = \frac{1}{n}\sum_{i=1}^{n} |f(Y_i) - g(Y_i)|^2 \tag{17.7}$$

is a normalized sum of n iid random variables almost-surely bounded by 1, with expectation

$$\mathbb{E}|f(Y_i) - g(Y_i)|^2 = \|f - g\|_{L^2(\mu)}^2 \geq \varepsilon^2.$$

From Hoeffding's inequality we have

$$\begin{aligned}\mathbb{P}(\|f - g\|_{L^2(\mu_n)} \leq \varepsilon/2) &\leq \mathbb{P}(|\|f - g\|_{L^2(\mu_n)}^2 - \mathbb{E}\|f - g\|_{L^2(\mu_n)}^2| > \varepsilon^2/2) \\ &\leq 2\exp(-c\varepsilon^4 n).\end{aligned}$$

Taking a union bound over all pairs, we have

$$\mathbb{P}(\exists f, g \in \mathcal{G} : f \neq g, \|f - g\|_{L^2(\mu_n)} \leq \varepsilon/2) \leq |\mathcal{G}|^2 2\exp(-c\varepsilon^4 n) \leq \exp(-\tfrac{1}{2}c\varepsilon^4 n)$$

if $n \geq C\varepsilon^{-4} \log |\mathcal{G}|$ for a sufficiently large constant C, and Claim 17.7 follows.

Now fix a set $\mathcal{Y}$ as in Claim 17.7, and let $\mathcal{G}' = \{g|_{\mathcal{Y}} : g \in \mathcal{G}\}$. From the claim we have that $|\mathcal{G}'| = |\mathcal{G}|$. Since the elements of $\mathcal{G}'$ are restrictions of elements of $\mathcal{G} \subseteq \mathcal{F}$ we have $d' := vc(\mathcal{G}') \leq vc(\mathcal{F})$. From Corollary 17.3, letting $d' = vc(\mathcal{G}')$, we have

$$|\mathcal{G}| = |\mathcal{G}'| \leq \Big(\frac{en}{d'}\Big)^{d'} = O\Big(\frac{\log |\mathcal{G}|}{\varepsilon^4 d'}\Big)^{d'}. \tag{17.8}$$

Then bounding $\frac{1}{d'} \log |\mathcal{G}| = 2 \log(|\mathcal{G}|^{1/2d'}) \leq 2|\mathcal{G}|^{1/2d'}$, substituting this bound on the right hand side above and rearranging yields

$$|\mathcal{G}| = O(\varepsilon^{-4})^{2d'} = O(1/\varepsilon)^{d'} \leq O(1/\varepsilon)^{O(vc(\mathcal{F}))}$$

giving (17.5), which concludes the proof. □

Exercise 17.2. Does Proposition 17.5 extend to the case that $\mathcal{X}$ has infinite cardinality? Prove or give a counterexample.

We can use Corollary 17.3 and Proposition 17.5 to control the supremum over empirical processes indexed by Boolean functions. To pass to control of increments in L^2 rather than L^∞ (as in the proof of Theorem 16.1) we use the following:

Lemma 17.8 (Symmetrization)**.** *Let $\mathcal{F}$ be a class of functions on a probability space $(\mathcal{X}, \Sigma, \mu)$, let $X_1, \dots, X_n$ be iid with distribution μ, and let $\varepsilon_i, i \geq 1$ be iid Rademacher variables, independent from $(X_i)_{i=1}^n$. We have*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n f(X_i) - \mathbb{E} f(X_i) \right| \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|. \tag{17.9}$$

The inequality still holds if we remove the absolute value bars on both sides.

Exercise 17.3.

(a) Prove Lemma 17.8.
(b) Show that the right hand side in (17.9) is further bounded by

$$\sqrt{2\pi}\mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n g_i f(X_i) \right| \tag{17.10}$$

where $g_1, \dots, g_n$ are iid standard Gaussians, and similarly with absolute value bars removed on both sides.
(c) Give a counter example for the first inequality in Vershynin Exercise 7.1.9. *(Hint: you can do this with $N = 1$.)*

Exercise 17.4 (Supremum of a sub-Gaussian process on the hypercube)**.** Let $\xi = (\xi_1, \dots, \xi_n)$ be a K-sub-Gaussian vector in $\mathbb{R}^n$, let $T \subset \{0,1\}^n$, and define a random process $(Y_t)_{t \in T}$ by $Y_t = \sum_{i=1}^n \xi_i t_i$. Let $R := \max_{t \in T} \|t\|_2$ and $d := vc(T)$ (viewing T as a class of Boolean functions on $[n]$).

(a) Use Lemma 16.3 and Corollary 17.3 to show

$$\mathbb{E} \sup_{t \in T} |Y_t| \lesssim KR\sqrt{d \log\Big(\frac{en}{d}\Big)}. \tag{17.11}$$

(b) Suppose that the ξ_i are iid Rademachers (so $K = 1$). Prove a lower bound $\gtrsim d$ for the left hand side of (17.11). In particular, (17.11) holds with equality (up to constant factors) for the case $V = \{0,1\}^n$.

(c) Show that Y_t has K-sub-Gaussian increments with respect to the ℓ^2-metric $d(s,t) = \|s-t\|_2 = (\sum_{i=1}^n (s_i - t_i)^2)^{1/2}$, and deduce

$$\mathbb{E} \sup_{t \in T} |Y_t| \lesssim K \int_0^{2R} \sqrt{\log N_{\ell^2}(T, r)} dr \,. \tag{17.12}$$

Combine this bound with Proposition 17.5 to conclude

$$\mathbb{E} \sup_{t \in T} |Y_t| \lesssim K \sqrt{dn}. \tag{17.13}$$

(Note that a ball of radius r under the ℓ^2-metric has radius $r/\sqrt{n}$ under the L^2-metric, where $\|s-t\|_{L^2} := (\frac{1}{n} \sum_{i=1}^n (s_i - t_i)^2)^{1/2}$.)

Combining Lemma 17.8 with (17.11), for a class $\mathcal{F}$ of Boolean functions on a probability space $(\mathcal{X}, \Sigma, \mu)$ of VC dimension $d = vc(\mathcal{F})$, and $X_1, \dots, X_n$ iid with distribution μ, we have

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X_i) \right| &\le 2 \mathbb{E} \left[\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \,\middle|\, X_1, \dots, X_n \right] \\ &\lesssim \sqrt{\frac{d}{n} \log\left(\frac{en}{d}\right)} \end{aligned} \tag{17.14}$$

where we applied (17.11) to the inner expectation, using that the VC dimension of the (random) family $\mathcal{F}' = \{f|_{\mathcal{X}_n} : f \in \mathcal{F}\}$ of functions restricted to the random set $\mathcal{X}_n = \{X_1, \dots, X_n\}$ is almost-surely bounded by d.

Using Proposition 17.5 and Dudley's chaining inequality instead of directly applying the Sauer–Shelah lemma (Corollary 17.3) and the union bound (Lemma 16.3), we get the following slight improvement over (17.14).

Theorem 17.9 (Uniform LLN for Boolean function classes)**.** *Let $\mathcal{F}$ be a class of measurable Boolean functions on a probability space $(\mathcal{X}, \Sigma, \mu)$ and $X_1, \dots, X_n$ iid with distribution μ. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X_i) \right| \lesssim \sqrt{\frac{vc(\mathcal{F})}{n}} \,. \tag{17.15}$$

Exercise 17.5. Prove Theorem 17.9. *(Hint: you can argue similarly as for the proof of (17.14), using (17.13) in place of (17.11).)*

Combining with esimates on VC dimensions such as the ones in Exercise 17.1 we can obtain discrepancy-type results for iid point clouds, such as the following.

Corollary 17.10 (Discrepancy bound for balls in $\mathbb{R}^d$)**.** *Let $X_1, \dots, X_n$ be iid samples from a distribution μ on $\mathbb{R}^d$, and let $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ be their empirical distribution. Then*

$$\mathbb{E} \sup_{x \in \mathbb{R}^d, r > 0} \left| \mu_n(B(x,r)) - \mu(B(x,r)) \right| \lesssim \sqrt{\frac{d}{n}} \tag{17.16}$$

where $B(x,r)$ is the Euclidean ball of radius r.

18. Mar 26: Generic chaining

Consider a centered process $(X_t)_{t\in T}$ with K-sub-Gaussian increments on a finite state metric space (T,d) (we've seen how we can always reduce to the finite case under separability conditions, which covers applications of interest). Dudley's inequality says

$$\mathbb{E}\sup_{t\in T} \lesssim K\sum_k \varepsilon_k\sqrt{\log|T_k|} \tag{18.1}$$

for any sequence of $\varepsilon_k = 2^{-k}$-nets T_k, which we can take to be the optimal size $|T_k| = |N_d(T,2^{-k})|$.

We can equivalently formulate Dudley's inequality by fixing the size of the T_k's and optimizing the precision ε_k. We say $(T_k)_{k\geq 0}$ is an *admissible sequence* if $|T_0|=1$ and $|T_k|\leq 2^{2^k}$ for all $k\geq 0$. Then setting

$$\varepsilon_k := \sup_{t\in T} d(t,T_k) \tag{18.2}$$

one can show (exercise!) that Dudley's inequality is equivalent to the bound

$$\mathbb{E}\sup_{t\in T} X_t \lesssim K\sum_{k\geq 0} 2^{k/2}\sup_{t\in T} d(t,T_k). \tag{18.3}$$

Now consider the following quantity, which in general is smaller than the right hand side above:

$$\gamma_2(T,d) := \inf_{(T_k)_{k\geq 0}}\sup_{t\in T}\sum_{k=0}^{\infty} 2^{k/2} d(t,T_k) \tag{18.4}$$

where the infimum is taken over all admissible sequences. This is called the γ_2 *functional* for (T,d) (following Talagrand). We have the following improvement of Dudley's inequality:

Theorem 18.1 (Talagrand–Fernique)**.** *For a centered process $(X_t)_{t\in T}$ with K-sub-Gaussian increments on a metric space (T,d), we have*

$$\mathbb{E}\sup_{t\in T} X_t \lesssim K\gamma_2(T,d). \tag{18.5}$$

Proof. See [Ver18, Section 8.5]. □

We point out that the proof of Theorem 18.1 is not much harder than the proof of Dudley's inequality. The following result deep result of Talagrand is harder and we do not give the proof. We refer to Talagrand's book [Tal21] for a thorough treatment of the theory of generic chaining and its applications.

Theorem 18.2 (Talagrand's majorizing measures theorem)**.** *For $(X_t)_{t\in T}$ a centered* Gaussian *process, we have*

$$\mathbb{E}\sup_{t\in T} X_t \asymp \gamma_2(T,d) \tag{18.6}$$

where $d(s,t) := \|X_s - X_t\|_{L^2}$ is the canonical metric for the process.

The γ_2 functional can be hard to compute in practice. (There are some open problems that reduce to such a computation by the above theorem.) The following important corollary does not involve the γ_2-functional.

Corollary 18.3 (Talagrand's comparison inequality)**.** *Let $(X_t)_{t\in T}, (Y_t)_{t\in T}$ be two centered processes on a set T, with Y Gaussian, and assume the increments satisfy the comparison bound*

$$\|X_s - X_t\|_{\psi_2} \le K\|Y_s - Y_t\|_{L^2} \quad \forall s,t \in T. \tag{18.7}$$

Then

$$\mathbb{E}\sup_{t\in T} X_t \lesssim K\mathbb{E}\sup_{t\in T} Y_t. \tag{18.8}$$

Proof. Denoting by d the canonical metric associated to the Gaussian process Y, from Theorem 18.1 and Corollary 18.3 we have

$$\mathbb{E}\sup_{t\in T} X_t \lesssim K\gamma_2(T,d) \lesssim K\mathbb{E}\sup_{t\in T} Y_t\,. \qquad \square$$

Corollary 18.3 allows one to obtain bounds for general sub-Gaussian processes by similar arguments to how we applied the Slepian and Sudakov–Fernique inequalities in the Gaussian case (we note however that unlike those comparison inequalities, the above loses a universal constant factor).

Exercise 18.1. Use Corollary 18.3 to give a new proof (see Theorem 8.2) of the bound $\mathbb{E}\|A\|_{\mathrm{op}} \lesssim \sqrt{n}$ for an $n\times n$ random matrix with independent centered 1-sub-Gaussian entries.

Recall the Gaussian width

$$w(T) := \mathbb{E}\sup_{t\in T}\langle g, t\rangle \tag{18.9}$$

for a subset $T \subseteq \mathbb{R}^n$, where $g \sim N(0, I_n)$. As a special case of Corollary 18.3 we get the following:

Corollary 18.4. *Let $(X_t)_{t\in T}$ be a centered process on $T \subseteq \mathbb{R}^n$ with sub-Gaussian increments under the Euclidean metric:*

$$\mathbb{E}|X_s - X_t|^2 \le K^2\|s-t\|_2^2 \quad \forall s,t\in T. \tag{18.10}$$

Then

$$\mathbb{E}\sup_{t\in T} X_t \lesssim Kw(T). \tag{18.11}$$

19. Mar 28: Entropy methods – subadditivity, LSI on the cube

The next few lecture's we'll develop the theory of entropy for functions on product spaces, and consider proofs and applications for some log-Sobolev inequalities and hypercontractivity inequalities. The main examples will be the uniform measure on the Boolean hypercube and the standard Gaussian measure on $\mathbb{R}^n$. The main references for this material are

- [BLM13, Chapters 4–5]
- [Led]
- [Led01]
- [vH]

These notes may at times have different emphasis from the above.

19.1. **Definitions and the duality formula.** We consider a measurable space $(\mathcal{X}, \Sigma)$ with a probability measure μ. (We prefer not to write Ω instead of $\mathcal{X}$, as we'll tend to view $\mathcal{X}$ as the *range* of a random element $X \in \mathcal{X}$ with distribution μ, following the probabilistic convention of hiding the sample space from view and leaving it flexible to accommodate additional sources of randomness.)

For $x \geq 0$ denote

$$\Phi(x) := x \log x \tag{19.1}$$

(taking $\Phi(0) = \lim_{x \downarrow 0} \Phi(x) = 0$). This is a convex function on $\mathbb{R}^+$.

For a non-negative function $f \in L^1(\mu)$ we define

$$\mathrm{Ent}(f) = \mathrm{Ent}_\mu(f) := \int \Phi(f) d\mu - \Phi(\int f d\mu) \in [0, +\infty]. \tag{19.2}$$

We can alternatively state this probabilistically as

$$\mathrm{Ent}(f) := \mathbb{E}\Phi(f(X)) - \Phi(\mathbb{E}f(X)) \tag{19.3}$$

for $X \sim \mu$. The lower bound $\mathrm{Ent}(f) \geq 0$ follows from Jensen's inequality and the convexity of Φ.

We note that the entropy is 1-homogeneous: for any constant $c \geq 0$,

$$\mathrm{Ent}(cf) = c\,\mathrm{Ent}(f)\,. \tag{19.4}$$

This will allow us to scale f to have mean 1 (i.e. $\inf f d\mu = \mathbb{E}f(X) = 1$) in many of the proofs.

We point out a connection to the relative entropy. For the case that

$$\int f d\mu = \mathbb{E}f(X) = 1 \tag{19.5}$$

we have that $d\nu = f d\mu$ is a probability measure on $(\mathcal{X}, \Sigma)$, i.e. a measure $\nu \ll \mu$ with Radon–Nikodym derivative $\frac{d\nu}{d\mu} = f$. Then we have

$$\mathrm{Ent}_\mu(f) = \int \Phi(f) d\mu = \int f \log f d\mu = \int (\log \frac{d\nu}{d\mu}) d\nu = \mathrm{D}(\nu \| \mu) \tag{19.6}$$

where we recall the *relative entropy* or *Kullback–Leibler divergence* of ν with respect to μ is defined as above when $\nu \ll \mu$, and taken to be $+\infty$ otherwise. Thus, the entropy of a non-negative function f of mean 1 is simply the relative entropy of the reweighted measure $f d\mu$ with respect to the reference measure μ.

We will generally have a fixed reference measure μ that will be a product measure (or even nicer – a uniform measure on a finite product space) and consider bounds on the entropy over classes of functions f.

We also note that the relative entropy makes sense when μ is not necessarily a probability measure – it is often convenient to consider entropies relative to an unnormalized Lebesgue or counting measure – the latter case leads to the better known *Shannon entropy* (up to flipping the sign) – we may review this connection in a later lecture, though we are running short on time!

Lemma 19.1 (Duality formula)**.** *We have*

$$\mathrm{Ent}_\mu(f) = \sup_g \left\{ \int f g d\mu : \int e^g d\mu \leq 1 \right\} = \sup_g \left\{ \int f g d\mu : \int e^g d\mu = 1 \right\} \tag{19.7}$$

and the supremum is attained at $g = \log f - \log \int f d\mu$.

(For the latter equality note that the objective function $g \mapsto \int f g d\mu$ is monotone under increasing g.)

We may use this identity in the probabilistic notation:

$$\mathrm{Ent}_\mu(f) = \sup_Y \left\{ \mathbb{E} f(X) Y : \mathbb{E} e^Y \leq 1 \right\}. \tag{19.8}$$

Proof. By the homogeneity property (19.4) we may assume $\int f d\mu = 1$. Fix an arbitrary g such that $\int e^g d\mu = 1$. Setting $d\nu = e^g d\mu$, we have

$$\begin{aligned} 0 \leq \mathrm{Ent}_\nu(f e^{-g}) &= \int f e^{-g} (\log f - g) e^g d\nu - \int f e^{-g} e^g d\nu \log \int f e^{-g} e^g d\nu \\ &= \int f \log f d\mu - \int f g d\mu = \mathrm{Ent}_\mu(f) - \int f g d\mu \end{aligned}$$

and the claim follows. □

19.2. **Subadditivity of the entropy on product spaces.** We will henceforth consider the case that $(\mathcal{X}, \Sigma, \mu)$ is a product probability space, i.e. $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ for probability spaces $(\mathcal{X}_i, \Sigma_i, \mu_i)$. We continue to write X for a random element of $\mathcal{X}$ with distribution μ. Thus, $X = (X_1, \ldots, X_n)$ has independent components.

Some notation: for given $x \in \mathcal{X}$ we'll write

$$x_{(i)} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \mathcal{X}_{(i)} := \prod_{j \neq i} \mathcal{X}_j. \tag{19.9}$$

When the product reference measure μ is clear from the context, we'll write

$$\mathrm{Ent}_i(f) := \mathrm{Ent}_{\mu_i}(f(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_n)) \tag{19.10}$$

That is, we fix all coordinates of x but the ith coordinate and take the entropy of the resulting function on $\mathcal{X}_i$ with respect to μ_i. Thus, $\mathrm{Ent}_i(f)$ is a function on $\mathcal{X}_{(i)} := \prod_{j \neq i} \mathcal{X}_j$. Viewed as a random variable, it is measurable under the σ-algebra $\sigma(X_{(i)})$ generated by the variables $\{X_j\}_{j \neq i}$, with

$$\mathrm{Ent}_i(f) = \mathbb{E}_{X_i} \Phi(f(X)) - \Phi(\mathbb{E}_{X_i} f(X)) \tag{19.11}$$

with $\mathbb{E}_{X_i}$ denoting expectation conditional on $X_{(i)}$.

Proposition 19.2 (Subadditivity / tensorization of the entropy)**.** *In the above setup, we have*

$$\mathrm{Ent}(f) \leq \int \sum_{i=1}^n \mathrm{Ent}_i(f) d\mu \,. \tag{19.12}$$

Or, probabilistically:

$$\mathrm{Ent}(f) \leq \mathbb{E} \sum_{i=1}^n \mathbb{E}_{X_i} \Phi(f(X)) - \Phi(\mathbb{E}_{X_i} f(X)) \,. \tag{19.13}$$

Remark 19.3. In the discrete case that $\mathcal{X}$ is countable, Proposition 19.2 is equivalent to Han's inequality for the Shannon entropy – see [BLM13, Chapter 4], where it is proved by an alternative route based on the chain rule for Shannon entropy.

Proposition 19.2 has a well-known analogue for the variance, under replacing Φ with the convex function $x \mapsto x^2$.

Corollary 19.4 (Effron–Stein inequality)**.** *In the above setup, for $f \in L^2(\mu)$ we have*

$$\operatorname{Var} f(X) \leq \mathbb{E} \sum_{i=1}^{n} (f(X) - \mathbb{E}_{X_i} f(X))^2. \tag{19.14}$$

The Effron–Stein inequality can be proved by considering a martingale difference sequence, but it is also a quick consequence of the subadditivity of the entropy.

Exercise 19.1. Prove Corollary 19.4 by applying Proposition 19.2 with the function $1 + \varepsilon f$ in place of f, and sending ε to zero.

Proof of Proposition 19.2. Fixing an arbitrary function g such that $\int e^g d\mu \leq 1$, by Lemma 19.1 it suffices to show

$$\int f g d\mu \leq \sum_{i=1}^{n} \operatorname{Ent}_i(f) d\mu. \tag{19.15}$$

Define random variables

$$Z_i := \log \frac{\mathbb{E}(e^{g(X)} | X_{\geq i})}{\mathbb{E}(e^{g(X)} | X_{>i})}, \quad 1 \leq i \leq n \tag{19.16}$$

where we write $X_{\geq i} := (X_i, \ldots, X_n)$, $X_{>i} := X_{\geq i+1}$. Then

$$\sum_{i=1}^{n} Z_i = g(X) - \log \mathbb{E} e^{g(X)} \geq g(X) \quad a.s.$$

since we assumed $\mathbb{E} e^{g(X)} \leq 1$. Furthermore, for each i we have

$$\mathbb{E}_{X_i} Z_i = 1. \tag{19.17}$$

Hence,

$$\int f g d\mu \leq \sum_{i=1}^{n} \mathbb{E} f(X) Z_i = \mathbb{E} \sum_{i=1}^{n} \mathbb{E}_{X_i} f(X) Z_i \leq \mathbb{E} \sum_{i=1}^{n} \mathbb{E}_{X_i} \Phi(f(X)) - \Phi(\mathbb{E}_{X_i} f(X)) \tag{19.18}$$

as desired, where in the final bound we applied Lemma 19.1 to the factors μ_i. □

19.3. **The log-Sobolev inequality for the discrete hypercube.** Now we consider the product space $\mathcal{X} = \{-1, 1\}^n$ with the uniform measure μ. Thus $X \sim \mu$ has independent Rademacher components X_i. Throughout this section we write $\operatorname{Ent}(f)$ for $\operatorname{Ent}_\mu(f)$.

For $x = (x_1, \ldots, x_n) \in \mathcal{X}$ we denote by

$$\overline{x}^{(i)} := (x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_n)$$

the vector with the ith coordinate flipped. We define the discrete gradient of a function $f : \mathcal{X} \to \mathbb{R}$ by

$$\nabla f(x) = (\nabla_1 f(x), \ldots, \nabla_n f(x)), \quad \nabla_i f(x) := \frac{1}{2}(f(x) - f(\overline{x}^{(i)})). \tag{19.19}$$

For $X \sim \mu$ we denote by

$$\widetilde{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n) \tag{19.20}$$

the vector with ith coordinate independently resampled, where $X' = (X'_1, \dots, X'_n)$ is an independent copy of X. Let

$$\mathcal{E}(f) := \frac{1}{2}\mathbb{E}\sum_{i=1}^{n}(f(X) - f(\widetilde{X}^{(i)}))^2 = \frac{1}{4}\mathbb{E}\sum_{i=1}^{n}(f(X) - f(\overline{X}^{(i)}))^2 = \mathbb{E}\|\nabla f(X)\|_2^2 . \tag{19.21}$$

Theorem 19.5 (Log-Sobolev inequality for the hypercube)**.** *For any $f : \{-1,1\}^n \to \mathbb{R}$,*

$$\mathrm{Ent}_\mu(f^2) \le 2\mathcal{E}(f) . \tag{19.22}$$

Proof. From Proposition 19.2,

$$\mathrm{Ent}(f^2) \le \int \sum_{i=1}^{n} \mathrm{Ent}_i(f^2) d\mu$$

where we recall $\mathrm{Ent}_i(f^2)$ is a function of $x_{(i)}$. Fixing an arbitrary i and conditioning on $X_{(i)}$, we see it suffices to show

$$\mathrm{Ent}_i(f^2) \le \frac{1}{2}\mathbb{E}_{X_i}(\nabla_i f(X))^2 \quad a.s.$$

which would follow from the case $n = 1$ of the theorem. So we are reduced to the case $n = 1$.

Now assuming $n = 1$, writing $a := f(1)$ and $b = f(-1)$, our aim is to show

$$\frac{1}{2}a^2\log(a^2) + \frac{1}{2}b^2\log(b^2) - \frac{a^2+b^2}{2}\log\frac{a^2+b^2}{2} \le \frac{1}{2}(a-b)^2 \tag{19.23}$$

for all $a, b \in \mathbb{R}$. By symmetry we may assume $0 \le b \le a$. For fixed $b \ge 0$ let $h_b : [b, \infty) \to \mathbb{R}$ be given be the difference of the left hand side and the right hand side above. One verifies that $h_b(b) = h'_b(b) = 0$ and that h_b is concave, and hence $h_b(a) \le 0$ for all $a \ge b$. The claim follows. □

Theorem 19.5 has the following generalization to non-centered product measures on the cube.

Theorem 19.6. *For μ_p the distribution of a vector X with iid components X_i with $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = -1) = p$, we have*

$$\mathrm{Ent}_{\mu_p}(f^2) \le C(p)\mathcal{E}(f)$$

for all $f : \{-1,1\}^n \to \mathbb{R}$, where $C(p) = \frac{1}{1-2p}\log\frac{1-p}{p}$. (Note that $C(p) \to 2$ as $p \to \frac{1}{2}$.)

Proof. See [BLM13], [Led]. □

20. Apr 02: Entropy methods – Gaussian LSI, the Herbst argument

- Deduction of Gaussian LSI from the LSI for the hypercube and CLT [Led]
- Herbst argument for sub-Gaussian concentration from LSI [BLM13]

21. Apr 04: Entropy methods – general Markov semigroups

- Source: [vH] Section 2.2 and Chapter 8

22. Apr 09: Entropy methods – hypercontractivity and threshold phenomena

- Source: [vH, Chapter 8]

23. Apr 11: Student presentations

- Sofia, Po-Ying and Yuanxin: *Sharp nonasymptotic bounds on the norm of random matrices with independent entries*, A. Bandeira and R. van Handel [BvH16]
- Victor, Bryan and Haotian: *Stein's method for concentration inequalities*, S. Chatterjee [Cha07]

24. Apr 16: Student presentations

- Jiayi, Yixin and Nathan: *Four Talagrand inequalities under the same umbrella*, M. Ledoux [Led]
- Kai, Angikar and Adway: *Testing for high-dimensional geometry in random graphs*, S. Bubeck, J. Ding, R. Eldan and M. Rácz [BDER16]

References

[AGZ10] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[AKV02] Noga Alon, Michael Krivelevich, and Van H. Vu. On the concentration of eigenvalues of random symmetric matrices. *Israel J. Math.*, 131:259–267, 2002.

[AS16] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., Hoboken, NJ, fourth edition, 2016.

[BDER16] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z. Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures Algorithms*, 49(3):503–532, 2016.

[BDF+11] Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, and Denka Kutzarova. Explicit constructions of RIP matrices and related problems. *Duke Math. J.*, 159(1):145–185, 2011.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[Bou14] Jean Bourgain. An improved estimate in the restricted isometry problem. In *Geometric aspects of functional analysis*, volume 2116 of *Lecture Notes in Math.*, pages 65–70. Springer, Cham, 2014.

[BvH16] Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 2016.

[Cha07] Sourav Chatterjee. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138(1-2):305–321, 2007.

[Cha14] Sourav Chatterjee. *Superconcentration and related topics*. Springer Monographs in Mathematics. Springer, Cham, 2014.

[CT06] Emmanuel J. Candes and Terence Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[Kom67] J. Komlós. On the determinant of (0, 1) matrices. *Studia Sci. Math. Hungar*, 2:7–21, 1967.

[Led] Michel Ledoux. Four talagrand inequalities under the same umbrella. arXiv:1909.00363.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[LPRTJ05] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Adv. Math.*, 195(2):491–523, 2005.

[MS86] Vitali D. Milman and Gideon Schechtman. *Asymptotic theory of finite-dimensional normed spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986. With an appendix by M. Gromov.

[RV08] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[RV10] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values; proceedings of the international congress of mathematicians. volume iii. pages 1576–1602, 2010.

[Ste97] J. Michael Steele. *Probability theory and combinatorial optimization*, volume 69 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[Tal95] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, (81):73–205, 1995.

[Tal96] Michel Talagrand. A new look at independence. *Ann. Probab.*, 24(1):1–34, 1996.

[Tal21] Michel Talagrand. *Upper and lower bounds for stochastic processes—decomposition theorems*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Cham, [2021] ©2021. Second edition [of 3184689].

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[Tik20] Konstantin Tikhomirov. Singularity of random Bernoulli matrices. *Ann. of Math. (2)*, 191(2):593–634, 2020.

[Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[vH] Ramon van Handel. Probability in high dimensions. Lectures notes. https://web.math.princeton.edu/∼rvan/APC550.pdf.

[Zei] Ofer Zeitouni. Gaussian fields. Lecture notes. https://cims.nyu.edu/∼zeitouni/notesGauss.pdf.

[Zei16] Ofer Zeitouni. Branching random walks and gaussian fields. In *Proceedings of Symposia in Pure Mathematics*, volume 91, pages 437–471. Amer. Math. Soc., Providence, R.I., 2016.

⋆Department of Mathematics, Duke University, 120 Science Dr, Durham, NC 27710

Email address: nickcook@math.duke.edu