STAT 216 NOTES

KEVIN CONG

1. Lecture 1: September 4

1.1. **Admin.** The following administrative facts will hold.

- Grading: 60% HW + 40% project
- Project in groups of 2; initial writeup due 9/30, midterm progress 10/31

Reference textbooks:

1.2. **Introduction and Course Plan.** We will roughly speaking have modules on the following topics in high-dimensional probability:

(1) Concentration Inequalities
(2) Non-asymptotic theory for random matrices
(3) Gaussian comparison: in applications, we have jointly distributed Gaussian variables, and we care about, say, the distribution of the supremum.
(4) Universality

Remark 1.1. Professor Yue M. Lu is teaching a related class, AM254, on the learning-theoretic areas and connections to machine learning. Lectures are on T/Th.

Remark 1.2 (Remark on Notation)**.** I will use RV to denote random variable or random vector, where the meaning is clear, and RM to denote random matrix.

1.3. **Introduction of Content.** Obviously, we start with the question: **what is high-dimensional probability?**

The core point of the field is to understand probability distributions on high-dimensional spaces. Intriguingly, we might care about usage of geometric methods, information-theoretic methods, etc.

Example 1.3. Consider the scenario where X_i are iid RVs, $T = f(X_1, \dots, X_n)$ is an RV for some 'nice' but also implicit function f. Typically, the variable T will not depend strongly on any X_i, so we can get generalizations of classical limit theorems, e.g. the SLLN.

In general, we will decouple T into

$$T = [T - \mathbb{E}T] + \mathbb{E}T.$$

We then separately consider these two terms: the expectation of T, and the distribution of the 'fluctuation' of T, i.e. its concentration around the mean.

Here is an illuminating example, as a sample for the types of questions we will see throughout the course.

Date: Fall 2024. *Instructor*: Prof. Subhabrata Sen.

Example 1.4. Let $G = (G_{ij})$ be an $N \times N$ symmetric random matrix. A standard example is to set $G_{ii} \overset{iid}{\sim} \mathcal{N}(0, 2)$, and $G_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1), i < j$.

A classical question is then to establish bounds on the eigenvalues, in particular the *largest* eigenvalue [cf. PCA, etc., which we care about in application]. We can set $T = \lambda_{\max}(G)$. In connection with the previous example, T is a function $f(G_{ij})$, but is not a function with an analytic expression.

Fact. The value $T = \lambda_{\max}(G)$ is roughly of order $\sqrt{N}$. In fact $\frac{T}{\sqrt{N}} \approx 2$. Formally, there exists $\epsilon_N \to 0$ such that

$$\mathbb{P}\left[\left|\frac{\lambda_{\max}(G)}{\sqrt{N}} - 2\right| > \epsilon_N\right] \to 0.$$

Our goal, then, is to be more quantitative about these nice 'limit' statements. We care about such bounds because in application, we would like actual estimates on, say, how close the value $\lambda_{\max}(G)$ is to 2, given the size of some RM we are working with. In practice, much more is known about this scenario, but this is a taste of the ideas and questions asked in the course.

1.4. **Characteristic Phenomena in High-Dim Prob.** Here are some phenomena which will occur frequently in HDP:

(1) Concentration of Measure. For instance, see Example 1.4
(2) Phase transitions / Threshold phenomena. e.g., the behavior of some functions will change drastically at certain points of the domain.
(3) Universality.

Example 1.4 illustrates concentration of measure. We give examples illustrating the other phenomena below.

Example 1.5 (Percolation on a complete graph, cf. phase transitions)**.** Take the complete graph K_n. Perform edge percolation: for each edge, keep the edge with probability $p = \frac{c}{n}$. This yields an Erdos-Renyi random graph. Denote $\mathcal{C}_1, \mathcal{C}_2, \ldots$ to be the connected components of the graphs, ordered in descending order of size. We can ask, **what is the size of $\mathcal{C}_1$?**

Fact. If $c < 1$, $|\mathcal{C}_1| = \mathcal{O}_p(\log n)$. However, if $c > 1$, then $|\mathcal{C}_1| = \mathcal{O}_p(n)$. Moreover, there is a constant $\psi(c)$ such that

$$\frac{|\mathcal{C}_1|}{n} \xrightarrow{\mathbb{P}} \psi(c).$$

At $c = 1$, $|\mathcal{C}_1| = \mathcal{O}_p(n^{\frac{2}{3}})$.

Thus c is a sort of critical value. In this present case, when $c < 1$, the number of edges is sufficiently small such that one expects small tree-like structures, whereas when $c > 1$ there emerges a 'giant' component.

Example 1.6. Take a symmetric RM, where $G_{ii} \overset{iid}{\sim} \mathcal{F}_1$, $G_{ij} \overset{iid}{\sim} \mathcal{F}_2$ are all independent. Here $\mathcal{F}_1, \mathcal{F}_2$ are distributions satisfying $\mathbb{E}[\mathcal{F}_1] = 0, \mathbb{E}\mathcal{F}_1^2 = 2$, $\mathbb{E}[\mathcal{F}_2] = 0, \mathbb{E}\mathcal{F}_2^2 = 1$. This is a generalized version of the setup in Example 1.4.

Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of G. We are quite interested in the empirical distribution μ_n of the eigenvalues, which is just a point mass at all of the normalized eigenvalues:

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

Note that this is a *random* measure, because the matrix G is random. This is a historic question from the work of Wigner, who was actually looking at physics systems, which with sufficiently many elements can be modeled via a random system. We know many interesting facts about the limiting properties of the distribution of μ_n, beginning with the Wigner Semicircle Law.

Fact (Wigner Semicircle Law)**.** We have

$$d_L(\mu_n, \mu_{sc}) \xrightarrow{\mathbb{P}} 0$$

provided that $\mathbb{E}|G_{ij}|^{4+\delta} < \infty$.

In particular, this law holds for non-normal entries. In general, we look at a well-behaved quantity for a specific case (e.g. normal RVs), and generalize to a universal class of RVs/RMs.

1.5. **Motivation in Application.** Let us argue for the importance of these concepts in practice, in particular in research in statistics and ML. In the high-dimensional setting, many 'obvious' facts in statistics that we take for granted are not so obvious. For instance, in the HD setting, many classical estimators, etc., are no longer effective; e.g., even the sample *mean* is not as obviously good!

Example 1.7 (multi-variate statistics)**.** Take samples $X_i \overset{iid}{\sim} \mathcal{N}(0, \mathbb{I}_p)$. We would like to estimate the sample covariance.

Traditionally, we just take the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\intercal,$$

hwere we have proper convergence guarantees $\hat{\Sigma} \to \mathbb{I}_p$ (in probability, here in the sense of operator norms), for *fixed* p and $n \to \infty$.

In modern applications, however, we have a nontrivially large dimension; we expect to have a large dataset, but the *ratio* is closer to a constant; i.e. $\frac{p}{N} \to \gamma \in (0, 1)$. We must replace $X_i \overset{iid}{\sim} \mathcal{N}(0, \frac{\mathbb{I}_p}{N})$ for normalization reasons. However, even then, suppose we take the same sample covariance matrix $\hat{\Sigma}$. In this case, the eigenvalue distribution converges to this 'Marchenko-Pastur Law'. We may discuss this more; however, the point is that in the limit as $N \to \infty$, the concentration of the eigenvalues is not at 1 (which would happen in the case of low-dimension).

Intuitively, with high dimension, we have much more noise in the system, and we may fail to have a proper concentration result. In analysis of data, we need to take into account this extra 'noise'.

2. Lecture 2: September 9

2.1. **Agenda.** We will finish discussing phase transitions and universality in statistical learning. We will then talk about the Efron-Stein inequality and its applications.

2.2. **Examples; Cont'd.** We provide some more examples, from the ideas discussed in the previous class.

Example 2.1 (PCA and phase transitions)**.** Let $M = \lambda vv^\intercal W$, where $W_{ii} \overset{iid}{\sim} \mathcal{N}(0, \frac{2}{n})$, $W_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$. This resembles the type of question in fundamental PCA: we have a matrix M, and we would like to find the projection direction v. A naive approach is to use PCA and take $\hat{v}$, the top eigenvector of M. The natural question is: **what is the performance of $\hat{v}$?**

As a technical note, we typically assume $v \sim \text{Unif}(S^{n-1})$ and $\lambda \sim \mathcal{O}(1)$. Intuitively, v is a direction vector and λ is the strength of the signal. This prevents some scaling/nonuniqueness issues. The following is known:

Fact. We have that:

$$\begin{cases} \lambda < 1 & |\langle \hat{v}, v \rangle| = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) \\ \lambda > 1 & |\langle \hat{v}, v \rangle| = \mathcal{O}_p(1) \end{cases}$$

Intuitively, suppose v is close to the 'north pole'; when the signal is small, $\hat{v}$ is close to the 'equator', and when the signal is large, $\hat{v}$ is distributed along a cone.

Remark 2.2. In this problem, the accuracy of $\lambda < 1$ is a 'true' information-theoretic limit as well.

Example 2.3 (LASSO regression and universality)**.** Consider a standard supervised learning setup, where our data is

$$\{(y_i, x_i) : 1 \le i \le n\}, y_i \in \mathbb{R}, x_i \in \mathbb{R}^p.$$

Fit a standard LASSO model:

$$\hat{\beta} = \text{argmin} \left\{ \frac{1}{2} ||y - X\beta||^2 + \lambda ||\beta||_1 \right\}.$$

One standard approach towards analyzing this model is to assume some reasonable model:

$$x_i \overset{iid}{\sim} \mathcal{N}(0, \frac{I_p}{n}), y_i = \langle x_i, \beta_0 \rangle + \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

However, in practice any guarantees on this setup is not too useful, because in practice the Gaussian assumption on data is problematic.

However, recent findings have suggested that the Gaussian assumption is not necessary. As long as we have iid and conditions on the first and second moments ($\mathbb{E}x_{ij} = 0, \mathbb{E}x_{ij}^2 = \frac{1}{n}$), the only necessary constraint is some sort of 'light tail' condition (formally this is the subgaussian condition). Intuitively, the Gaussian condition is very easy to work with, but we have this 'universality' concept such that many statements which hold for Gaussian RVs/RMs generalize naturally to a wider class of RVs/RMs.

Remark 2.4. The universality heuristic depends on distribution / functional form. As intuition, in the CLT the important condition is that the function at hand is not strongly dependent on any specific variable.

2.3. **Concentration Inequalities.** We will now begin discussing concentration inequalities. The standard reference is the text [BLM13].

In general, let $Z = f(X_1, \ldots, X_n)$ for independent $X_1, \ldots, X_n$. We want to get some expression of the type $\mathbb{P}[|Z - \mathbb{E}Z| > \epsilon_n] \le \mathcal{O}(\text{expr})$ (for some expression expr, typically e.g. converging to 0, etc.). The point is to establish formally the concept that Z is closely concentrated about its mean.

A first idea is to control the variance of Z. Then we can use methods such as Markov/Chebyshev to control this probability. To do this, we have the nice *Efron-Stein Inequality*:

Theorem 2.5 (Efron-Stein Inequality)**.** Let X be a measurable space, $f : \mathcal{X}^n \to \mathbb{R}$, $X_1, \dots, X_n$ are independent $\mathcal{X}$-valued RVs. Let $Z = f(X_1, \dots, X_n)$. Assume that $\mathbb{E}Z^2 < \infty$. Then

$$\begin{aligned}\mathrm{Var}(Z) &\le \sum_{i=1}^n \mathbb{E}\left[(Z - \mathbb{E}[Z \mid X_{-i}])^2\right] \\ &= \sum_{i=1}^n \mathbb{E}\left[\mathrm{Var}(Z \mid X_{-i})\right] =: v.\end{aligned}$$

If $X' = (X_1', \dots, X_n')$ is an iid copy of X, let $Z_i' = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$. Then

$$v = \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[(Z - Z_i')^2\right] = \sum_{i=1}^n \mathbb{E}\left[(Z - Z_i')_+^2\right] = \sum_{i=1}^n \mathbb{E}\left[(Z - Z_i')_-^2\right].$$

Note that X_{-i} denotes the collection $\{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n\}$.

Remark 2.6. Here is a general heuristic about concentration inequalities. Usually, we care both about the generality and efficacy of new inequalities. In particular, the result of Efron-Stein is quite general, so we should expect it to be powerful. If it were

Remark 2.7 (Efron-Stein and McDiarmid)**.** Note that unlike McDiarmid (which gives an Azuma-type exponential tail bound), Efron-Stein does not require worst case control. It only requires some sort of 'average' control.

Proof of Theorem 2.5. We will try to use a martingale difference (MD) decomposition, each term of which is easy to bound, and any two terms of which are uncorrelated. First, write

$$\mathbb{E}_i(Z) = \mathbb{E}(Z|X_1, ...X_i), \mathbb{E}_n Z = Z, \mathbb{E}_0 Z = \mathbb{E}Z.$$

We see that $E(\mathbb{E}_i(Z) - \mathbb{E}_{i-1}(Z)) = Z - Z = 0$. Moreover, we have by a telescoping sum argument that

$$Z - \mathbb{E}Z = \sum_{i=1}^n (\mathbb{E}_i Z - \mathbb{E}_{i-1} Z) := \sum \Delta_i,$$

from which we find that

$$\mathrm{Var}(Z) = \mathbb{E}\left[\left(\sum \Delta_i\right)^2\right] = \sum_{i=1}^n \mathbb{E}(\Delta_i^2) + 2\sum_{j>i} \mathbb{E}(\Delta_i \Delta_j) = \sum_{i=1}^n \mathbb{E}(\Delta_i^2) + 0.$$

The last equality follows from the fact that the cross terms reduce to zero in expectation; indeed, suppose $j > i$; then

$$\mathbb{E}(\Delta_i \Delta_j) = \mathbb{E}(\mathbb{E}(\Delta_i \Delta_j \mid X_1, \dots, X_i)) = \mathbb{E}(\Delta_i E(\Delta_j \mid X_1, \dots, X_i)) = 0,$$

where the last equality follows from the fact that

$$E(\Delta_j \mid X_1, \dots, X_i) = E(Z \mid X_1, \dots, X_i) - E(Z \mid X_1, \dots, X_i)$$

(cf., smaller sigma algebra wins). Note that thus far, we have not used independence; however, this condition will become critical in what follows. Now, we make the following key observation:

$$\mathbb{E}_i[\mathbb{E}[Z|X_{-i}]] = \mathbb{E}[\mathbb{E}[Z|X_{-i}]|X_1, \dots, X_i] = \mathbb{E}_{i-1}[Z],$$

hence

$$\Delta_i = \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z] = \mathbb{E}_i[Z - \mathbb{E}[Z|X_{-i}]].$$

Now by Jensen's inequality, we have

$$\mathbb{E}[\Delta_i^2] = \mathrm{Var}(Z) \leq \mathbb{E}[(Z - \mathbb{E}[Z|X_{-i}])^2] = \sum \mathbb{E}[\mathrm{Var}(Z|X_{-i})] = \frac{1}{2}\sum \mathbb{E}[(Z - Z_i')^2].$$

The last equality follows because we know that $\mathrm{Var}(Z \mid X_{-i}) =$

This gives us the first form of Efron-Stein. Now m Note we used the fact that for X, Y i.i.d; $Var(X) = 1/2E[(X-Y)^2]$

□

2.4. **History and Applications.** This inequality actually arose from an applied statistical context. Initially, B. Efron and C. Stein were investigating the *jackknife*. The setup is as follows: let $X_i \overset{iid}{\sim} \mathcal{F}$, and our goal is to estimate $\mathcal{O}(\mathcal{F})$ (here $\mathcal{O}$ denotes the general space of facts about $\mathcal{F}$). Usually, we are given some general statistic, $Z = f(X_1, \ldots, X_n)$. We are interested in the bias and variance.

The *jackknife* estimate for the variance is $\sum_{i=1}^n (Z - Z_i)^2$, where Z_i is the estimate of θ based on X_{-i}. Intuitively, the difference $(Z - Z_i)^2$ should be related to the variance contribution of X_i. The implication of Theorem 2.5 is then exactly that this jackknife estimate is biased positively!

We now provide some applications.

2.4.1. *Functions with bounded differences.* We define the following functional property.

Definition 2.8 (Bounded Difference Property)**.** Let $f : \mathcal{X}^n \to \mathbb{R}$. We say that f has the *bounded difference property* with $c_1, \ldots, c_n \geq 0$ if

$$\sup_{x_{-i} \in \mathcal{X}^{n-1}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i.$$

Efron-Stein immediately gives the following.

Corollary 2.9. If f has the bounded difference property, then

$$\mathrm{Var}(Z) \leq \frac{1}{4}\sum_{i=1}^n c_i^2.$$

Proof. From Efron-Stein,

$$\mathrm{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}\left[\mathrm{Var}(Z \mid X_{-i})\right].$$

Each individual term essentially a bounded function of X_i which is within an interval of length c_i, so the variance is at most $\frac{1}{4}c_i^2$. Summing these gives the result. □

Intuitively, if we have some worst-case control, then the Efron-Stein average case control will certainly exist.

Example 2.10 (Bin-Packing Problem)**.** We have $X_1, \ldots, X_n \in [0, 1]$ independent, and we would like to 'pack' them in boxes where each box has size ≤ 1. Let Z be the minimum number of boxes required.

If one swaps X_i with another variable, the number of boxes can change by at most 1 (either we can save one box or need to use one more box). Then Corollary 2.9 above implies $\mathrm{Var}(Z) \le \frac{n}{4}$. This bound is tight if $X_i \sim \mathrm{Bern}(\frac{1}{2})$.

2.4.2. *Largest Eigenvalue of Random Symmetric Matrices.* Let $A = (X_{ij})$ be an $n \times n$ random symmetric matrix, where $X_{ij} \in [-1, 1]$. Let

$$Z = \lambda_{\max}(A) = \sup_{||u||=1} u^\intercal A u = v^\intercal A v$$

be the largest eigenvalue. Let A'_{ij} be the result of replacing X_{ij} by X'_{ij}, and $Z'_{ij} = \lambda_{\max}(A'_{ij})$. Then Efron-Stein implies that

$$\mathrm{Var}(Z) \le \sum_{i \le j} \mathbb{E}\left[(Z - Z'_{ij})^2\right].$$

The individual terms are easy to bound: we know as $Z'_{ij} \ge v^\intercal A v$ that

$$Z - Z'_{ij} \le v^\intercal (A - A'_{ij}) v \cdot \mathbb{I}(Z > Z'_{ij}) \le 2 v_i v_j (X_{ij} - X'_{ij})_+ \le 4|v_i v_j|.$$

It follows that

$$\mathbb{E}\sum (Z - Z'_{ij})_+^2 \le \sum 4|v_i v_j|^2 \le 16 \left(\sum_{i=1}^n v_i^2\right)^2 = 16.$$

This bound is actually of the right order! If a matrix is the adjacency matrix of an Erdos-Renyi graph, then the largest eigenvalue is asymptotically normal with $\mathcal{O}(1)$ fluctuation.

Remark 2.11. Here is an interesting related monograph reference: *Superconcentration and Related Topics*, S. Chatterjee.

3. Lecture 3: September 11

First Efron-Stein was proven; for clarity, we have included the material alongside the statement, which was presented in Lecture 2. We now look at the Gaussian Poincare Inequality, which is in a sense a 'continuous version' of Efron-Stein.

3.1. Gaussian Poincare Inequality.

Theorem 3.1 (Gaussian Poincare Inequality)**.** Let $X = (X_1, \ldots, X_n), X_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ be a random Gaussian vector, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Then we have the inequality

$$\mathrm{Var}(f(X)) \le \mathbb{E}[|\nabla F(X)|^2].$$

Remark 3.2. Note the intuition here is similar to that of Efron-Stein. In both cases, we are saying that the variance of the entire function can be bounded via information about the change in the function occurring due to a small perturbation. Efron-Stein is a sort of discrete version, Gaussian-Poincare is a continuous version. The latter is specifically for Gaussians, but yields an easier-to-compute upper bound, whereas Efron-Stein is harder to use.

Proof. WLOG, we may assume $\mathbb{E}[|\nabla f(X)|^2] < \infty$. First, the idea is that Efron-Stein allows us to reduce to the case $n = 1$. Indeed, suppose the statement holds for $n = 1$. Then we find that

$$\mathrm{Var}(f(X)) \le \sum_{i=1}^n \mathbb{E}\left[\mathrm{Var}\left(f(X) \mid X_{-i}\right)\right] \le \sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\partial f(X)}{\partial x_i}\right)^2\right]\right] = \mathbb{E}[|\nabla f(X)|^2],$$

where the second inequality follows noting that the conditional variance is a function of only x_i, so when applying the $n = 1$ case, the gradient $|\nabla f(X)|^2$ reduces to $\left(\frac{\partial f(X)}{\partial x_i}\right)^2$.

Now we prove the result for $n = 1$. First, note that it suffices to show the result when $f \in C_c^2$ is twice continuously differentiable with compact support.[1] We will now utilize properties of the Gaussian; in particular, we will use the CLT.[2] The idea is that by the CLT, we can approximate a Gaussian by a sum of simpler RVs; then applying Efron-Stein again and taking limits will yield the result.

In particular, let us use a sum of Rademacher random variables:

$$\epsilon_1, ...\epsilon_n \overset{iid}{\sim} \text{Rademacher} := \begin{cases} 1 & p = \frac{1}{2} \\ -1 & p = \frac{1}{2} \end{cases},$$

and also let their scaled sum be

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i.$$

By Efron-Stein, we have

$$\text{Var}(f(S_n)) \leq \sum \mathbb{E}[\text{Var}(f(S_n) \mid \epsilon_{-i}].$$

Now, conditioning on ϵ_{-i}, note that $f(S_n))$ is essentially a one-variable function of $\epsilon_i = \pm 1$. Thus, we have

$$\text{Var}(f(S_n) \mid \epsilon_{-i}) = \frac{1}{4}\left[f\left(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\right) - f\left(S_n + \frac{1+\epsilon_i}{\sqrt{n}}\right)\right]^2.$$

It follows that

$$\mathbb{E}[\text{Var}(f(S_n) \mid \epsilon_{-i}] = \frac{1}{4}\sum_{i=1}^{n} \mathbb{E}\left[\left(f\left(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\right) - f\left(S_n + \frac{1+\epsilon_i}{\sqrt{n}}\right)\right)^2\right].$$

Summing across all i, we find that

$$\text{Var}(f(S_n)) \leq \frac{1}{4}\sum_{i=1}^{n} \mathbb{E}[\text{Var}(f(S_n) \mid \epsilon_{-i}] = \frac{1}{4}\mathbb{E}\left[\left(f\left(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\right) - f\left(S_n + \frac{1+\epsilon_i}{\sqrt{n}}\right)\right)^2\right].$$

By Taylor expansion, we know that

$$\left|f\left(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\right) - f\left(S_n + \frac{1+\epsilon_i}{\sqrt{n}}\right)\right| \leq \frac{2}{\sqrt{n}}|f'(S_n)| + \frac{2}{n}\|f''\|_\infty.$$

Therefore,

$$\frac{1}{4}\left(f\left(S_n + \frac{1-\epsilon_i}{\sqrt{n}}\right) - f\left(S_n + \frac{1+\epsilon_i}{\sqrt{n}}\right)\right)^2 \leq f'(S_n)^2 + \frac{2\|f\|_\infty}{\sqrt{n}}|f'(S_n)| + \frac{\|f''\|_\infty^2}{n}.$$

Plugging this into the above inequalities yields

$$\text{Var}(f(S_n)) \leq \mathbb{E}f'(S_n)^2 + \mathbb{E}\left[\frac{2\|f\|_\infty}{\sqrt{n}}|f'(S_n)|\right] + \frac{\|f''\|_\infty^2}{n}.$$

[1] These functions are dense, hence we can approximate any f in the problem by such f_n and take $f_n \to f$ and apply standard convergence results.

[2] It is perhaps also possible to appeal to calculus methods, but it would be very ugly.

Now let $n \to \infty$; by the CLT we know that $S_n \to X$ in distribution. Applying the continuous mapping theorem, we thus find that the latter two terms on the right hand side go to 0, and therefore

$$\mathrm{Var}(f(X)) \leq \mathbb{E} f'(X)^2,$$

as desired. This completes the proof. □

3.2. **Extensions to General RVs.**

Theorem 3.3. Let $X = (X_1, \ldots, X_n), X_i \in [0,1]$ be independent bounded random variables and $f : \mathbb{R}^n \to \mathbb{R}$ be a separately convex function (i.e. convex in each variable). If $f \in C^1$ then

$$\mathrm{Var}(f(X)) \leq \mathbb{E}[||\nabla f(X)||^2].$$

Proof. By Efron-Stein, we know that

$$\mathrm{Var}(f(X)) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}(f(X) \mid X_{-i}\right].$$

Set $Z = f(X)$. If we condition on all but the $i-$th variable, we essentially have a function of one variable. Recall that $\mathrm{Var}(X) \leq \mathbb{E}\left[(X-c)^2\right]$ for any c, with equality at $c = \mathbb{E}U$. Using the conditional form of this, we have that

$$\mathbb{E}\left[\mathrm{Var}(f(X) \mid X_{-i}\right] \leq \sum \mathbb{E}\left[\mathbb{E}\left[(Z - Z_i)^2 \mid X_{-i}\right]\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[(f(X) - f(\overline{X_i}))^2\right].$$

Now, choose $Z_i' = \inf_{x_i'} f(X_{-i}, x_i')$. In a compact setting $[0,1]^n$, the infimum is attained as a mininum. Set

$$X_i' = \arg\min_{x_i'} f(X_{-i}, x_i'), \qquad \overline{X_i} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n).$$

Now we need the following observation: in a convex function of one variable, if (x_0, y_0) is the argmin, from any other point, drawing a tangent line to the curve will intersect $x = x_0$ at (x_0, y_0') for $y_0' < y_0$. This will subsequently justify the choice of X_i' as a minimizer.

In particular, we need to bound $f(X) - f(\overline{X_i})$. First by definition of X_i as minimizer, $f(X) - f(\overline{X}_i) \geq 0$. Moreover, by the above observation, we know that

$$f(X) - f(\overline{X}_i) \leq \frac{\partial f(x)}{\partial x_i} \left|X - X_i\right|.$$

We therefore establish that

$$\begin{aligned}
\sum \mathbb{E}\left[(f(X) - f(\overline{X_i}))^2\right] &\leq \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{\partial f(x)}{\partial x_i}\right)^2 (X_i - X_i')^2\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{\partial f(x)}{\partial x_i}\right)^2\right] \\
&= \mathbb{E}\left[||\nabla f(x)||^2\right],
\end{aligned}$$

where the second inequality follows from $X_i, X_i' \in [0,1]$. Combining with the previously displayed inequalities, we conclude that

$$\mathrm{Var}(f(X)) \leq \mathbb{E}[||\nabla f(x)||^2],$$

as desired.

□

Remark 3.4. One can apply some version of a 'subgradient inequality' if we do not have guarantees of differentiability, etc.

‘

4. Lecture 4: September 16

4.1. **Wrapping up Theorem 3.3.** We first finished the proof of Theorem 3.3; this is in the previous section for readability. Let us now see an application of this inequality.

Example 4.1 (Operator Norm of a random matrix)**.** Let

$$A = (X_{ij}) \in \mathbb{R}^{m\times n}, X_{ij} \in [0,1] \text{ independent.}$$

Let

$$Z = \sqrt{\lambda_{\max}(A^\intercal A)} = \sup_{||u||=1} ||Au||$$

be the operator norm of A. We are interested in establishing some results on Z.

Note that $f : X \to Z$ is separately convex. Indeed, under the function, $X_{ij} \to Au$ is convex (actually linear!), the norm function is convex, and the sup of convex functions is convex; the composition $X_{ij} \to Au \to ||Au|| \to \sup ||Au||$ is therefore convex. Moreover, everything is differentiable, so directly by Theorem 3.3 we can establish $\mathrm{Var}(Z) \le \mathbb{E}\left[||\nabla f(X)||^2\right]$.

Additionally, the following statement holds.

Fact. Let $P = (P_{ij}), Q = (Q_{ij}) \in \mathbb{R}^{m\times n}$, Z_p is the operator norm of P, Z_q is the operator norm of Q. Then

$$(Z_P - Z_Q)^2 \le \sum_{ij} (P_{ij} - Q_{ij})^2.$$

For reference, see Boucheron, Lugosi, Massart. This is a purely linear algebra fact! The importance, anyway, is that $A \to Z_A$ is Lipschitz. We can thus establish a very good bound on the gradient of f! In particular, we find that $||\nabla f|| = \mathcal{O}(1)$. So the variance is bounded.

Remark 4.2. The operator norm here is also the largest singular value, so this result applies to that quantity as well.

Remark 4.3. This example also shows the nontriviality of Theorem 3.3. In particular, the existence of nontrivial implied functions of interest, for instance the operator norm with respect to the entries of a matrix, yields relevant applications of this result.

4.2. **A general inequality to look for.** At this point, we will also make the following remark. We are interested in stronger versions of the type of results we have just proven. In particular, we will at some point arrive at the following inequality.

Theorem 4.4 (Lipschitz Concentration Inequality)**.** Let $g \sim \mathcal{N}(0, \mathbf{I}_n)$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz with constant $L = ||F||_{\mathrm{Lip}}$; that is,

$$L = \inf\{\ell : |F(x) - F(y)| \le \ell |x - y|\}.$$

Then we have

$$\mathbb{P}\left[|F(g) - \mathbb{E}F(g)| \ge t\right] \le 2\exp\left(-\frac{t^2}{4||F||_{\mathrm{Lip}}^2}\right)$$

This is a type of theorem we are working towards, which somewhat accomplishes the stated goal at the beginning of this course. In particular, we can then properly separate the expectation $\mathbb{E}F(g)$ and the fluctuations, which are of much smaller order, and then analyze them separately.

In order to arrive here, however, we will need to discuss a few more methods.

4.3. **Stein's Method (Gaussian Integration by Parts).** Here is the setting of this lemma.

Lemma 4.5 (Stein's Lemma)**.** Let $Z \sim \mathcal{N}(0, \sigma^2)$. Then if $E|f'(Z)| < \infty$, we have

$$\mathbb{E}[Zf(Z)] = \sigma^2 \mathbb{E}[f'(Z)].$$

Proof. WLOG assume $\sigma^2 = 1$, otherwise it suffices to scale via $Z \leftarrow \frac{Z}{\sigma}$. The standard proof is essentially a calculus expansion via integration by parts. Here is a similar but slightly more succinct proof. Write

$$\mathbb{E}[f'(Z)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(\zeta) e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(\zeta) e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta.$$

We can work with both parts separately. For the first,

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(\zeta) e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta &= -\frac{1}{2\pi} \int_{-\infty}^{0} \int_{-\infty}^{\zeta} f'(\zeta) x e^{-\frac{x^2}{2}} \mathrm{d}x \mathrm{d}y \\
&= -\frac{1}{2\pi} \int_{-\infty}^{0} \int_{x}^{0} f'(\zeta) x e^{-\frac{x^2}{2}} \mathrm{d}y \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} (f(x) - f(0)) x e^{-\frac{x^2}{2}} \mathrm{d}x
\end{aligned}$$

where the second-to-last expression is by Fubini (the conditions are satisfied in that $f'(Z)$ is absolutely integrable). The exact same argument yields

$$\frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(\zeta) e^{-\frac{\zeta^2}{2}} \mathrm{d}\zeta = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} (f(x) - f(0)) x e^{-\frac{x^2}{2}} \mathrm{d}x.$$

Combining the inequalities, we find

$$\mathbb{E}[f'(Z)] = \mathbb{E}[Z(f(Z) - f(0))] = \mathbb{E}[Zf(Z)],$$

as desired. □

Remark 4.6. The usual proof is via an integration by parts, but the first line of the expansion, replacing $e^{-\frac{\zeta^2}{2}}$ with $\int_{-\infty}^{\zeta} x e^{-\frac{x^2}{2}} \mathrm{d}x$, is essentially the same idea.

We can extend this to the multivariate setting.

Lemma 4.7 (Multivariate Stein's Lemma)**.** Let $g = (g_\ell)_{\ell=1}^n \sim \mathcal{N}(0, \Sigma)$ be Multivariate Gaussian, and $F : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Set $F_\ell = \frac{\partial F}{\partial x_\ell}$ to be the partial derivatives of F with respect to its arguments. Then

$$\mathbb{E}[g_1 F(g)] = \sum_{\ell=1}^{n} \mathbb{E}[g_1 g_\ell] \mathbb{E}[F_\ell(g)]$$

if for all $\ell \leq n$, either $\mathbb{E}[|F_\ell(g)|] < \infty$ or $\mathbb{E}[g_1 g_2] = 0$.

Proof. Intuitively, if the expressions were solely a function of g_1, then we can easily establish this type of result. We will therefore try to regress the other variables over g_1. We will now be more formal.

Let $\sigma^2 = \mathbb{E}g_1^2$. Obviously if $\sigma^2 = 0$ we are done because both sides clearly equal 0. Otherwise, let

$$g'_\ell = g_\ell - \lambda_\ell g_1, \quad \lambda_\ell = \frac{\mathbb{E}[g_1 g_\ell]}{\sigma^2},$$

which is the standard regression β from g_1 to g_ℓ. In particular, g'_ℓ is independent of g_1 for all ℓ. Letting $\lambda = (\lambda_\ell)_{\ell=1}^n$, we can write $g = \lambda g_1 + g'$. Now condition on g', which is independent of g_1. Then

$$\mathbb{E}_{g_1}[g_1 F(g)] = \sigma^2 \mathbb{E}_{g_1}[g_1 F(\lambda g_1 + g')] = \mathbb{E}_{g_1}\left[\sum_{\ell=1}^n \lambda_\ell F_\ell(g)\right],$$

where the second line follows from the single variable Stein's Lemma proven before, along with the chain rule. Taking the expectation of both sides, we find

$$\mathbb{E}[g_1 F(g)] = \sum_{\ell=1}^n \mathbb{E}[g_1 g_\ell]\mathbb{E}[F_\ell(g)],$$

as desired. □

Why do we care about these results? Recall the general Chernoff-type bound:

$$\mathbb{P}\left[F(g) \geq \mathbb{E}F(g) + t\right] \leq \mathbb{E}\left[e^{\lambda(F(g)-\mathbb{E}F(g))}\right] e^{-\lambda t}.$$

The equalities we have just shown will help us get a tighter bound on the MGF expression $\mathbb{E}\left[e^{\lambda(F(g)-\mathbb{E}F(g))}\right]$.

5. Lecture 5: September 18

We continue to adapt our tools now. We begin with Gaussian Interpolation, an important tool used in analyzing Gaussian RVs and RMs.

5.1. Gaussian Interpolation.

The setup is as follows. Let X, Y be two mean zero Gaussian RVs in $\mathbb{R}^n$, and

$$a_{ij} = \mathbb{E}X_i X_j, \quad b_{ij} = \mathbb{E}Y_i Y_j.$$

Let

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Y \sim \mathcal{N}(0, t\mathbf{a} + (1-t)\mathbf{b})$$

be a family of Gaussians ranging from $Z(0) = Y$ to $Z(1) = X$. Let $F(t) \in C^\infty$ be a smooth function, and set

$$f(t) = \mathbb{E}F(Z(t)) = \mathbb{E}[F(\sqrt{t}X + \sqrt{1-t}Y)].$$

We then have the following result.

Lemma 5.1 (Gaussian Interpolation Lemma)**.** Suppose that there exists $c_1, c_2 > 0$ such that $\left|\frac{\partial F}{\partial x_i}\right| \leq c_1 e^{c_2\|x\|}$, and either $\left|\frac{\partial F}{\partial x_i \partial x_j}\right| \leq c_1 e^{c_2\|x\|}$ or $a_{ij} = b_{ij}$. Then

$$f'(t) = \frac{1}{2}\sum_{i,j\leq n}(a_{ij} - b_{ij})\mathbb{E}\frac{\partial^2 F}{\partial x_i \partial x_j}Z(t).$$

Remark 5.2. Intuitively, we want to understand the behavior of function $f(Y)$ when Y is some nontrivially complicated Gaussian vector. In order to do this, we can try to compare $\mathbb{E}[f(Y)]$ to $\mathbb{E}[F(X)]$, for a suitable choice of X. In order to establish this formally, we need some way to estimate how close these quantities are. It suffices, therefore, to have estimates of the derivative $f'(t)$, which are easy to compute via the above result. We then just need to look at the behavior change in $f(t)$ when t ranges from 0 to 1.

Proof. The point is to differentiate under the integral sign. We will not provide a rigorous justification of this, but in general, it will follow from Leibniz's rule; note that the exponential bound on the derivatives ensures absolute integrability of the partial derivatives, so the conditions are met.

By performing differentiation under the integral sign,

$$f'(t) = \mathbb{E}\left[\frac{d}{\mathrm{d}t}F(Z(t))\right] = \sum_{i\le n}\mathbb{E}\left[\frac{\partial F}{\partial x_i}\left(Z(t)\right)Z_i'(t)\right].$$

Now we recognize the form of the Gaussian Integration by Parts method. In particular, from Theorem 4.7, we have that

$$\mathbb{E}\left[\frac{\partial F}{\partial x_i}\left(Z(t)\right)Z_i'(t)\right] = \sum_{j=1}^{n}\mathbb{E}\left[Z_i'(t)Z_j(t)\right]\mathbb{E}\left[\frac{\partial^2 F}{\partial x_i\partial x_j}\left(Z(t)\right)\right].$$

Now, note that direct computation yields

$$\mathbb{E}[Z_i'(t)Z_j(t)] = \mathbb{E}\left[\left(\frac{1}{2\sqrt{t}}X_i - \frac{1}{2\sqrt{1-t}}Y_i\right)\left(\sqrt{t}X_i + \sqrt{1-t}Y_i\right)\right] = \frac{1}{2}(a_{ij}-b_{ij}).$$

Combining these results, we find precisely that

$$f'(t) = \frac{1}{2}\sum_{i,j\le n}(a_{ij}-b_{ij})\mathbb{E}\frac{\partial^2 F}{\partial x_i\partial x_j}Z(t),$$

as desired. □

This method of smoothly varying Gaussians turns out to be a very fundamental result. In particular, it allows us to prove general results on Gaussians. For instance, we can now prove the aforementioned Theorem 4.4! We will see this now.

5.2. **Proof of Theorem 4.4.** As intuition, by the Chernoff method,

$$\mathbb{P}[F(g) - \mathbb{E}F(g) > t] \le \mathbb{E}\left[e^{\lambda(F(g)-\mathbb{E}F(g))}\right]e^{-\lambda t}.$$

So it suffices to bound the MGF expression, which will be done via Gaussian interpolation.

Proof of Theorem 4.4. First, assume that F is differentiable and $\|\nabla F\| \le L$. We will see how to generalize this later (intuitively, we can approximate Lipschitz functions with such F). Fix $\lambda > 0$, and define for $t \in [0,1]$ the function

$$f(t) = \mathbb{E}\exp\left[\lambda(F(\sqrt{t}g^{(1)} + \sqrt{1-t}g) - F(\sqrt{t}g^{(2)} + \sqrt{1-t}g))\right],$$

where $g^{(1)}, g^{(2)}$ are independent copies of g. Now, define a function $G : \mathbb{R}^{2n} \to \mathbb{R}$, with

$$G(x_1,\dots,x_{2n}) = \exp\left[\lambda(F(x_1,\dots,x_n) - F(x_{n+1},\dots,x_{2n}))\right].$$

Then we can write

$$f(t) = \mathbb{E}\left[G(\sqrt{t}X + \sqrt{1-t}Y)\right],$$

for $X = \begin{pmatrix} g^{(1)} \\ g^{(2)} \end{pmatrix}$, $Y = \begin{pmatrix} g \\ g \end{pmatrix}$.

Note that $X \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \overbrace{\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}}^{A}\right)$, $Y \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \overbrace{\begin{pmatrix} I & I \\ I & I \end{pmatrix}}^{B}\right)$.

We will apply the previous lemma. Using the same notation, note that

$$a_{ij} - b_{ij} = \begin{cases} -1 & 1 \le i \le n, j = i+n \text{ or } 1 \le j \le n, i = j+n \\ 0 & \text{otherwise} \end{cases}.$$

Moreover, we can compute the partial derivatives

$$\frac{\partial G}{\partial x_i \partial x_{i+n}} = -\lambda^2 G \cdot F_i(x_1 \ldots, x_n) \cdot F_i(x_{n+1}, \ldots, x_{2n}).$$

Applying the Gaussian Interpolation Lemma, we find after simplification via the previous formulae that

$$f'(t) = \lambda^2 \mathbb{E}\left[G(\sqrt{t}X + \sqrt{1-t}Y) \cdot \left(\sum_{i=1}^{n} F_i(\sqrt{t}g^{(1)} + \sqrt{1-t}g) F_i(\sqrt{t}g^{(2)} + \sqrt{1-t}g)\right)\right]$$

Now we will use the Lipschitz property of the function. By Cauchy-Schwarz, note that the summation in the above equation is bounded by

$$\sum_{i=1}^{n} F_i(\sqrt{t}g^{(1)} + \sqrt{1-t}g) F_i(\sqrt{t}g^{(2)} + \sqrt{1-t}g) \le \sup_x \|\nabla F(x)\|^2 \le L^2$$

(the LHS is upper bounded by a product of two sum-of-squares expressions, both of which are the the magnitude of the gradient of F at separate points).

Applying this, we immediately get the magical simplification

$$f'(t) \le \lambda^2 L^2 f(t),$$

as the summation is bounded by L^2 and the expectation after taking out the L^2 term is just $\mathbb{E}G(\sqrt{t}X + \sqrt{1-t}Y) = f(t)$ exactly!

Finally, we just need some basic manipulations to turn this bound on f' into the desired form of the inequality. Consider the inequality

$$\frac{d}{\mathrm{d}t}\overbrace{\left(f(t)e^{-\lambda^2 L^2 t}\right)}^{h(t)} = e^{-\lambda^2 L^2 t}\left(f'(t) - \lambda^2 L^2 f(t)\right) \le 0.$$

The above implies $h' \le 0$, so $h(1) \le h(0)$. But we easily find $h(0) = f(0)$ and $h(1) = f(1)e^{-\lambda^2 L^2}$. We conclude that

$$f(1) \le f(0)e^{\lambda^2 L^2}.$$

However, we also have that $f(0) = 1$ clearly and $f(1) = \mathbb{E}e^{\lambda(F(g^{(1)}) - F(g^{(2)}))}$. So we obtain that

$$\mathbb{E}e^{\lambda(F(g^{(1)}) - F(g^{(2)})} \le e^{\lambda^2 L^2}.$$

This looks almost exactly like the MGF term now. To finish, note that e^{-x} is convex; by Jensen we know that

$$e^{\lambda\left[F(g^{(1)})-\mathbb{E}F(g^{(2)})\right]} \leq e^{\lambda\left[F(g^{(1)})-F(g^{(2)})\right]}.$$

Taking expectations with respect to $g^{(1)}$ yields

$$Ee^{\lambda\left[F(g^{(1)})-\mathbb{E}F(g)\right]} \leq e^{\lambda^2 L^2}.$$

Going back to the initial motivating bound from the Chernoff method, we conclude that

$$\mathbb{P}\left[F(g) - \mathbb{E}F(g) > t\right] \leq e^{\lambda^2 L^2 - \lambda t}.$$

Setting $\lambda = -\frac{t}{2L^2}$ (which optimizes the RHS) gives precisely the theorem statement, completing the proof.

Addendum. Note that we have assumed above that F has bounded gradient and is smooth. We will now show how to generalize to the case where F is an arbitrary Lipschitz function. The main idea is to 'smooth' F via convolution with a small Gaussian! In particular, it turns out that considering the smooth family $F_\epsilon(x) = \mathbb{E}_g F(x+\epsilon g)$ is sufficient.

We need only the following claim:

Claim. The following hold:

(1) $|F_\epsilon(x) - F_\epsilon(y)| \leq \mathbb{E}_g|F(x+\epsilon g) - F(y+\epsilon g)| \leq L|x-y|$. (So still Lipschitz).
(2) $F_\epsilon(x) = \frac{1}{(\epsilon\sqrt{2\pi})^n}\int_{\mathbb{R}^n} F(y)e^{-\frac{1}{2\epsilon^2}\|y-x\|^2}\mathrm{d}y$ (note differentiability is sufficient)
(3) $\|F_\epsilon - F\|_\infty \leq L\epsilon\mathbb{E}\|g\| \to 0$ a.s. as $\epsilon \to 0$. We will finish this next time.

Proof. (1) Since F is $L-$Lipschitz, we can write

$$\begin{aligned}|F_\epsilon(x) - F_\epsilon(Y)| &= |\mathbb{E}[F(x+\epsilon g)] - \mathbb{E}[F(y+\epsilon g)]| \\ &\leq \mathbb{E}|F(x+\epsilon g) - F(y+\epsilon g)| \\ &\leq L\|x-y\|,\end{aligned}$$

where the second-to-last inequality follows from Jensen on $|\cdot|$ and the last follows from $L-$Lipschitzness.

(2) It is clear that

$$F_\epsilon(x) = \frac{1}{(\epsilon\sqrt{2\pi})^n}\int_{\mathbb{R}^n} F(y)e^{-\frac{1}{2\epsilon^2}\|y-x\|^2}\mathrm{d}y.$$

One can differentiate under the integral sign due to the absolute integrability of F, so it follows that $F_\epsilon \in C^\infty$.

(3) Note that

$$\begin{aligned}|F_\epsilon(x) - F(x)| &= |\mathbb{E}[F(x+\epsilon g) - F(x)]| \\ &\leq \mathbb{E}|F(x+\epsilon g) - F(x)| \\ &\leq L\epsilon\mathbb{E}\|g\| \overset{\epsilon\to 0}{\longrightarrow} 0\end{aligned}$$

This completes the proof of the claim.

Lastly, we can apply the theorem for F_ϵ and take $\epsilon \to 0$; it suffices to then apply DCT and the theorem is proven. □

6. Lecture 6: September 23

First Theorem 4.4 was proven for general Lipschitz functions (not necessarily C^1). We now discuss applications of this result.

6.1. **Norms of Gaussian Vectors.** We can now properly analyze the fluctuations of Gaussian vectors.

Claim 6.1. Let $X \sim \mathcal{N}(0, \Sigma)$ be a random vector. Then the L^p norm $Z = \|X\|_p$ is nicely concentrated: that is,

$$\mathbb{P}\left[|Z - \mathbb{E}Z| > t\right] \le 2e^{-\frac{t^2}{4L^2}},$$

where

$$L = \|\Sigma^{\frac{1}{2}}\|_{\ell_2 \to \ell_p} = \sup_{\|x\|_2 = 1} \|\Sigma^{\frac{1}{2}} x\|_p$$

Proof. As usual we can let $X = \Sigma^{\frac{1}{2}} Y$ where $Y \sim \mathcal{N}(0, \mathbf{I}_n)$. Let $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|\Sigma^{\frac{1}{2}} x\|_p$. We can compute a bound for the Lipschitz norm:

$$|f(x) - f(y)| \le \|\Sigma^{\frac{1}{2}}\|_{\ell_2 \to \ell_p} \|x - y\|_2.$$

The result follows from Theorem 4.4. □

Remark 6.2. Here is an interesting special case: when $p = 2, \Sigma = I$, then $L = \|I\| = 1$. The above claim implicates (e.g. via classical facts about subgaussian tails) that the fluctuations of $\|X\|_2$ are $\mathcal{O}(1)$.

Comparing this to the asymptotic value of $\|X\|_2$, we see that

$$\mathbb{E}\|X\|_2 = \mathbb{E}\sqrt{x_1^2 + \cdots + x_n^2} \sim \sqrt{n}.$$

This is striking! It gives us the intuition that for high-dimensional Gaussians, most of the mass is within a small shell around the sphere $\|x\| = n$.

6.2. **Concentration of suprema of Gaussian processes.** We first have the following claim, which is only for a finite set of Gaussians.

Claim 6.3. Let $X \sim \mathcal{N}(0, \Sigma)$. Let $Z = \max_{1 \le i \le n} X_i$. Then

$$\mathbb{P}[|Z - \mathbb{E}Z| \ge t] \le 2e^{-\frac{t^2}{2L^2}},$$

where $L^2 = \max_{1 \le i \le n} \mathbb{E}X_i^2$.

Proof. Again let $X = \Sigma^{\frac{1}{2}} Y$ for $Y \sim \mathcal{N}(0, \mathbf{I})$. Let $F : \mathbb{R}^n \to \mathbb{R}$, $F(y) = \max_{1 \le i \le n} \langle s_i, Y \rangle$, where s_i are the rows of $\Sigma^{\frac{1}{2}}$. We can check the Lipschitz property: via Cauchy-Schwarz we find

$$|F(y_1) - F(y_2)| \le \left(\max_{1 \le i \le n} \|S_i\|_2 \right) \|y_1 - y_2\|_2.$$

So we can set $L^2 = \max_{1 \le i \le n} \|S_i\|_2^2 = \max_{1 \le i \le n} \mathbb{E}[X_i^2]$. □

Remark 6.4. A few points are in store. First, we do not get any results on the expectation $\mathbb{E}Z$ itself. Later in the course, we will analyze $\mathbb{E}Z$ itself.

Secondly, look at $X_i \overset{iid}{\sim} \mathcal{N}(0, 1)$. Then it is well-known that $\mathbb{E}Z = \sqrt{2 \log n}$. So the bound is already nontrivially of lower order; however, it is not optimal, because in fact one can show that $\mathrm{Var}(Z) \to 0$. However, if we take X_i to all be the same, then $\mathrm{Var}(Z) = \mathcal{O}(1)$. Intuitively, we see that the more dependence there is, the better this type of bound may be.

Lastly, the result can be extended to continuous Gaussian processes, assuming some nice analytical properties (under which we can take the limit of this finite set).

This completes our discussion of concentration of Lipschitz functions on Gaussians. These bounds are generally pretty good, and moreover dimension-independent.

6.3. **Concentration on the Uniform Spherical Distribution.** We have now spent some time discussing concentration of functions on independent random variables. Our working intuition is that if the dependence is not strong in each individual variable, then we can get some bounds based on the individual terms, cf. Efron-Stein.

We now shift to looking at a new distribution: the uniform distribution measure ν_n on the $n-$sphere S^{n-1}. The easiest way to think of this distribution is by sampling a Gaussian and then normalizing.

Fact. Let $g \sim \mathcal{N}(0, \mathbf{I})$; then $\frac{g}{\|g\|} \sim \nu_n$.

Proof Sketch. The joint PDF of this Gaussian is spherically symmetric; in particular, a sufficient statistic is $\sum g_i^2$. The result follows. □

Theorem 6.5. There exist constants $C, A > 0$ independent of n such that for any function $F : S^{n-1} \to \mathbb{R}$ which is $L-$Lipschitz in the geodesic metric on S^{n-1}, we have

$$\nu_n \left(F - \mathbb{E}_{\nu_n}[F] \geq t\right) \leq Ce^{-\frac{nt^2}{AL^2}}.$$

Remark 6.6. Of course, one can get two-sided bounds by applying the statement to $-F$. Also, this demonstrates our intuition that we can get similar results to Theorem 4.4 when the dependence on our variables is weak.

We omit the proof in this course, as it is similar to that of Theorem 4.4. A full proof is given in the reference notes [Lal].

We now move to applications of this topic.

6.4. **Application: Data Compressing and Dimension Reduction.** A classical technique in statistics and machine learning is dimension reduction, for a dataset of m points in $\mathbb{R}^n$ for relatively large m, n. Of course, one can always reduce to 0 dimensions, which is trivial but not useful: we would like to maintain some amount of information. We therefore modify our goal:

Goal: is thus to attain dimension-reduction subject to some pairwise distance constraints.
Question: How low can the projection dimension be?

In this direction, there is the celebrated Johnson-Lindenstrauss lemma.

Lemma 6.7 (Johnson-Lindenstrauss)**.** There is a universal constant $D > 0$ (independent of n) such that the following holds: Given m points $x_j \in \mathbb{R}^n$ and $\epsilon > 0$, for any $k > D\epsilon^{-2} \log m$, there exists a $k-$dimensional projection $A : \mathbb{R}^n \to \mathbb{R}^k$ such that

$$(1+\epsilon)^{-1}\|x_i - x_j\| \leq \sqrt{\frac{n}{k}}\|Ax_i - Ax_j\| \leq (1+\epsilon)\|x_i - x_j\|.$$

7. Lecture 7: September 25

We finish discussing the Johnson-Lindenstrauss Lemma, and move on to the deep and beautiful connection between concentration and isoperimetry.

7.1. **Finishing up Johnson-Lindenstrauss.** We now prove the Johnson-Lindenstrauss Lemma. We first make a few remarks regarding the result itself:

- The result intuitively states that $\{\sqrt{\frac{n}{k}}Ax_i, 1 \le i \le n\}$ is an approximate isometry,
- importantly, $k = \mathcal{O}(\log m)$ suffices; n does not matter. Therefore, our bottlenecks are in the amount of data; and
- this will lead to the start of the idea of 'sketching'.

The main idea is an application of the probabilistic method by taking a *uniformly random projection*. In particular, let $Y_1, \dots, Y_k \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_n)$, and $V = \text{span}\{Y_1, \dots, Y_k\}$. Then choose A to be the projection *to* the subspace V. This subspace is uniformly random chosen in a spherically symmetric sense: for an orthogonal matrix $\mathbf{O}$, if $V' = \mathbf{O}V$, then the distribution is the same.[3]

We then have the following intermediate lemma, at the heart of the proof of Lemma 6.7.

Lemma 7.1. Let A be the orthogonal projection of $\mathbb{R}^n$ onto a uniformly random $k-$dimensional subspace. Then for any $x \neq 0 \in \mathbb{R}^n$,

$$\mathbb{P}\left[\left|\frac{\|Ax\|}{\|x\|} - \sqrt{\frac{k}{n}}\right| \le \epsilon\sqrt{\frac{k}{n}}\right] \le \| - C' \exp\left(-B'k\epsilon^2\right),$$

where C', B' do not depend on n, k, or the matrix A.

Remark 7.2. Intuitively, $\mathbb{E}\|Ax\|^2 = \frac{k}{n}$ because when projecting randomly onto k dimensions, we should maintain on average $\frac{k}{n}$ of the variance. Then this lemma is just asserting that $\|Ax\|^2$ concentrates near its mean to fluctuations on the same order.

Proof of Lemma 7.1. Recall that we think of A as a fixed orthogonal matrix, and x as random. Then the map $F : x \to \|Ax\|$ is $1-$Lipschitz[4] in $\mathbb{R}^n$. Now, note that we have two equivalent views of this random projection:

- with A fixed and $x \sim \text{Unif}(S^{n-1})$, and
- with A a random projection to a $k-$dimensional subspace and x fixed.

The distribution of Ax is the same in both cases, by recalling the earlier remark when defining a uniformly random projection. This is an interesting perspective shift, which allows us to think of A as fixed and x following the uniform spherical distribution. But now we can immediately apply concentration results: in particular, by Theorem 6.5, we find

$$\mathbb{P}\left[|\|Az\| - \mathbb{E}\|Az\|| > t\right] \le Ce^{-Bnt^2}, \tag{1}$$

where $z \sim \text{Unif}(S^{n-1})$. This is similar to our desired result: now it suffices to establish the 'closeness' of $\mathbb{E}\|Az\|$ and $\sqrt{\frac{k}{n}}$. By spherical symmetry, we know that $\mathbb{E}\|Az\|^2 = \frac{k}{n}$ (for instance, we can take fixed orthogonal bases and 'pair together' the $k-$subsets). By Jensen, we find that $\mathbb{E}\|Az\| \le \sqrt{\frac{k}{n}}$. Now, by integrating the tail bound (1) and using the fact that

[3]The reasons we use Gaussians is for this spherical symmetry property. Intuitively, if we take random Gaussians in high dimension, we expect them to be 'almost orthogonal'. This further motivates this probabilistic approach.

[4]For instance, if A is orthogonal, its eigenvalues are all complex of magnitude 1; the operator norm in this case coincides with the spectral norm.

for nonnegative Y, $\mathbb{E}Y^2 = \int_0^\infty 2Y\mathbb{P}[Y > t]\mathrm{d}t$, we find that

$$\mathrm{Var}\left(\|Az\|\right) \leq \int_0^\infty Cte^{-Bnt^2}\mathrm{d}t \leq \frac{C}{Bn}.$$

Here B, C denote arbitrary constants; for notational simplicity we use the same letters though the constants may change inline by other fixed factors. It follows that

$$\mathbb{E}\left[\|Az\|\right] \geq \sqrt{\mathbb{E}\left[\|Az\|^2\right] - \frac{C}{Bn}} \sim \sqrt{\frac{k}{n}} + \frac{D}{\sqrt{kn}}. \tag{2}$$

Combining inequalities (1) and (2), we find that

$$\mathbb{P}\left[\left|\|Az\| - \sqrt{\frac{k}{n}}\right| > t\right] \leq C\exp\left(-Bn\left(t - \frac{D}{\sqrt{nk}}\right)^2\right)$$

for $t > \frac{D}{\sqrt{nk}}$. Setting $\epsilon\sqrt{\frac{k}{n}}$ and replacing z with x completes the proof.

□

We are now ready to prove the Johnson-Lindenstrauss Lemma.

Proof of Lemma 6.7. We can clearly ignore points which are equivalent or zero, since they do not affect the correctness of the necessary inequality, zero is easy to deal with separately. Therefore WLOG suppose X is a set of m distinct, nonzero points in $\mathbb{R}^n$. Let Y be the set of all pairwise differences: $Y = \{x_i - x_j, i < j\}$. Of course $|Y| = \binom{m}{2}$. Moreover, define $T = \sqrt{\frac{n}{k}}A$.

Intuitively, we would like to see how much $\sqrt{\frac{n}{k}}A$ distorts the distances. Thus, for $y \in Y$, say that T *distorts* y if $|\|Ty\| - \|y\|| \geq \epsilon\|y\|$. By Lemma 7.1, for each $y \in Y$,

$$\mathbb{P}[Y \text{ distorted}] \leq C'\exp\left(-B'k\epsilon^2\right).$$

Then by a union bound,

$$\begin{aligned}\mathbb{P}[\exists Y, Y \text{ distorted}] &\leq \binom{m}{2}C'\exp\left(-B'k\epsilon^2\right) \\ &= \mathcal{O}(m^2\exp(-B'D\log m)) \\ &= \mathcal{O}(m^{2-B'D}).\end{aligned}$$

Choosing D sufficiently large shows the result. □

Remark 7.3. The proof here is one example of why concentration inequalities are powerful. The point of Lemma 7.1 is to provide an explicit bound on these small possibilities of something going wrong. Then we can just apply a simple union bound and obtain a nontrivial result.

Moreover, this result takes a different perspective as the previous results we have shown. Previously, we have shown concentration *despite* the randomness in high-dimensional situations; in this case, however, we explicitly *leverage* this randomness to provide a clean probabilistic proof.

7.2. **Concentration and Isoperimetry.** Interestingly, many of the concentration-type results we have seen were motivated by a geometric and functional analysis perspective. In particular, we have been motivated by fixing a distribution $X \sim \mu$ and showing results of the form $f \sim \mathbb{E}f$ for different classes of functions $f : \mathbb{R}^n \to \mathbb{R}$. However, we can also take the perspective of fixing f and picking μ.

It turns out to be best to work with a metric space $(\mathcal{X}, d)$, and let μ be a probability distribution[5] of $(\mathcal{X}, d)$. Take some fixed 'measurable' set $A \subseteq X$. Then consider the regions $A^\epsilon = \{x \in \mathcal{X} : d(x, A) \leq \epsilon\}$.

Definition 7.4. In general, if $A \subseteq X$, we say that μ satisfies a $\sigma-$*isoperimetric inequality* if whenever $\mu(A) \geq \frac{1}{2}$, then for all $\epsilon > 0$ we have

$$\mu(A^\epsilon) \geq 1 - Ce^{-\frac{\epsilon^2}{2\sigma^2}}.$$

Interestingly, isoperimetric inequalities are closely related to the Lipschitz inequalities we have seen. In fact, there is an essential equivalence, seen by the following:

Exercise (Isoperimetric Inequality equivalent to Lipschitz Inequality)**.** If μ satisfies a $\sigma-$isoperimetric inequality, then

$$\mathbb{P}[f - \text{med}(f) \geq t]Ce^{-\frac{t^2}{2\sigma^2}}$$

for all $t \geq 0$, when f is $1-$Lipschitz. Moreover, the *converse* holds!

This problem will appear in problem set two. Some hints:

- For the first direction, look at $A = \{f \leq \text{med}(f)\}$.
- For the second direction, one can choose any 'test' function f; it turns out the function $f = d(X, A)$ as the distance function is sufficient.

For more details, see problem set two.

8. Lecture 8: September 30

In this lecture we will conclude our discussion of concentration; we then move onto the topic of Gaussian suprema and comparison inequalities. First we discussed the hints for Exercise 7.2. In light of this result, we should find important the following

Question: How does one establish isoperimetry?

Such results would carry over to nontrivial inequalities. There has been a rich literature of study in this area, with contributions of Marton, Talagrand, Bobkov-Gotze, and others. There are interesting interactions, moreover, in techniques in isoperimetry with information theory, optimal transport, convex analysis. For a reference, see Ramon Van Handel's notes for High-Dimensional Probability [H18].

8.1. **Isoperimetry in the Hamming Metric.** In our situation, we will 'cheat' a bit and look at a particular case, namely that of isoperimetry with respect to the Hamming metric.

The setup is as follows: let $X_1, \ldots, X_n \overset{iid}{\sim} \mu$, with $X_i \in [0, 1]$. Let $X = (X_1, \ldots, X_n)$ be the vector of X_is. Recall that the *Hamming metric* on $[0, 1]^n$ is defined by

$$d_H(x, y) = \sum_{i=1}^{n} \mathbb{I}(x \neq y).$$

[5] we will not define this rigorously, but one can consider a natural analogue of the Borel sigma-algebra to a general metric spaces, e.g. that defined by open balls.

The Lipschitz property with respect to the Hamming metric is familir. Indeed, $f : [0,1]^n \to \mathbb{R}$ is $1-$Lipschitz with respect to the Hamming metric d_H iff

$$\sup_{\substack{x_{-i}\in[0,1]^{n-1}\\ x_i,x_i'\in[0,1]}} |f(x_{-i},x_i) - f(x_{-i},x_i')| \le 1.$$

But this is precisely the bounded differences property! By Azuma-Hoeffding we immediately find that

$$\mathbb{P}\left[f(X) - \mathbb{E}f(X) > t\right] \le e^{-\frac{2t^2}{n}}.$$

This already resembles the form of Exercise 7.2, with the median replaced by expectation.

Corollary 8.1. Let $\mathcal{A} \subseteq [0,1]^n$ be measurable. Then we have the following bound on the Hamming distance between X and $\mathcal{A}$:

$$\mathbb{P}[d_H(X,\mathcal{A}) \ge t + \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}(X\in\mathcal{A})}}] \le e^{-\frac{2t^2}{n}}.$$

Remark 8.2. Let us interpret this result. Suppose $\mathbb{P}(X \in \mathcal{A}) = \epsilon > 0$. Then the second term is roughly of size $\sqrt{n}$. If one chooses $t \sim C_0\sqrt{n}$ for large C_0, then the RHS is on the order of $e^{-2C_0^2} \overset{C_0\to\infty}{\longrightarrow} 0$. Thus, intuitively, if X has some mass at $\mathcal{A}$, then almost all of the mass of X is concentrated near $\mathcal{A}$ (to a $\mathcal{O}(\sqrt{n})$ order).

Proof of Corollary 8.1. First, note that $x \to d_H(x,A)$ is $1-$Lipschitz. Recall via the previous result that

$$\mathbb{P}[\mathbb{E}d_H(X,\mathcal{A}) - d_H(X,\mathcal{A}) \ge t] \le e^{-\frac{2t^2}{n}}.$$

Choose $t = \mathbb{E}d_H(X,\mathcal{A})$. Then the event becomes on the LHS becomes $\{d_H(X,\mathcal{A}) \le 0\}$. However, because the Hamming distance is a discrete metric,

$$d_H(X,\mathcal{A}) \le 0 \iff d_H(X,\mathcal{A}) = 0 \iff X \in \mathcal{A}.$$

Hence we find that

$$\mathbb{P}(X\in A) \le e^{-\frac{2\left(\mathbb{E}d_H(X,\mathcal{A})\right)^2}{n}} \implies \mathbb{E}d_H(X,\mathcal{A}) \le \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}(X\in\mathcal{A})}}.$$

Applying Azuma-Hoeffding again, we find that

$$\mathbb{P}[d_H(X,\mathcal{A}) \ge \mathbb{E}d_H(X,\mathcal{A}) + t] \le e^{-\frac{2t^2}{n}}.$$

Combining this with the bound on $\mathbb{E}d_H(X,\mathcal{A})$, we find the desired result. □

The important part of this reuslt is the set-blowup intuition. In particular, as noted in the earlier remark, once there is a set of constant mass, it suffices to 'blow up' the set, which immediately yields a mass of $1-\epsilon$.

8.2. **Gaussian Suprema and Comparison Inequalities.** Let $\{X_t, t \in \mathcal{T}\}$ be a multivariate cnetered Gaussian where $\mathcal{T}$ is a finite index set.[6]

We have seen through our Lipschitz concentration result that if $\max_{t\in\mathcal{T}} \mathbb{E}[X_t^2] < \sigma^2$, then

$$\mathbb{P}\left[\left|\max_{i\in\mathcal{T}} X_i - \mathbb{E}\left[\max_{i\in\mathcal{T}} X_i\right]\right| > t\right] \le 2e^{-\frac{t^2}{2\sigma^2}}.$$

[6]Just think of $X_1,\dots,X_n$. The use of $\mathcal{T}$ is mostly for general notational convenience.

This concentration result resolves questions about the tail fluctuation; however, we still do not know what $\mathbb{E}[\max_{i\in\mathcal{T}} X_i]$ is. Our remaining task is therefore to analyze this quantity.

We can of course ask the question, *why suprema?* We provide some examples to justify this use.

Example 8.3 (Eigenvalues of Random Matrix)**.** Random matrix W symmetric $n\times n$ matrix, with $W_{ii} \overset{iid}{\sim} \mathcal{N}(0, \frac{2}{n})$ and $W_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$ for $i \neq j$. Again we want to analyze the behavior of the largest eigenvalue $\lambda_{\max}$. Note that this quantity is $\lambda_{\max} = \sup_{\|x\|=1} H(x)$, where $H = \sum_{i,j} W_{ij}x_ix_j$. In this case, analyzing the inf/sup proves useful.

Note that we can also apply the result instead for $\lambda_{\min}$, or even the $i-$th eigenvalue λ_i, if one references the Courant-Fisher minimax result. Then we can write λ_i as a combination of inf, sup, and H.

Example 8.4 (Penalized Regression)**.** We consider the well-known and classical problem of penalized regression. Our setup is the following: let $(y_i, x_i)_{i=1}^n$, $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$, with the linear model $y_i = \langle x_i, \beta_0\rangle + \epsilon_i$ for $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Recall that the *least-squares regularized regression* parameter

$$\hat{\beta} = \underset{\hat{\beta}\in\mathbb{R}^p}{\arg\min}\left[\frac{1}{2}\|y - X\beta\|^2 + \lambda\ell(\beta)\right],$$

for ℓ a standard convex penalty (e.g., L^1, L^2).

In the high-dimensional setting, we are interested in when the dimension of our data is comparable to its size, e.g. when $\frac{p}{n} \to \kappa \in (0, \infty)$. Moreover, a reasonable approximation, and an interesting toy problem, is to assume $X_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$.

A good first step is to understand the minimum loss itself:

$$\begin{aligned}
\mathcal{L} &= \min_{\hat{\beta}\in\mathbb{R}^p}\left[\frac{1}{2}\|y - X\beta\|^2 + \lambda\ell(\beta)\right] \\
&= \min_{\hat{\beta}\in\mathbb{R}^p}\left[\frac{1}{2}\|X(\beta - \beta_0) + \epsilon\|^2 + \lambda\ell(\beta)\right] \\
&= \min_{\beta\in\mathbb{R}^p}\max_{u\in\mathbb{R}^n}\left[u^\intercal X(\beta - \beta_0) + u^\intercal\epsilon - \frac{\|u\|^2}{2} + \lambda\ell(\beta)\right].
\end{aligned}$$

The last three terms represent a Gaussian process, so the crux of this problem actually lies in understanding a type of minimax expression on a Gaussian process. Recent progress in this direction has come from the setting of Gaussian comparison inequalities.

In the next few lectures, we will try to find a way to systematically analyze this type of result, and prove such results.

9. Lecture 9: October 2

We continue our discussion of Gaussian comparison inequalities. Recall the setup: we have some multivariate Gaussian vector $X \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^n$, and we want to analyze the maximum, $Y = \max_{i\leq n} X_i$.

The first idea is to compare our multivariate Gaussian to a 'simpler' process. We should expect independent Gaussians to lead to the largest maximum. In this direction, we already have one tool: Gaussian Interpolation, Lemma 5.1. We will use this to derive new results.

The second idea is to attempt to understand the dependence of the maximum on the 'complexity' of the index set. We will see this in later lectures.

9.1. **Two Technical Results.** In the vein of the first idea, we will show two results. The proofs of both results will rely on using Gaussian Interpolation and justifying the sign of a derivative. First, we have *Slepian's Inequality*:

Theorem 9.1 (Slepian's Inequality)**.** Let X, Y be two centered Gaussian vectors in $\mathbb{R}^n$ satisfying:

- $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$ for all $i \leq n$, and
- $\mathbb{E}X_iX_j \leq \mathbb{E}Y_iY_j$ for all $i \neq j$.

Then for all sequences $(\lambda_i)_{i=1}^n \in \mathbb{R}^n$, we have the inequality

$$\mathbb{P}(X_i \leq \lambda_i \,\forall i) \leq \mathbb{P}(Y_i \leq \lambda_i \,\forall i).$$

Proof. First note that $\mathbb{I}\left(\bigcap_{i=1}^n \{X_i \leq \lambda_i\}\right) = \prod_{i=1}^n \mathbb{I}(X_i \leq \lambda_i)$. Now, we can approximate each of the indicators by smooth functions. In particular, let $\phi_i(X_i)$ be decreasing smooth functions such that $\phi_i(X_i) \geq \mathbb{I}\left(X_i \leq \lambda_i\right) \geq 0$. Let $\phi(X) = \prod_{i=1}^n \phi_i(X_i)$. Now, let X, Y be independent, and let

$$f(t) = \mathbb{E}[\phi(\overbrace{\sqrt{t}X + \sqrt{1-t}Y}^{Z(t)})].$$

By Gaussian Interpolation and the first condition,

$$\begin{aligned} f'(t) &= \frac{1}{2}\sum_{i,j}\left(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j\right)\mathbb{E}\left(\frac{\partial^2\phi}{\partial X_i \partial X_j}(Z(t))\right) \\ &= \frac{1}{2}\sum_{i\neq j}\left(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j\right)\mathbb{E}\left(\frac{\partial^2\phi}{\partial X_i \partial X_j}(Z(t))\right). \end{aligned}$$

Now we can compute

$$\frac{\partial^2\phi}{\partial X_i \partial X_j}(x) = \left[\prod_{k\notin\{i,j\}} \phi_k(X_k)\right]\frac{\partial\phi_i}{\partial X_i}(X_i)\frac{\partial\phi_j}{\partial X_j}(X_j) \geq 0.$$

Th final inequality results because the first term is nonnegative as $\phi_k \geq 0$, and the second and third are both negative as ϕ_i is decreasing by definition. By the second condition, we find that

$$f'(t) \leq 0 \implies f(1) \leq f(0).$$

We now attain that $\mathbb{E}\phi(X) \leq \mathbb{E}\phi(Y)$. Now, applying MCT and letting the ϕ_i approach $\mathbb{P}(X_i \leq \lambda_i)$ yields that

$$\mathbb{P}(X_i \leq \lambda_i \,\forall i) \leq \mathbb{P}(Y_i \leq \lambda_i \,\forall i).$$

□

We have the following immediate corollary, which is often also listed under the statement of Theorem 9.1.

Corollary 9.2. Under the conditions of Theorem 9.1, we have the inequality

$$\mathbb{E}(\max X_i) \geq \mathbb{E}(\max Y_i).$$

Proof. By letting $\lambda_i = \lambda$ in the conclusion of Theorem 9.1, we find that $\mathbb{P}(\max X_i \le \lambda) \le \mathbb{P}(\max Y_i \le \lambda)$. Equivalently, we have that $\mathbb{P}(\max X_i > \lambda) \ge \mathbb{P}(\max Y_i > \lambda)$. The result is then intuitively obvious. Rigorously, we can define some variables $U_n = \min(\max X_i, -n)$ and $V_n = \min(\max Y_i, -n)$. Then $\mathbb{E}U_n \to \mathbb{E}(\max X_i)$, $\mathbb{E}V_n \to \mathbb{E}(\max Y_i)$ (indeed, the only difference in the expectations is in the tails, which can be seen to approach 0). Then we find that

$$\mathbb{E}(U_n + n) = \int \mathbb{P}(U_n + n > \lambda) \ge \int \mathbb{P}(V_n + n > \lambda) = \mathbb{E}(V_n + n).$$

it follows that $\mathbb{E}U_n \ge \mathbb{E}V_n$, from which the result follows by taking the limit as $n \to \infty$. □

We next prove the Sudakov-Fernique inequality, which is of a similar vein.

Theorem 9.3 (Sudakov-Fernique)**.** Let X, Y be two centered Gaussians in $\mathbb{R}^n$. Suppose that

$$\mathbb{E}(X_i - X_j)^2 \ge \mathbb{E}(Y_i - Y_j)^2 \ \forall i, j.$$

Then $\mathbb{E}(\max X_i) \ge \mathbb{E}(\max Y_i)$.

Remark 9.4. Note that $\mathbb{E}(X_i - X_j)^2 = \mathbb{E}X_i^2 + \mathbb{E}X_j^2 - 2\mathbb{E}X_iX_j$. Therefore if $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, then the Slepian Inequality conditions also hold. Conversely, if the Slepian assumptions hold, so do the Sudakov-Fernique assumptions.

Proof. The idea is to approximate the maximum with smooth functions, similar to the first problem in the first problem set. Let

$$F(X) = \frac{1}{\beta} \log \left(\sum_{i=1}^{n} e^{\beta X_i} \right)$$

for $\beta > 0$ be the smooth maximum functions. It is easy to see that

$$\max X_i \le F(X) \le \max X_i + \frac{\log n}{\beta}.$$

We will now prove that $\mathbb{E}F(X) \ge \mathbb{E}F(Y)$. Then letting $\beta \to \infty$ and applying DCT (with $|F(X)| \le |\max X_i| + |\max X_i + \frac{\log n}{\beta}|$) will yield the result.

As in the proof of Slepian, choose X, Y independent and let $Z(t) = \sqrt{t}X + \sqrt{1-t}Y$, $f(t) = \mathbb{E}F(Z(t))$. For notational simplicity, write $a_{ij} = \mathbb{E}X_iX_j$ and $b_{ij} = \mathbb{E}Y_iY_j$. By Gaussian Interpolation, we again have

$$f'(t) = \frac{1}{2} \sum_{i,j} (a_{ij} - b_{ij}) \mathbb{E} \left(\frac{\partial^2 F}{\partial X_i \partial X_j}(Z(t)) \right).$$

Now, note the following formulas from direct expansion:

$$p_i(x) = \frac{\partial F}{\partial X_i}(X) = \frac{e^{\beta X_i}}{\sum_j e^{\beta X_j}}$$

$$\frac{\partial^2 F}{\partial X_i^2}(X) = \beta(p_i(X) - p_i(X)^2)$$

$$\frac{\partial^2 F}{\partial X_i \partial X_j}(X) = -\beta p_i(X) p_j(X), i \ne j.$$

Using these to simplify the expression for f', we find that[7]

$$f'(t) = \frac{\beta}{2}\sum_{i=1}^{n}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}[(p_i - p_i^2)(Z(t))] - \frac{\beta}{2}\sum_{i\neq j}(\mathbb{E}X_iX_j - \mathbb{E}Y_iY_j)\mathbb{E}[(p_ip_j)(Z(t))].$$

Moreover, note that $\sum p_k(X) = 1 \implies p_i(X) - p_i(X)^2 = \sum_{j\neq i} p_i(X)p_j(X)$. Therefore, we see that

$$\sum_i(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}_{Z(t)}(p_i - p_i^2) = \sum_{i\neq j}(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}_{Z(t)}(p_ip_j) = \sum_{i\neq j}(\mathbb{E}X_j^2 - \mathbb{E}Y_j^2)\mathbb{E}_{Z(t)}(p_ip_j).$$

It follows that

$$\sum_i(\mathbb{E}X_i^2 - \mathbb{E}Y_i^2)\mathbb{E}_{Z(t)}(p_i - p_i^2) = \frac{1}{2}\sum_{i\neq j}(\mathbb{E}X_i^2 + \mathbb{E}X_j^2 - \mathbb{E}Y_i^2 - \mathbb{E}Y_j^2)\mathbb{E}_{Z(t)}(p_ip_j).$$

We conclude that

$$f'(t) = \frac{\beta}{4}\sum_{i\neq j}\mathbb{E}_{Z(t)}(p_ip_j)[\mathbb{E}X_i^2 + \mathbb{E}X_j^2 - 2\mathbb{E}X_iX_j - \mathbb{E}Y_i^2 - \mathbb{E}Y_j^2 + 2\mathbb{E}Y_iY_j] \geq 0$$

by the condition. Hence we find that $f(1) \geq f(0)$, i.e. $\mathbb{E}F(X) \geq \mathbb{E}F(Y)$, as desired. □

Remark 9.5. Note that the statement holds as long as $\mathbb{E}X_i = \mathbb{E}Y_i$ for all i, even if the Gaussians are not centered.

Remark 9.6. These theorems also extends to infinite index sets (i.e., Gaussian processes), provided that said processes are continuous and that the index set satisfies some 'nice' conditions, in particular, if it is a compact set within a metric space. We will not provide the full proof here, but it is simply by discretization. Indeed, noting that the processes will be uniformly continuous, we can apply the discrete versions of these inequalities, and take a limit as our discretization becomes more fine.

10. Lecture 10: October 7

We first completed the discussion of the Sudakov-Fernique inequality. We now discuss several applications.

10.1. **Norms of Gaussian Random Matrices.**

We can now show the asymptotically optimal result on the operator norm of a Gaussian RM!

Lemma 10.1. Let A be an m times n matrix, where $A_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$. THen

$$\mathbb{E}\|A\| \leq \sqrt{m} + \sqrt{n}.$$

[7] In the below equations, an expression like $(p_i - p_i^2)(Z(t))$ is the application of the function $p_i - p_i^2$ to $Z(t)$.

Proof. Note that $\|A\| = \max_{u\in S^{m-1}, v\in S^{n-1}} \langle Au, v\rangle = \max_{\langle u,v\rangle \in T} X_{uv}$, where $T = S^{m-1}\times S^{n-1}$ and $X_{uv} = \langle Au, v\rangle$. Now, note that

$$\begin{aligned}
\mathbb{E}(X_{uv} - X_{wz})^2 &= \mathbb{E}\left[(\langle Au, v\rangle - \langle Aw, z\rangle)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j} A_{ij}(u_jv_i - w_jz_i)\right)^2\right] \\
&= \sum_{i,j}(u_jv_i - w_jz_i)^2 \\
&\le \|u-w\|^2 + \|v-z\|^2.
\end{aligned}$$

Note that the last inequality follows from an expansion and the Cauchy-Schwarz inequality. In particular,

$$\begin{aligned}
\sum_{i,j}(u_jv_i - w_jz_i)^2 &= \sum_i\sum_j (u_jv_i - w_jz_i)^2 \\
&= \sum_i (v_i^2 + z_i^2 - 2v_iz_i\sum_j u_jw_j) \\
&= 2 - 2\sum_i v_iz_i \sum_j u_jw_j \\
&\le 4 - 2\sum_i v_iz_i - 2\sum_j u_jw_j \\
&= \|u-w\|^2 + \|v-z\|^2.
\end{aligned}$$

The second-to-last inequality here follows from rearranging into $(1-\sum v_iz_i)(1-\sum_j u_jw_j) \ge 0$ and noting that both terms are nonnegative by Cauchy-Schwarz.

Now, define $Y_{uv} = \langle g, u\rangle + \langle h, v\rangle$, where $g, h \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_n)$ are independent Gaussians.[8] Since $\mathbb{E}X_{uv} = 0$ and g, h are independent we find that Y_{uv} is also a centered Gaussian process. Therefore, we see that

$$\mathbb{E}(Y_{uv} - Y_{wz})^2 = \mathbb{E}[(\langle g, u-w\rangle + \langle h, v-z\rangle)^2] = \|u-w\|^2 + \|v-z\|^2.$$

Recalling that $S^{m-1}\times S^{n-1}$ is compact and X_{uv}, Y_{uv} are linear and hence continuous Gaussian Processes, we can apply the Sudakov-Fernique inequality to find that

$$\mathbb{E}\|A\| = \mathbb{E}(\sup_{(u,v)\in T} X_{uv}) \le \mathbb{E}(\sup_{(u,v)\in T} Y_{uv}).$$

Now, the right hand side is bounded by

$$\begin{aligned}
\mathbb{E}(\sup_{(u,v)\in T} Y_{uv}) &= \mathbb{E}[\sup_{(u,v)\in T}(\langle g, u\rangle + \langle h, v\rangle)] \\
&= \mathbb{E}\|g\| + \mathbb{E}\|h\| \\
&\le \sqrt{\mathbb{E}\|g\|^2} + \sqrt{\mathbb{E}\|h\|^2} \\
&= \sqrt{m} + \sqrt{n}.
\end{aligned}$$

This completes the proof. □

[8] They should also be independent from the space of Gaussians $u, v \in T$.

Remark 10.2. First, note that this bounded is asymptotically tight. Indeed, as $m, n \to \infty$ with $\frac{m}{n} \to \gamma$, the Bai-Yin Law implies that

$$\frac{\|A\|}{\sqrt{n}} \overset{a.s.}{\to} 1 + \sqrt{\gamma}.$$

This illustrates the strength of the Gaussian comparison inequalities we have shown. Indeed, the comparison-based results are nonasymptotic, which is convenient. However, the obvious con is the Gaussianity requirement. For instance, Bai-Yin holds for other distributions with 'nice' moments; one should expect similar results in these non-Gaussian settings.

It turns out that similar bounds can be derived for matrices with subgaussian entries, though we will not do so in this course. See [Ver11] for a reference.

10.2. **Gordon's Inequality.** We now mention one more comparison inequality, known as *Gordon's Inequality*.

Theorem 10.3 (Gordon's Inequality)**.** Let X_{ij} and Y_{ij} be $m \times n$ arrays of joint Gaussians with equal means. Assume that the following three conditions hold:

- $\mathrm{Cov}(X_{ij}, X_{i\ell}) \geq \mathrm{Cov}(Y_{ij}, Y_{i\ell})\ \forall i, j, \ell$,
- $\mathrm{Cov}(X_{ij}, X_{k\ell}) \leq \mathrm{Cov}(Y_{ij}, Y_{k\ell})\ \forall i, j, k, \ell$ and $i \neq k$, and
- $\mathrm{Var}(X_{ij}) = \mathrm{Var}(Y_{ij})\ \forall i, j$.

Then the following hold:

(1) For any $t_{ij} \in \mathbb{R}$, we have

$$\mathbb{P}\left[\bigcap_i \bigcup_j \{X_{ij} > t_{ij}\}\right] \leq \mathbb{P}\left[\bigcap_i \bigcup_j \{Y_{ij} > t_{ij}\}\right], \text{ and}$$

(2) Choosing all the t_{ij} to be equal, we find that

$$\mathbb{E}[\min_i \max_j X_{ij}] \leq \mathbb{E}[\min_i \max_j Y_{ij}].$$

Remark 10.4. Note that if $n = 1$, then this reduces directly to Slepian's inequality. Moreover, there exists an extension of this result where the conditions depend only on increments of the Gaussians; however, this extension is tedious.

The proof of this inequality will be assigned in problem set two. As a hint, one should follow similar steps as in the proof of Slepian's inequality. In particular, it will suffice to use Gaussian interpolation with

$$f(x) = \prod_i \left[1 - \prod_j h(X_{ij})\right],$$

where $h(X_{ij})$ is a smooth approximation to $\mathbb{I}(X_{ij} \geq t_{ij})$.

We now move on to discussing a different side of comparison inequalities; in particular, the extent of their strength.

10.3. **The Convex Gaussian Minimax Theorem.** Now that we have an arsenal of comparison inequalities, there is an obvious question: *how strong are these comparison inequalities?* It turns out that the answer is *quite a bit*, as demonstrated by the following relatively recent result, known as the *Convex Gaussian Minimax Theorem*.

Theorem 10.5 (Convex Gaussian Minimax Theorem (CGMT)[9])**.** Let $S_w \subseteq \mathbb{R}^N$, $S_u \subseteq \mathbb{R}^n$ be compact, and $Q : S_w \times S_u \to \mathbb{R}$ be continuous. Let the matrix $G = (G_{ij})$ have IID $\mathcal{N}(0,1)$ entries. Moreover, let $g \sim \mathcal{N}(0, \mathbf{I}_N)$ and $h \sim \mathcal{N}(0, \mathbf{I}_n)$ be independent Gaussians. Lastly, let

$$C^*(G) = \min_{w \in S_w} \max_{u \in S_u} [u^\intercal G w + Q(w,u)],$$

and

$$L^*(g,h) = \min_{w \in S_w} \max_{u \in S_u} \left[\|u\| \langle g, w \rangle + \|w\| \langle h, u \rangle + Q(w,u) \right].$$

Then the following hold:

(1) For all $t \in \mathbb{R}$, we have $\mathbb{P}(C^*(G) \le t) \le 2\mathbb{P}(L^*(g,h) \le t)$, and
(2) If S_w, S_u are convex and Q is convex-concave,[10] then for all $t \in \mathbb{R}$, we have

$$\mathbb{P}(C^*(G) \ge t) \le 2\mathbb{P}(L^*(g,h) \ge t).$$

Remark 10.6. This is a very general and quite strong result. It will turn out to yield very tight bounds on many relevant estimators in statistics and ML, justifying the power of the method of comparison inequalities.

11. Lecture 11: October 9

We will continue to discuss the Convex Gaussian Minimax Theorem (CGMT).

The purpose of this result is as follows. We often care about some type of minimax expression such as in C^* (for instance, in evaluating the operator norm, etc.). In general, these are difficult to work with. However, we can find some L^* which is the sum of terms, each of which can be independently analyzed, and is therefore easier to work with.

Typically, CGMT will give us a formal relation between the tails of C^* and L^*; hence, tail bounds on L^* will give us tail bounds on C^*. Moreover, if $L^* \xrightarrow{D} \mu^*$ then $C^* \xrightarrow{D} \mu^*$ as well.

Why do we care? It turns out that this result is motivated by methods in penalized regression in machine learning, and in particular gives us strong bounds.

Example 11.1. Recall the standard regression setup: $x_i \in \mathbb{R}^p$ and $y \in \mathbb{R}$, where

$$y_i = \langle x_i, \beta_0 \rangle + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Our goal is to understand the regression estimator

$$\hat{\beta} = \arg\min \frac{1}{2} \|y - X\beta\|^2 + \lambda \ell(\beta)$$

for a convex penalization loss ℓ, in the high dimensional setting $\frac{p}{n} \to \kappa \in (0, \infty)$. This problem is highly general and difficult; to make it more concrete, we make the Gaussianity assumption $x_{ij} \sim \mathcal{N}(0,1)$.

[9] From [TOH15], and earlier work by Stojnik.

[10] That is, $Q(\cdot, u)$ is convex and $Q(w, \cdot)$ is concave.

The key idea is to use Theorem 10.5. Set

$$\begin{aligned} C^* &= \frac{1}{n}\left[\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|^2 + \lambda\ell(\beta)\right] \\ &= \frac{1}{n}\left[\min_{\beta\in\mathbb{R}^p}\max_{u\in\mathbb{R}^n} u^\intercal(y - X\beta) - \frac{\|u\|^2}{2} + \lambda\ell(\beta)\right] \\ &= \frac{1}{n}\left[\min_{\beta\in\mathbb{R}^p}\max_{u\in\mathbb{R}^n} u^\intercal X(\beta_0 - \beta) + \overbrace{u^\intercal\epsilon - \frac{\|u\|^2}{2}}^{Q(x)} + \lambda\ell(\beta)\right]. \end{aligned}$$

where the second line is the standard trick of rewriting the loss as a minimax expression.[11]

Now, the point is to apply CGMT and analyze the resulting expression L^* to find $C^* \approx L^* \to \mu$. Note that this $\approx$ can be made concrete. We thus have guarantees on the distribution of C^*.

Next, we would like to understand the actual estimator $\hat{\beta}$. The idea is the following: suppose $\|\hat{\beta} - \beta_0\|^2 > p\delta_0$ is large, i.e. that $\hat{\beta}$ is far from β_0. To this end, consider the loss function

$$C_1^* = \frac{1}{n}\left[\min_{\beta:\frac{1}{p}\|\beta-\beta_0\|^2>\delta} \frac{1}{2}\|y - X\beta\|^2 + \lambda\ell(\beta)\right].$$

If we apply Gordon's Comparison Inequality, some amount of analysis will lead to a result of the form $\delta_0 > \delta_1 \implies C_1^* > \mu^* + \epsilon$ where $\epsilon = F(\delta)$ is some function of δ. For large δ, our assumption will imply that $C^* = C_1^* > \mu^* + \epsilon$, contradicting our guarantees that $C^* \approx \mu^*$. This yields that for large n, p we have a bound $\frac{1}{p}\|\hat{\beta} - \beta_0\|^2 \le \delta_0$.

We thus obtain a quantitative understanding of $\hat{\beta}$. We can also obtain the following concrete applications.

(1) We can get precise characterizations of δ_0, which can inform the choice of the penalization parameter λ, and
(2) We can understand the empirical average of the optimizer, and attain 'law of large numbers' type of results, e.g. $\frac{1}{p}\sum_{j=1}^p f(\hat{\beta}_{0,j}) \approx \eta$. Here, $\hat{\beta}_{0,j}$ is the jth coordinate of $\hat{\beta}$.

Now, we will move onto the proof of this result. The first part will be an application of Gordon's Inequality, and the second will rely on convexity. First, we will need to state a general version of Gordon's Inequality.

Theorem 11.2 (Gordon's Inequality, General Form)**.** Let $D_u \subseteq \mathbb{R}^n$, $D_v \subseteq \mathbb{R}^n$ be compact sets, and let $Q : D_u \times D_v \to \mathbb{R}$ be continuous. Consider two centered Gaussian processes

$$\{X(u, v) : u \in D_u, v \in D_v\},\ ,\{Y(u, v) : u \in D_u, v \in D_v\}$$

with continuous trajectories.[12] Suppose that

- $\mathbb{E}X(u, v)^2 = \mathbb{E}Y(u, v)^2$ for all $(u, v) \in D_u \times D_v$,

[11]This is referred to by 'linearization' of such a quadratic form. Analogously if we had a general convex function, one can take a 'dual'.

[12]That is, $(u, v) \to X(u, v)$ is continuous a.s.

- $\mathbb{E}[X(u,v)X(u,v')] \geq \mathbb{E}[Y(u,v)Y(u,v')]$ for all $u \in D_u, v \neq v' \in D_v$, and
- $\mathbb{E}[X(u,v)X(u',v')] \leq \mathbb{E}[Y(u,v)Y(u',v')]$ for all $u \neq u' \in D_u, v, v' \in D_v$.

Then we have the inequality

$$\mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(Y(u,v)+Q(u,v)) \leq t\right] \leq \mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(X(u,v)+Q(u,v)) \leq t\right].$$

This is of course a direct continuous analog of Gordon. For a reference, see the appendices to [MM18].

Proof Sketch. As usual, discretize the index set and apply Gordon to the discrete process. By compactness of the index sets, X, Y are uniformly continuous. The result follows by taking limits. □

Now we are ready to prove Theorem 10.5.

Proof of Theorem 10.5. We will of course start with the first item then prove the second. To make the application of Gordon clear, we are going to butcher some notation and replace $(w,u) \to (u,v)$ in the theorem statement.

Item (1). The goal is to use Gordon's inequality. However, note that the result contains a factor of 2, which we will need to create.

First, define two Gaussian processes:

$$X(u,v) = \|v\|g^\intercal u + \|u\|h^\intercal v,\ Y(u,v) = v^\intercal G u + \|u\|\|v\|\zeta.$$

Here, $\zeta \sim \mathcal{N}(0,1) \perp\!\!\!\perp G$ is some independent Gaussian; this is introduced for the sake of the proof.

We now verify the conditions of Gordon's inequality. First, note that because G, ζ, g, h are independent Gaussians, we have

$$\begin{aligned}\mathbb{E}[Y(u,v)Y(u',v')] - \mathbb{E}[X(u,v)X(u',v')] &= (u^\intercal u')(v^\intercal v') + \|u\|\|u'\|\|v\|\|v'\| \\ &\quad - \|v\|\|v'\|(u^\intercal u') - \|u\|\|u'\|(v^\intercal v') \\ &= (\|u\|\|u'\| - u^\intercal u')(\|v\|\|v'\| - v^\intercal v') \\ &\geq 0\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz. This shows the third condition. However, note that if $u = u'$ or $v = v'$ then in fact the above inequality is an equality. This shows the first two conditions. By Gordon's Inequality, we have

$$\mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(Y(u,v)+Q(u,v)) \leq t\right] \leq \mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(X(u,v)+Q(u,v)) \leq t\right].$$

Now this is almost the desired result; however, we will need to address the term $\|u\|\|v\|\zeta$. However, note that ζ is symmetric. In particular, we can remove the ζ term whenever $\zeta < 0$,

but $\zeta \geq 0$ occurs with probability $\frac{1}{2}$. We can thus write

$$\begin{aligned}\mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(Y(u,v)+Q(u,v))\leq t\right] &\geq \frac{1}{2}\mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(Y(u,v)+Q(u,v))\leq t \mid \zeta\geq 0\right]\\ &\geq \frac{1}{2}\mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(v^\intercal Gu+Q(u,v))\leq t \mid \zeta\geq 0\right]\\ &= \frac{1}{2}\mathbb{P}\left[\min_{u\in D_u}\max_{v\in D_v}(v^\intercal Gu+Q(u,v))\leq t\right].\end{aligned}$$

Combining the above results, we find that

$$\mathbb{P}(C^*(G)\leq t)\leq 2\mathbb{P}(L^*(g,h)\leq t),$$

as desired.

Item (2). We must now make use of the convexity condition. First, apply the result in item (1) with $(u,v,Q,t)\to(v,u,-Q,-t)$. We then find

$$\mathbb{P}\left[\min_{v\in D_v}\max_{u\in D_u} v^\intercal Gu - Q(u,v)\leq -t\right]\leq 2\mathbb{P}\left[\min_{v\in D_v}\max_{u\in D_u}\|v\|g^\intercal u+\|u\|h^\intercal v-Q(u,v)\leq -t\right].$$

Recall that S_u, S_v are convex and Q is convex-concave, and that $-G=G$. By standard minimax theorems, we can swap the min and max, and then negate, yielding

$$\begin{aligned}\mathbb{P}\left[\min_{v\in D_v}\max_{u\in D_u} v^\intercal Gu - Q(u,v)\leq -t\right] &= \mathbb{P}\left[\max_{u\in D_v}\min_{v\in D_u} v^\intercal Gu - Q(u,v)\leq -t\right]\\ &= \mathbb{P}\left[\min_{u\in D_v}\max_{v\in D_u} v^\intercal - Gu + Q(u,v)\geq t\right]\\ &= \mathbb{P}\left[\min_{u\in D_v}\max_{v\in D_u} v^\intercal Gu + Q(u,v)\geq t\right].\end{aligned}$$

Moreover, note that the Gaussians $g,h \overset{d}{=} g,h$. Then the RHS equals

$$\begin{aligned}\mathbb{P}\left[\min_{v\in D_v}\max_{u\in D_u}\|v\|g^\intercal u+\|u\|h^\intercal v-Q(u,v)\leq -t\right] &= \mathbb{P}\left[\max_{v\in D_v}\min_{u\in D_u}(-\|v\|g^\intercal u-\|u\|h^\intercal v+Q(u,v))\geq t\right]\\ &= \mathbb{P}\left[\max_{v\in D_v}\min_{u\in D_u}(\|v\|g^\intercal u+\|u\|h^\intercal v+Q(u,v))\geq t\right]\\ &\leq \mathbb{P}\left[\min_{v\in D_v}\max_{u\in D_u}(\|v\|g^\intercal u+\|u\|h^\intercal v+Q(u,v))\geq t\right],\end{aligned}$$

where the last inequality follows from $\min\max\cdot\leq\max\min\cdot$. Combining all our inequalities, we find that

$$\mathbb{P}(C^*(G)\geq t)\leq 2\mathbb{P}(L^*(g,h)\geq t),$$

as desired. This completes the proof of both claims of the theorem. □

12. Lecture 12: October 16

Recall that our current approach, Gaussian comparison, relies on relating our current Gaussian process to a simpler one. This is very strong when it works, and provides good non-asymptotic guarantees (as seen in the previous lectures); however, it is quite restrictive and relies strongly on the Gaussianity assumption.

12.1. **Gaussian Suprema via Metrics.** We now move onto a separate approach to analyze the suprema of Gaussian processes: in particular, we take a metric-based approach. We will also introduce the Dudley entropy bound. For a reference, see [H18], Chapter 5.

Recall the basic setup: we have a centered Gaussian process $\{X_t, t \in T\}$, and we would like to analyze the quantity $\mathbb{E}[\sup_{t\in T} X_t]$.

Remark 12.1. We make this remark now: that $\sup X_t$ is not necessarily measurable. This is okay if we make reasonably nice assumptions; for instance, either countability, or some type of compactness assumption. We will more or less ignore this issue with now on; in general applications, this can be proven in a case-by-case basis.

We now introduce the following lemma. This addresses the *finite* case of the problem.

Lemma 12.2 ([H18], Lemma 5.1, adapted)**.** Suppose that ψ is a differentiable convex function such that $\psi(0) = \psi'(0) = 0$ and

$$\log \mathbb{E}[e^{\lambda X_t}] \le \exp(\psi(\lambda))$$

for all $\lambda \ge 0, t \in T$. Let $\psi^*(x) = \sup_{\lambda\ge 0}(\lambda x - \psi(\lambda))$ be the Fenchel conjugate of ψ. Then[13]

$$\mathbb{E}[\sup_{t\in T} X_t] \le (\psi^*)^{-1}(\log |T|).$$

Proof. Via the conditions and Jensen's inequality, we find that

$$\begin{aligned}
\mathbb{E}[\sup_{t\in T} X_t] &= \frac{1}{\lambda}\mathbb{E}[\log e^{\lambda \sup_{t\in T} X_t}] \\
&\le \frac{1}{\lambda}\log \mathbb{E}[e^{\lambda \sup_{t\in T} X_t}] \\
&\le \frac{1}{\lambda}\log \mathbb{E}[\sum_{t\in T} e^{\lambda X_t}] \\
&\le \frac{1}{\lambda}\log \mathbb{E}[|T| e^{\psi(\lambda)}] \\
&= \frac{\log |T|}{\lambda} + \frac{\psi(\lambda)}{\lambda}.
\end{aligned}$$

Now, we find that

$$\mathbb{E}[\sup_{t\in T} X_t] \le \inf_{\lambda}\left(\frac{\log |T|}{\lambda} + \frac{\psi(\lambda)}{\lambda}\right).$$

The last term is exactly $(\psi^*)^{-1}(\log |T|)$. Indeed, it suffices to note that if $z = (\psi^*)^{-1}(\log |T|)$, then $z \le \frac{\psi^*(z)+\psi(\lambda)}{\lambda} = \frac{\log |T|+\psi(\lambda)}{\lambda}$ by the definition of ψ^*. Moreover, as λ approaches the optimizer, equality is attained.[14] We thus find that

$$\mathbb{E}[\sup_{t\in T} X_t] \le (\psi^*)^{-1}(\log |T|),$$

[13] assuming that the set T is finite, i.e. $|T| < \infty$.

[14] Two remarks are in store. First, that ψ^* is invertible, and second that the optimizer λ^* is not at 0 (otherwise the statement is not obviously trure, because we divide by $\lambda \to \lambda^* = 0$). For this, see the proof of Lemma 5.1 in the reference [H18]. The point is that the function ψ^* is strictly increasing except when the optimizer is 0, in which case one finds $x \le 0$; moreover, we have exactly that $\psi^*(0) = 0$. This justifies the first remark, as well as the second when $x > 0$. For $x = 0$, just note that setting $\lambda \to 0$ and using the condition that $\psi'(0) = 0$ yields the infimum to be at most 0 anyway.

as desired. □

Note that we immediately find the following corollary.

Corollary 12.3. If X_t is σ^2-subgaussian, then

$$\mathbb{E}[\sup_{t\in T} X_t] \leq \sqrt{2\sigma^2 \log |T|}.$$

Proof. Note that we can apply the result of the lemma to $\psi(\lambda) = \frac{\sigma^2\lambda^2}{2}$). The Fenchel conjugate is $\psi^* = \frac{x^2}{2\sigma^2}$; hence, $(\psi^*)^{-1}(z) = \sqrt{2\sigma^2 z}$. The application of the lemma then immediately yields the desired result. □

Remark 12.4. This bound is reasonably nice; however, it can be loose. For instance, if we take $X_1 = \cdots = X_n = Z$, $Z \sim \mathcal{N}(0,1)$, then $\mathbb{E}\max X_i = \mathbb{E}Z = \mathcal{O}(1)$.

This remark suggests some sort of reliance of the maximum quantity on dependence. Note that this phenomenon is also touched upon in the statement of the previous comparison inequalities, which rely on bounds on the covariances. Nonetheless, this phenomenon will arise further.

In particular, we ask the following natural **question:** *What is the behavior of the supremum* $\mathbb{E}[\sup_{t\in T} X_t]$ *when the sample set* T *is infinite?* Note that in the IID setting, the answer is clearly ∞, but if say $X_t : t \in [0,1] \sim$ BM is Brownian Motion, then $\sup_{t\in[0,1]} X_t = \mathcal{O}(1)$. To analyze this, we will consider the more general setting of subgaussian processes in metric spaces. In this setting, we will find geometric techniques to be important; in particular, we will introduce the concepts of *packing* and *covering*.

12.2. **Packing and Covering.** Let (T,d) be a metric space (where d is the distance metric). We have two definitions.

Definition 12.5 (ϵ-net and covering numbers)**.** A set $\mathcal{N} \subseteq T$ is called an $\epsilon-$*net* for T if for all $t \in T$, there is some $\pi(t) \in \mathcal{N}$ such that $d(t,\pi(t)) \leq \epsilon$. The smallest cardinality of an $\epsilon-$net of (T,d) is called the *covering number*:

$$N(T,d,\epsilon) = \inf\{|\mathcal{N}| : \mathcal{N} \text{ is } \epsilon-\text{net}\}.$$

Intuitively, an $\epsilon-$net is one such that if we take spheres centered at the points in the net with radius ϵ, they will *cover* the space T.

Definition 12.6 (ϵ-packing and packing numbers)**.** A set $\mathcal{N} \subseteq T$ is called an $\epsilon-$*packing* of (T,d) if $d(t,t') > \epsilon$ for all $t,t' \in \mathcal{N}$, $t \neq t'$. The largest cardinality of an $\epsilon-$packing of (T,d) is called the *packing number*:

$$D(T,d,\epsilon) = \sup\{|\mathcal{N}| : \mathcal{N} \text{ is } \epsilon-\text{packing}\}.$$

Intuitively, an $\epsilon-$packing is one such that if we take spheres centered at the points in the packing with radius ϵ, no two will intersect. This is analogous to the familiar notion of *sphere packing*, which this generalizes to arbitrary metric spaces.

Remark 12.7. These two notions are dual to each other, and are related in concrete ways. We will see more of this below.

The following lemma relates covering numbers and packing numbers, and suggests that they are asymptotically equivalent.

Lemma 12.8. For every $\epsilon > 0$, we have

$$D(T, d, 2\epsilon) \leq N(T, d, \epsilon) \leq D(T, d, \epsilon).$$

Proof Sketch. Suppose $\{p_i\}$ forms a $2\epsilon-$packing; then in any $\epsilon-$net $\{q_i\}$, there exists for each j at most one i such that $d(p_i, q_j) < \epsilon$ (as $d(p_i, p_{i'}) \leq d(p_i, q_j) + d(p_{i'}, q_j) < 2\epsilon$ would be a contradiction). However, the q_j must also completely cover all the p_i. It follows that $|\{q_i\}| \geq |\{p_i\}|$. The left inequality follows immediately.

Now, take any maximal $\epsilon-$packing $\{p_i\}$. This must also be an $\epsilon-$net. Indeed, taking any other point $x \in T$, if $d(x, p_i) > \epsilon$ for all ϵ, then appending x to $\{p_i\}$ contradicts maximality. The right inequality follows immediately. □

Remark 12.9. Heuristically, we should expect the covering number and packing number of the ball with induced Euclidean metric in n dimensions to be on the order of ϵ^{-n}.

Lastly, we have the following definitions, to make our object of study more concrete.

Definition 12.10 (Subgaussian Process on Metric Spaces)**.** A *subgaussian process* $\{X_t : t \in T\}$ on a metric space (T, d) is one such that $\mathbb{E}X_t = 0$ for all t and

$$\mathbb{E}[e^{\lambda(X_t - X_s)}] \leq e^{\frac{\lambda^2 d(t,s)^2}{2}}$$

for all $\lambda \geq 0$ and $s, t \in T$.

Definition 12.11 (Separable Processes)**.** A random process $\{X_t : t \in T\}$ is called *separable* if there exists a countable set $T_0 \subseteq T$ such that almost surely, the event

$$\{\forall t \in T, \exists s_n \in T_0; X_{s_n} \to X_t\}$$

holds.

Remark 12.12. Two important remarks. First, via the Chernoff bound, any subgaussian process satisfies

$$\mathbb{P}[X_t - X_s > x d(s, t)] < C_1 e^{-\frac{x^2}{C_2}}.$$

Secondly, any separable process satisfies $\sup_{t \in T} X_t = \sup_{t \in T_0} X_t$ almost surely, by continuity of the supremum.

With these prerequisites, we can finally state a general theorem of Dudley, giving a strong upper bound on the suprema of subgaussian processes based on the covering number.

Theorem 12.13 (Dudley)**.** Let $\{X_t : t \in T\}$ be a separable, subgaussian process on a metric space (T, d). Then we have

$$\mathbb{E}[\sup_{t \in T} X_t] \leq 6 \sum_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}.$$

13. Lecture 13: October 21

13.1. **Dudley's Theorem: Proof and Implications.** Recall the statement of Dudley's Theorem, Theorem 12.13, from the previous class. In this class, we will prove the theorem and provide some applications.

First, we have the following lemma, which is immediate from working with volumes in a rough sense. It is not directly necessary for the proof of Dudley, but gives a way to estimate the upper bound resulting from Dudley.

Lemma 13.1. Let $B_2^n = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ be the ball in n dimensions. Then

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(B_2^n, \|\cdot\|, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^n.$$

Proof. We first show the upper bound. Let D be a $2\epsilon-$packing of B_2^n. Then note that the sets $\{B_\epsilon(t) : t \in D\}$ of radius ϵ spheres centered at the points in D are disjoint, by definition. Moreover, clearly $\bigcup_{t\in D} B_\epsilon(t) \subseteq B_{1+\epsilon}(0)$. Let μ denote the Lebesgue measure; we find that

$$\mu(B_{1+\epsilon}(0)) \geq \mu(\bigcup_{t\in D} B_\epsilon(t)) = |D|\mu(B_\epsilon(0)).$$

It follows that

$$|D|\epsilon^n \leq (1+\epsilon)^n \implies |D| \leq \left(1 + \frac{1}{\epsilon}\right)^n$$

The upper bound follows by replacing 2ϵ with ϵ.

Now we show the lower bound. Let $\mathcal{N}$ be an $\epsilon-$net of B_2^n. Then note that

$$\mu(B_2^n) \leq \mu(\bigcup_{t\in\mathcal{N}} B_\epsilon(t)) \leq |\mathcal{N}|\mu(B_\epsilon(0)).$$

It follows that

$$|\mathcal{N}| \geq \left(\frac{1}{\epsilon}\right)^n.$$

The lower bound follows. □

Now, we can finally move onto the proof of Theorem 12.13.

Proof of Theorem 12.13. First we will address the finite case, where $|T| < \infty$; the idea is to consider consecutive approximations to X_t via elements in the $2^{-k}-$nets, write $\sup X_t$ as a summation of the approximation residuals, and bound each term in the summation via Corollary 12.3, which holds for finite suprema of subgaussian variables. Then, intuitively we will only need to justify taking limits for the infinite case.

Let k_0 be the largest integer such that $2^{-k_0} > \operatorname{diam}(T)$. Note that for any $t_0 \in T$, $\{t_0\} = \mathcal{N}_{k_0}$ is a $2^{-k_0}-$net of T; pick some fixed such t_0. For any $k > k_0$, let N_k be a $2^{-k}-$net such that $|N_k| = \mathcal{N}(T, d, 2^{-k})$.

Now, note that by subadditivity of the supremum, we have the inequality

$$\mathbb{E}[\sup_{t\in T} X_t] \leq \mathbb{E}[X_{t_0}] + \sum_{k=k_0+1}^{n} \mathbb{E}[\sup_{t\in T}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)})] + \mathbb{E}[\sup_{t\in T}(X_t - X_{\pi_n(t)})],$$

where $\pi_k(t)$ is the closest element of $\mathcal{N}_k$ to t.[15]

To simplify this expression, note that $\mathbb{E}X_{t_0} = 0$. Moreover, observe that since $|T| < \infty$, there exists some sufficiently large n_0 such that $|\mathcal{N}_{n_0}| = T$. Therefore, for all $n > n_0$, the last term in the above expression also equals 0.

Now, the idea is to apply Corollary 12.3 to Lemma 12.2. Indeed, for each term in the summation, note that the supremum is over at most $|\mathcal{N}_k| \cdot |\mathcal{N}_{k-1}| \leq |\mathcal{N}_k|^2$ terms. Moreover, because X_t is a subgaussian process, we see that $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ is subgaussian with parameter

$$d(\pi_k(t), \pi_{k-1}(t)) \leq d(t, \pi_k(t)) + d(t, \pi_{k-1}(t)) \leq 2^{-k} + 2^{-(k-1)} = 3 \cdot 2^{-k}.$$

[15] ties broken arbitrarily

Applying the aforementioned Corollary, we find that

$$\mathbb{E}[\sup_t\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\}] \le \sqrt{2(3\cdot 2^{-k})^2 \log|\mathcal{N}_k|^2} = 6\cdot 2^{-k}\sqrt{\log|\mathcal{N}_k|}.$$

We thus find that

$$\mathbb{E}[\sup_{t\in T} X_t] \le 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}_k} = 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})},$$

as desired.

Now, we will consider the general case, dropping the assumption that $|T| < \infty$. By separability, there exists a countable set $T_0 \subseteq T$ such that $\sup_{t\in T} X_t = \sup_{t\in T_0} X_t$ almost surely. Write $T_0 = \{v_1, v_2, \ldots\}$ and its $k-$th truncation $T_0^k = \{v_1, \ldots, v_k\}$. By MCT[16]we find that

$$\mathbb{E}[\sup_{t\in T} X_t] = \mathbb{E}[\sup_{t\in T_0} X_t] = \sup_{k\ge 1}\mathbb{E}[\sup_{t\in T_0^k} X_t].$$

Now, recall that

$$\begin{aligned}\mathbb{E}[\sup_{t\in T_0^k} X_t] &\le 6\sum_{\ell\in\mathbb{Z}} 2^{-\ell}\sqrt{\log \mathcal{N}(T_0^k,d,2^{-\ell})}\\ &\le 6\sum_{\ell\in\mathbb{Z}} 2^{-\ell}\sqrt{\log \mathcal{N}(T,d,2^{-\ell})}.\end{aligned}$$

Combining these inequalities, we find that

$$\mathbb{E}[\sup_{t\in T} X_t] \le 6\sum_{\ell\in\mathbb{Z}} 2^{-\ell}\sqrt{\log \mathcal{N}(T,d,2^{-\ell})},$$

as desired. This completes the proof of the theorem. □

Note that the 2^{-k} is a reasonable but relatively specific choice. For a more natural statement, we have the following immediate corollary.

Corollary 13.2. Let $\{X_t : t\in T\}$ be a separable subgaussian process. Then

$$\mathbb{E}[\sup_{t\in T} X_t] \le 12\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\epsilon)}\mathrm{d}\epsilon.$$

Proof Sketch. This is merely a continuous form of the discrete variant above, where 12 is exactly the constant needed. Note that $\mathcal{N}(T,d,\epsilon)$ is decreasing in ϵ, so in the range $\epsilon \in (2^{-\ell}, 2^{-(\ell-1)})$ we can lower bound this portion of the integral by

$$12\cdot 2^{-\ell}\sqrt{\log \mathcal{N}(T,d,2^{-(\ell-1)})} = 6\cdot 2^{-(\ell-1)}\sqrt{\log \mathcal{N}(T,d,2^{-(\ell-1)})}.$$

[16]In theory, this is not quite right, because the variables are not necessarily nonnegative. However, X_t have subgaussian tails. So first the negative tail of all our random variables is uniformly bounded in magnitude by $X_{v_1}^-$ which is absolutely integrable. Now, truncating X_t at $-N$, we know that $\mathbb{P}(\sup_{t\in T_0^k} X_t > -N) \to 0$ uniformly as $N\to\infty$. We can apply MCT on any RV bounded below (by adding the lower-bound to make it nonnegative). More rigorously, we can find $\mathbb{E}[\sup_{t\in T_0} X_t \mathbb{I}(\sup_{t\in T_0} > -N)] = \sup_{k\ge 1}\mathbb{E}[\sup_{t\in T_0^k} X_t\mathbb{I}(\sup_{t\in T_0^k} > -N)]$. The main point is that either the RHS of the desired bound is infinite, in which case there is nothing to prove, or we find that $\mathbb{E}[(\sup_{t\in T} X_t)^+]$ is bounded by setting $N=0$ and recalling that the negative tail is bounded. It follows that the integrand on the LHS has absolutely integrable positive part. Hence, the integrands on the RHS are all bounded by the LHS on its positive part and X_{v_1} on its negative part; hence DCT then completes the proof.

Summing across the discrete regions $\epsilon \in (2^{-\ell}, 2^{-(\ell-1)})$ will yield the result. $\square$

We have now formally established a nice bound on the suprema of Gaussian processes in a metric space. Let us now look at an application.

13.2. **Uniform Laws of Large Numbers.** We aim to prove a uniform version of the standard LLN. Let $X_1, \dots, X_n \overset{iid}{\sim} \mu$ where μ is a distribution supported on $[0,1]$. Let $\mu_n = \frac{1}{n}\sum_{k=1}^n \delta_{X_k}$ be the empirical distribution. Then we have

$$W_1(\mu_n, \mu) \xrightarrow{p} 0.$$

Here, recall that W_1 is the Wasserstein 1-distance, given via Kantorovich-Rubinstein duality by

$$W_1(\mu, \nu) = \sup_{f \in \text{Lip}(1)} \left(\int f d\mu - \int f d\nu \right).$$

Here Lip(1) is the set of 1-Lipschitz functions.

Now, note that $\sup \int f \mathrm{d}\mu_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and hence $\int f \mathrm{d}\mu_n - \int f \mathrm{d}\mu = \frac{1}{n}\sum_{i=1}^n [f(X_i) - \mathbb{E}f(X_i)]$. We already know from the WLLN that

$$\int f \mathrm{d}\mu_n - \int f \mathrm{d}\mu \xrightarrow{P} 0.$$

Now the claim is that this statement is uniform over all f.

Proof of the above claim. Let $X_f = \frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}_\mu[f(X_i)])$ for ease of notation. First, by Azuma-Hoeffding, we have

$$\mathbb{E}[e^{\lambda(X_f - X_g)}] \le \exp(\frac{\lambda^2 \|f-g\|_\infty^2}{n}).$$

Now, note that X_f remains unchanged if we subtract a constant from f; therefore, we can restrict the supremum in the expression for W_1 to all $f \in \text{Lip}(1)$, $0 \le f \le 1$ by subtracting $\inf f$.

Now, look at $\{X_f : f \in [\text{Lip}(1), 0 \le f \le 1] = \mathcal{F}\}$. This is a subgaussian process with respect to the metric on the relevant *function space* $\mathcal{F}$, in particular $d(f,g) = \frac{\|f-g\|_\infty}{\sqrt{n}}$. It is immediate that this space is separable. By the corollary to Dudley's Theorem, we find that

$$\mathbb{E}[W_1(\mu_n, \mu)] \le 12 \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, d, \epsilon)} \mathrm{d}\epsilon.$$

Moreover, remark that by scaling, $\mathcal{N}(T, d/c, \epsilon) = \mathcal{N}(T, d, c\epsilon)$ for $c > 0$. Moreover, since (T,d) is a bounded metric space, for large enough n[17] we see that

$$12 \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, d, \sqrt{n}\epsilon)} \mathrm{d}\epsilon = 12 \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \sqrt{n}\epsilon)} \mathrm{d}\epsilon = \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon)} \mathrm{d}\epsilon.$$

Now, it suffices to bound the covering number. Break the x- and y-axes into intervals I_k, J_ℓ of length $\frac{\epsilon}{2}$, ϵ, respectively. For $x \in I_k$, let $\pi(f)(x) = \frac{\max J_\ell + \min J_\ell}{2}$, where $f(\min I_k) \in J_\ell$. We

[17] actually only $n > 2$ is needed

find that

$$\begin{aligned}|f(x)-\pi(f)(x)| &\le |f(x)-f(\min I_k)| + |f(\min I_k) - \frac{\max J_\ell + \min J_\ell}{2}| \\ &\le |x - \min I_k| + \frac{\max J_\ell - \min J_\ell}{2} \\ &\le \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= \epsilon.\end{aligned}$$

Here we have used the fact that f is 1-Lipschitz. Thus we have indeed constructed an ϵ-net. Naively, we already have by this construction and no further analysis that

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \le \left(\frac{1}{\epsilon}\right)^{C/\epsilon}.$$

In practice, one can see that the values assigned to consecutive intervals cannot differ by more than $\frac{3}{2}\epsilon$, hence we can attain a better bound of $\frac{1}{\epsilon}\cdot 3^{1/\epsilon}$.[18] Anyways, combining this with our previous inequalities, we find that

$$\mathbb{E}[W_1(\mu_n, \mu)] \le 12 n^{-\frac{1}{2}} \int_0^1 \sqrt{\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right)} \le \mathcal{O}(n^{-\frac{1}{2}}).$$

This completes the proof. □

To summarize, we have established a quantitative uniform LLN with guarantees that the 1-Wasserstein distance between the empirical CDF and the true distribution decays at a rate of $n^{-\frac{1}{2}}$. This is a very strong, and very nice result! In fact, this decay rate also turns out to be asymptotically optimal, showing the strength of Dudley.[19]

14. Lecture 14: October 23

We first finished discussing and proving the application of Dudley to uniform laws of large numbers.[20]

We are now finished with our discussion of suprema of Gaussian processes, via Gaussian comparison and metric geometry. We now move onto a discussion of *universality*.

14.1. **Universality.** We are interested in the following setup. Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ be vectors with independent coordinates, and $f : \mathbb{R}^n \to \mathbb{R}$ be a 'nice' function, which we will describe later. The question is: *when is* $f(X) \overset{d}{\approx} f(Y)$*?*

The motivation for and most classical example of this is of course the central limit theorem (CLT).

Example 14.1 (CLT)**.** In our setup, let $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$, $\mathbb{E}Y_i = 0$, $\mathbb{E}Y_i^2 = 1$ where Y_i are normal. Then we know that

$$\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \overset{d}{\approx} \frac{\sum_{i=1}^n Y_i}{\sqrt{n}} \overset{d}{=} \mathcal{N}(0,1).$$

[18] See [H18] chapter 5, lemma 5.16 for the details.

[19] See [H18] chapter 5 for a brief discussion.

[20] This section is short because this discussion was the majority of lecture, and was placed in the previous section for readability.

Another example is the Bai-Yin law.

Example 14.2 (Random Matrices and the Bai-Yin law)**.** Let X be a symmetric $n \times n$ random matrix, where $X_{ij} \overset{iid}{\sim} F$. Suppose that F for $i \leq j$, $\mathbb{E}X_{ij} = 0, \mathbb{E}X_{ij}^2 = 1$, and $\mathbb{E}|X_{ij}|^4 < \infty$. Then

$$\frac{\lambda_{\max}(X)}{\sqrt{n}} \xrightarrow{P} 2$$

In the next lecture, we will continue with result which formalize this intuition.

15. Lecture 15: October 28

We continue our discussion with universality-type results. Recall our setup: we have two random variables X, Y which have independent coordinates, and some function $f : \mathbb{R}^n \to \mathbb{R}$, and we would like to establish some type of result like $f(X) \overset{d}{\approx} f(Y)$.

What is a common case where this may occur? First, we might want to impose the conditions $\mathbb{E}X_i = \mathbb{E}Y_i$ and $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, otherwise by the CLT, we will not obtain approximate distributional equality. Now, intuitively, we also want that f does not depend strongly on any individual coordinate. In this setting, we will find that $f(X) \overset{d}{\approx} f(Y)$.[21] In particular, we have the following theorem.

Theorem 15.1. [Lindeberg Principle] Let X, Y be random vectors in $\mathbb{R}^n$ with independent coordinates. Assume that $\mathbb{E}X_i = \mathbb{E}Y_i$, $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$ for all $1 \leq i \leq n$. Then for any thrice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \leq \frac{1}{6}\left(\sum_{i=1}^{n} \|\frac{\partial^3 f}{\partial x_i^3}\|_\infty\right) \cdot \left[\mathbb{E}|X_i|^3 + \mathbb{E}|Y_i|^3\right].$$

Remark 15.2. Generally, we can expect two standard applications. We give these now.

First, in the special case where $f(X) \approx c$ is constant with high probability. It then follows immediately that $f(y) \approx c$ is constant with high probability, in that the third derivative term should be small.

More generally, suppose we expect $f(x) \overset{d}{\approx} Z$. Applying Theorem 15.1

In order to prove Theorem 15.1, we will first need the following lemma.

Lemma 15.3. Let X, Y be as in Theorem 15.1. Then

$$|\mathbb{E}f(X) - \mathbb{E}f(Y)| \leq \frac{1}{6}\|f'''\|_\infty \cdot \left[\mathbb{E}|X|^3 + \mathbb{E}|Y|^3\right].$$

Proof. The proof is just by Taylor expansion. We thus know that

$$|f(X) - f(0) - f'(0)X - \frac{1}{2}f''(0)X^2| \leq \frac{1}{6}\|f'''\|_\infty |X|^3$$

and

$$|f(Y) - f(0) - f'(0)Y - \frac{1}{2}f''(0)Y^2| \leq \frac{1}{6}\|f'''\|_\infty |Y|^3.$$

[21] We can expect some higher-order moments to match with lower-order terms. The theorem following this will be proven via Taylor-expansion methods. Moreover, there is actually nothing special out matching the first two moments. In general, one can obtain results for matching any number of moment terms, but in practice, the first two is the most common setting which arises.

We thus get bounds on $|f(X)|$, $|f(Y)|$. In particular, we find that

$$f(X) - f(Y) = f'(0)(X-Y) + \frac{1}{2}f''(0)(X^2 - Y^2) + C$$

for $|C| \le \frac{1}{6}\|f'''\|_\infty(|X|^3 + |Y|^3)$. Taking expectations and applying the triangle inequality, we find that

$$\begin{aligned}|\mathbb{E}f(X) - \mathbb{E}f(Y)| &= |f'(0)\mathbb{E}(X-Y) + \frac{1}{2}f''(0)\mathbb{E}(X^2-Y^2) + \mathbb{E}C| \\ &\le \mathbb{E}|C| \\ &\le \frac{1}{6}\|f'''\|_\infty \cdot (\mathbb{E}|X|^3 + \mathbb{E}|Y|^3).\end{aligned}$$

This completes the proof. □

We now prove Theorem 15.1.

Proof of Theorem 15.1. The idea is that by Lemma 15.3, we can slowly 'interpolate' from X to Y, changing one coordinate at a time. That is, define intermediate variables $Z_i = (X_i, \dots, X_i, Y_{i+1}, \dots, Y_n$. Then $Z_0 = Y$, $Z_n = X$, and transitioning from Z_i to Z_{i+1} changes only one coordinate. Now, we find that

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| = \left|\sum_{i=1}^n \left\{\mathbb{E}[f(Z^i)] - \mathbb{E}[f(Z^{i-1}]\right\}\right|,$$

so it suffices to derive a bound on each of the individual differences. However, conditional on all the variables other than X_i, Y_i for a fixed i, we can now apply Lemma 15.3 to the $i-$th difference in the above summation, since it is now only a function of X_i, Y_i. Formally, let $\mathcal{F}^{(i)} = \sigma(X_j, Y_j, j \ne i)$. Then by Lemma 15.3, we have

$$\left|\mathbb{E}[f(Z^i) \mid \mathcal{F}^{(i)}] - \mathbb{E}[f(Z^{i-1}) \mid \mathcal{F}^{(i)}]\right| \le \frac{1}{6}\|\frac{\partial^3 f}{\partial x_i^3}\|_\infty \cdot \left[\mathbb{E}|X_i|^3 + \mathbb{E}|Y_i|^3\right].$$

We thus find via this bound and Jensen's inequality that

$$\begin{aligned}|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| &\le \sum_{i=1}^n \mathbb{E}\left|\mathbb{E}[Z^i \mid \mathcal{F}^{(i)}] - \mathbb{E}[Z^{i-1} \mid \mathcal{F}^{(i-1)}]\right| \\ &\le \frac{1}{6}\sum_{i=1}^n \|\frac{\partial^3 f}{\partial x_i^3}\|_\infty \cdot (\mathbb{E}|X_i|^3 + \mathbb{E}|Y_i|^3),\end{aligned}$$

as desired. □

We now discuss some applications. Historically, Theorem 15.1 arose with Lindeberg's method of proving the CLT.

Example 15.4 (Proof of the CLT)**.** It suffices to show that $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \overset{d}{\approx} \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$, since then one can just take Y_i to be normal (in which case we obtain exactly the desired CLT behavior). Now, it suffices to show that for all bounded, continuous test functions g, we have that

$$\left|\mathbb{E}g\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\right) - \mathbb{E}g\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\right)\right|$$

is small. Roughly, it is sufficient to take bounded, smooth function g with bounded third derivative $\|g'''\|_\infty \le 1$.[22] But by Theorem 15.1, we know that

$$\left|\mathbb{E}g\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\right) - \mathbb{E}g\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\right)\right| \le \frac{1}{6}\frac{1}{n^{\frac{3}{2}}}\sum_{i=1}^n \|g'''\|_\infty \cdot [\mathbb{E}|X_i|^3 + \mathbb{E}|Y_i|^3].$$

Note that the extra factor of $n^{\frac{3}{2}}$ arises because we are applying the result to the composition $g(\frac{1}{\sqrt{n}}\sum X_i)$. Now, the right hand side will be of order $\frac{1}{\sqrt{n}}$. We thus obtain a uniform bound on this difference! Therefore, we get an *quantitative* bound on CLT.

Note that the bounded continuous test functions g will also approximate the bounded measurable functions. Then take $g(T) = \mathbb{I}(T \in A)$ for some measurable A, and apply the bound above. Take a supremum over all A, we then find that

$$d_{\text{TV}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i, \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\right) \le \mathcal{O}(n^{-\frac{1}{2}}).$$

So this is a first step in results of the Berry-Esseen type, attained quite easily!

We now discuss a deeper example, namely universality in the setting of the prominent SK model in statistical physics.

15.1. **General State of the Sherrington-Kirkpatrick (SK) Model.** We first give the setup. Take $J_{ij} \overset{iid}{\sim} \mu$, $\mathbb{E}J_{ij} = 0$, $\mathbb{E}J_{ij} = 1$ for all $i < j$. Then let

$$Z_n = \max_{\sigma\in\{\pm1\}^n} \sum_{i<j} \frac{J_{ij}}{\sqrt{n}}\sigma_i\sigma_j.$$

Intuitively, consider n points in space. Then J_{ij} represents some 'weights', and the $\sigma_i \in \pm 1$ are usually the magnetic spins; effectively one is miniziming some interpretation of energy. [23] We then have the following claim:

Claim 15.5. Z_n is universal to the law of J_{ij}.

More formally, let

$$Z_n^{(1)} = \frac{1}{\sqrt{n}} \max_{\sigma\in\{\pm1\}^n} \sum_{i<j} X_{ij}\sigma_i\sigma_j$$

and

$$Z_n^{(2)} = \frac{1}{\sqrt{n}} \max_{\sigma\in\{\pm1\}^n} \sum_{i<j} Y_{ij}\sigma_i\sigma_j,$$

then we claim that

$$\frac{1}{n}\left|\mathbb{E}[Z_n^{(1)}] - \mathbb{E}[Z_n^{(2)}]\right| \to 0.$$

[22] It turns out that this is also enough to show convergence in distribution, via approximation results of bounded, smooth functions.

[23] A non-physics intuition, given in class, is to consider people in two groups with certain affinities, and determining how to maximize friendship.

This is enough, because the $Z_n^{(i)}$ will be roughly of order n, so this result states that these two normalized expectations, which are $\mathcal{O}(1)$, are close.[24] Now, we have a discrete maximum. Therefore, we will approximate this with a smooth max, via a log-sum-exp. In particular, for $\beta > 0$, define

$$F_n^{(1)}(\beta) = \frac{1}{n\beta}\mathbb{E}\log\sum_{\sigma\in\{\pm 1\}} e^{\frac{\beta}{\sqrt{n}}\sum_{i<j} X_{ij}\sigma_i\sigma_j}, F_n^{(2)}(\beta) = \frac{1}{n\beta}\mathbb{E}\log\sum_{\sigma\in\{\pm 1\}} e^{\frac{\beta}{\sqrt{n}}\sum_{i<j} Y_{ij}\sigma_i\sigma_j},$$

By very crude bounding (e.g. either only taking one term in the summation or bounding all by their max), we find that

$$\frac{1}{n}\mathbb{E}[Z_n^{(j)}] \le F_n^{(j)}(\beta) \le \frac{1}{n}\mathbb{E}[Z_n^{(j)}] + \frac{\log 2}{\beta}.$$

Therefore, $F_n^{(j)}(\beta)$ is close to $\frac{1}{n}\mathbb{E}Z_n^{(j)}$. It therefore suffices to show that

$$|F_n^{(1)}(\beta) - F_n^{(2)}(\beta)| \overset{n\to\infty}{\longrightarrow} 0.$$

By the Lindeberg principle, we find that

$$|F_n^{(1)}(\beta) - F_n^{(2)}(\beta)| \le \frac{1}{6}\left(\sum_{i<j}\left\|\frac{\partial^3 F_n}{\partial X_{ij}^3}\right\|_\infty\right)\cdot(\mathbb{E}|X_{ij}^3| + \mathbb{E}|Y_{ij}^3|).$$

Now, consider the probability mass function[25] of σ, namely

$$\mu(\sigma) = \frac{e^{\frac{\beta}{\sqrt{n}}\sum_{i<j} X_{ij}\sigma_i\sigma_j}}{\sum_{\tau\in\{\pm 1\}^n} e^{\frac{\beta}{\sqrt{n}}\sum_{i<j} X_{ij}\sigma_i\sigma_j}}\ \forall\sigma\in(\pm 1)^n.$$

Sequentially expanding the partial derivatives, we find that

$$\frac{\partial F_n}{\partial X_{ij}} = \frac{1}{n\beta}\frac{e^{\frac{\beta}{\sqrt{n}}\sum_{i<j} X_{ij}\sigma_i\sigma_j}}{\sum_{\tau\in\{\pm 1\}^n} e^{\frac{\beta}{\sqrt{n}}\sum_{i<j} X_{ij}\sigma_i\sigma_j}} = n^{-\frac{3}{2}}\mu(\sigma_i\sigma_j),$$

$$\frac{\partial^2 F_n}{\partial X_{ij}^2} = \frac{\beta}{n^2}\mathrm{Var}_\mu(\sigma_i\sigma_j),$$

and

$$\frac{\partial^3 F_n}{\partial X_{ij}^3} = \frac{\beta^2}{n^{\frac{5}{2}}}\mu((\sigma_i\sigma_j - \mu(\sigma_i\sigma_j)^3).$$

We thus find that

$$\left\|\frac{\partial^3 F_n}{\partial X_{ij}^3}\right\|_\infty \le \frac{\beta^2}{n^{\frac{5}{2}}}.$$

[24]From problem 4 of problem set 1, we find that the fluctuations of $Z_n^{(j)}$ are of order 1, so they will be small; in other words, we have two variables concentrated with exponentially small tails at their expectations, which are close. It is not totally obvious whether their tails are identical, but they are certainly small and comparably negligible.

[25]In this context, the formula below is sometimes referred to as the *Hamiltonian*.

Plugging this back into the Lindeberg bound above, we conclude that

$$\begin{aligned}|F_n^{(1)}(\beta) - F_n^{(2)}(\beta)| &\le \frac{1}{6}\beta^2 n^{-\frac{5}{2}} \sum_{i<j} (\mathbb{E}|X_{ij}|^3 + \mathbb{E}|Y_{ij}|^3) \\ &\le \frac{C\beta^2}{\sqrt{n}} \max_{i<j} \mathbb{E}|X_{ij}|^3 + \mathbb{E}|Y_{ij}|^3 \\ &\overset{n\to\infty}{\longrightarrow} 0,\end{aligned}$$

where convergence happens provided that $\max_{i<j} \mathbb{E}|X_{ij}|^3 + \mathbb{E}|Y_{ij}|^3 < o(\sqrt{n})$. By taking $n \to \infty$, and then $\beta \to \infty$, the result follows.

Remark 15.6. This is a very nice and strong result: we essentially attain that the expectations

$$n^{-\frac{3}{2}} \mathbb{E}\left[\max_{\sigma\in\{\pm 1\}^n} \sum_{i<j} X_{ij}\sigma_i\sigma_j\right]$$

are close independent of the choice of X. It turns out we choose $J_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$, then actually

$$n^{-\frac{3}{2}}\left[\max_{\sigma\in\{\pm 1\}^n} \sum_{i<j} J_{ij}\sigma_i\sigma_j\right] \overset{a.s.}{\to} P_* \approx 0.7.$$

This is a very deep result; it was first 'shown' by Parisi, via heuristic methods in statistical physics, and since then formally by Talagrand.[26] Nonetheless, despite not knowing the limiting value at all, we can obtain some sort of universality result!

16. Lecture 16: October 30

We first completed the discussion of the application of the Lindeberg principle to universality in the SK model. We now discuss another example, namely universality in regularized regression.

16.1. **Universality for Regularized Regression.** The work for this section comes from [HS23]. Recall the setup of regularized regression: we have some linear model

$$y = Ax_0 + \epsilon, A \in \mathbb{R}^{m\times n}, y \in \mathbb{R}^m, x \in \mathbb{R}^n,$$

and we fit with a regularized loss function

$$\hat{x} = \arg\min\left\{\frac{1}{2}\|y - Ax\|^2 + \sum_{i=1}^n \ell(x_i)\right\}.$$

Recall that the CGMT (Theorem 10.5) provides already sharp characterizations on the loss and therefore the behavior of $\hat{x}$. However, the CGMT critically relies on the assumption that A_{ij} are iid normal. While normality is a reasonable assumption, we would like this type of result to be at least somewhat robust to the distribution of A; otherwise, the results would not be as applicable.

In fact, we have the following result.

[26] In fact, it has been cited in their respective Nobel, Abel prizes, as significant contributions.

Theorem 16.1. [Corollary 2.6 in [HS23]] Suppose that $\tau \in (0,1)$ is a parameter such that $\tau \le \frac{m}{n} \le \frac{1}{\tau}$. Moreover, define

$$X(u,w,A) = u^\intercal A w + Q(u,w),$$

for $u \in \mathbb{R}^m, w \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}, Q : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$. Let $A_0, B_0 \in \mathbb{R}^{m\times n}$ be matrices with independent entries of mean 0 and variance 1, and let

$$M_0 = \max_{i,j}\{\mathbb{E}|A_{0,ij}|^3 + \mathbb{E}|B_{0,ij}|^3\}$$

be the Lindeberg upper bound term. Lastly, take the normalizations $A = \frac{A_0}{\sqrt{m}}$, $B = \frac{B_0}{\sqrt{m}}$.

Then there exists some $c_0 = c_0(\tau, M_0) > 0$ such that for any $\mathcal{S}_u \subseteq [-L_u, L_u]^m, \mathcal{S}_w \subseteq [-L_w, L_w]^n$, $L_u, L_w \ge 1$, and $g \in C^3(\mathbb{R})$, we have the inequality

$$\left|\mathbb{E}g(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X(u,w,A)) - \mathbb{E}g(\max_{u\in\mathcal{S}_u}\min_{w\in\mathcal{S}_w} X(u,w,B))\right| \le C_0 K_g \mathfrak{r}_n,$$

where

$$\mathfrak{r}_n = \inf_{0\le\delta\le n^{-1}}\left\{M_Q(L,\delta) + \frac{L^2\log^{\frac{2}{3}}\left(\frac{L}{\delta}\right)}{n^{\frac{1}{6}}}\right\},$$

$$K_g = 1 + \max_{\ell\in\{0,1,2,3\}}\|g^{(\ell)}\|_\infty,$$

$L = L_u + L_w$, and $M_q(L,\delta)$ is the modulus of continuity of Q.[27] Moreover, the same conclusion holds if we swap max min with min max.[28]

Proof Sketch. The proof uses similar ideas as seen previously. The key idea is to apply the Lindeberg principle, Theorem 15.1, to a smoothed max min. The smoothing is done similarly to the example of the SK model; via a log sum exp (with temperature parameter). See Sections 4.4, 4.5 of [HS23] for the details. □

Intuitively, the idea of the theorem is the following. First, X is analogous to the expression present in the CGMT. Then, the idea is that for different test functions A, B, the expectations of a reasonable function will provide similar values.

Remark 16.2. Note that this theorem itself only gives universality guarantees for the loss. For instance, we obtain the following. Suppose two run LASSO with A of Gaussian entries, and LASSO with A of non-Gaussian entries matching the moments. Then Theorem 16.1 implies that the *loss* of these two regressions are similar.

Now, we ultimately want to obtain universality results for the minimizer; for instance, we would like to bound the difference between the minimizers of the above two regressions. It turns out universality of the loss in this general context in fact implies universality results of the minimizer. However, attaining this strong result requires significant extra work, which is a main part of the contribution of [HS23].

Second, note that another technical challenge is to actually prove L^∞ contraints on the minimizers. In particular, applications of Theorem 16.1 require the 'test space' $\mathcal{S}_u, \mathcal{S}_w$ to be

[27] That is, $M_q(L,\delta) = \sup|Q(u,w) - Q(u',w')|$, where the supremum is taken over $u, u' \in [-L,L]^m$, $w, w' \in [-L,L]^n$, and $\|u-u'\|_\infty, \|v-v'\|_\infty \le \delta$.

[28] The statement also holds if there exists some $S \subseteq [m]\times[n]$ such that $A_{ij} = B_{ij} = 0$ for $(i,j) \subseteq S$. That is, zeroing out some of the terms of the matrices does not affect the result.

rectangular sets. Another technical challenge the authors of [HS23] address is to show that the a priori unbounded regression problem is actually well-bounded.

Remark 16.3. We can also take the following different and recent perspective on this type of result. In particular, we can consider that in this setting, we have $\mathcal{O}(mn)$ iid entries, i.e. mn 'bits' of randomness. A natural question is: *can we handle more structured randomness?*

An important observation is that it turns out that structured matrices sometimes 'behaves' like random matrices. For a reference, see [DSL24]. The idea of their proofs is to show that *gradient descent* is universal. We can then run gradient descent for finite iterations to converge to the minimizer; universality of this *process* then implies that the minimizer is universal.

This wraps up our discussion on universality. We will now move onto discussing Stein's method as a tool for proving limit theorems.

17. Lecture 17: November 4

We now begin our discussion of Stein's Method. This technique originates from a paper by Charles Stein in 1972, and is surprisingly powerful in proving limit theorems. We will begin by discussing the prerequisites and simple examples. The strength of the method both lies in its ability to attain quantitative results and its ability to work with locally dependent random variables.

17.1. **Integral Probability Metrics.** First, in order to attain any sort of quantitative CLT results, we need to quantify the 'closeness' of two distributions. There are many ways to do this, but broadly, the idea is to define an *integral probability metric*, i.e. a distance metric between distributions. The most general way to do this is as follows.

Definition 17.1. Let μ, ν be two probability measures on a fixed probability space Ω. Let $\mathcal{H}$ be a set of measurable functions. Then we can define a distance metric

$$d_{\mathcal{H}}(\mu, \nu) = \sup_{h \in \mathcal{H}} \left| \int h \mathrm{d}\mu - \int h \mathrm{d}\nu \right|.$$

Example 17.2. Many specific choices give known and classical examples of distance:

(1) If $\mathcal{H} = \{\mathbb{I}(\cdot \leq x) \forall x \in \mathbb{R}\}$, then we obtain the *Kolmogorov metric* d_K.
(2) If $\mathcal{H}$ is the set of 1-Lipschitz functions, then we obtain the *1-Wasserstein metric* d_W.
(3) If $\mathcal{H}$ is the set of all indicators, then we obtain the *total variation (TV) distance* d_{TV}.

These distance metrics turn out to all be similar. For example, we have the following lemma.

Lemma 17.3. We have the following:

(1) $d_K(W, Z) \leq d_{\mathrm{TV}}(W, Z)$, and
(2) $d_K(W, Z) \leq \sqrt{2C d_W(W, Z)}$ if Z has PDF bounded by C.

Therefore, it will suffice to bound the Wasserstein distance, which we will do via tools based on *Stein's Lemma*, which we have actually seen! It is Lemma 4.5.

17.2. **Stein's Lemma.** At the heart of Stein's Method is *Stein's Lemma*. Part of this result was stated and proven in an earlier lecture, namely as Lemma 4.5, under the context of using it in Gaussian interpolation. However, we will need more than just the statement there.

Lemma 17.4 (Stein's Lemma, Full Version)**.** Define $\mathcal{A}$ to be the operator

$$\mathcal{A}f(x) = f'(x) - xf(x).$$

Then:

(1) If $Z \sim \mathcal{N}(0,1)$, then $\mathbb{E}[\mathcal{A}(f(Z))] = 0$ for all absolutely continuous[29] functions f with $\mathbb{E}|f'(Z)| < \infty$.
(2) The converse holds: if $\mathbb{E}[\mathcal{A}(f(Z))] = 0$ for all absolutely continuous f with $\mathbb{E}|f'(Z)| < \infty$, then $Z \sim \mathcal{N}(0,1)$.

The first statement is just Lemma 4.5; for the second, we will need another lemma.

Lemma 17.5 (Choice of f)**.** Let Φ be the standard normal CDF. Then the differential equation

$$f_x'(w) - wf_x(w) = \mathbb{I}(w \le x) - \Phi(x)$$

has a unique *bounded* solution, of the form

$$\begin{aligned} f_x(w) &= e^{w^2/2} \int_w^\infty e^{-t^2/2}(\Phi(x) - \mathbb{I}[t \le x])\, dt \\ &= -e^{w^2/2} \int_{-\infty}^w e^{-t^2/2}(\Phi(x) - \mathbb{I}[t \le x])\, dt. \end{aligned}$$

Moreover, f_x is smooth and satisfies $\|f_x\|_\infty \le 1, \|f_x'\|_\infty \le \sqrt{\frac{2}{\pi}}, \|f_x''\|_\infty \le 2$.

Proof. See [Ros11]; one applies the method of integrating factors and then there is some subsequent computation. The point is that the left hand side can be written as a derivative via product rule with $e^{-\frac{w^2}{2}}$. The bounds can be proven explicitly. □

Now we can give the full proof of Lemma 17.4.

Proof of Lemma 17.4. Statement (1) was already proven in Lemma 4.5. To show statement (2), we apply the choice of $f = f_x$ given by Section 18.1. The point is that f_x satisfies the necessary conditions, and is designed such that $\mathbb{E}\mathcal{A}(f)$ is exactly the difference between Φ and the CDF of Z. In particular, we have that

$$0 = \mathbb{E}[f_x'(Z) - Zf_x(Z)] = \mathbb{P}(Z \le x) - \Phi(x).$$

This completes the proof. □

In other words, in order to prove Gaussianity of some random variable, it suffices to prove that $\mathbb{E}\mathcal{A}(f(Z)) = 0$ for the necessary family of test functions. This doesn't give an immediate quantitative bound; however, it gives us intuition for how we can try to bound the distance between Z and a standard normal; in particular, if $W \overset{d}{\approx} \mathcal{N}(0,1)$, then W should satisfy an 'approximate Stein's identity'. We will see this in the next lecture.

[29]That is, f' exists almost everywhere and $f(b) - f(a) = \int_a^b f' \mathrm{d}x$ (i.e. FTOC holds).

18. Lecture 18: November 6

18.1. **Stein's Method Continued.** We will continue our discussion of Stein's method. Recall from the last time we have from Lemma 17.4 that normality of Z is closely tied to the functional operator $\mathcal{A}$.

Now, we will give the formal idea of Stein's method. Suppose we are interested in proving that $W \overset{d}{\approx} \mathcal{N}(0,1)$. Now, suppose we can find a class $\mathcal{F}$ of functions such that

$$d_W(W,Z) = \sup_{g\in \mathrm{Lip}(1)} |\mathbb{E}g(W) - \mathbb{E}g(Z)| \leq \sup_{f\in\mathcal{F}} |\mathbb{E}[f'(W) - Wf(W)]|\,. \tag{3}$$

Then it suffices to show that W satisfies an 'approximate' Stein's identity.

Now, given some fixed $g \in \mathrm{Lip}(1)$, the easiest way to show Equation (3) is to find f_g such that

$$g(x) - \mathbb{E}g(Z) = f'(x) - xf(x).$$

Note that this differential equation is a generalization of Section 18.1.[30] In particular, we have the following stronger version of Section 18.1.

Lemma 18.1 (Choice of f_g)**.** Let $Z \sim \mathcal{N}(0,1)$ be the standard normal CDF. Then the differential equation

$$f_g'(w) - wf_g(w) = g(w - \mathbb{E}[g(Z)])$$

has a solution

$$\begin{aligned} f_g(w) &= e^{\frac{w^2}{2}} \int_w^\infty e^{-\frac{t^2}{2}} (\mathbb{E}g(Z) - g(t))\mathrm{d}t \\ &= -e^{\frac{w^2}{2}} \int_{-\infty}^w e^{-\frac{t^2}{2}} (\mathbb{E}g(Z) - g(t))\mathrm{d}t. \end{aligned}$$

Moreover,

(1) If g is bounded, we have

$$\|f_g\| \leq \sqrt{\frac{\pi}{2}} \|g(\cdot) - \mathbb{E}g(Z)\|, \text{ and } \|f_g\|' \leq 2\|g(\cdot) - \mathbb{E}g(Z)\|.$$

(2) If g is absolutely continuous, then

$$\|f_g\| \leq 2\|g'\|, \|f_g'\| \leq \sqrt{\frac{2}{\pi}} \|g'\|, \text{ and } \|f_g''\| \leq 2\|g'\|.$$

Proof. See [Ros11], which actually references another result. It is similar to the proof of but requires more technical details. □

From this lemma, it follows that by choosing

$$\mathcal{F} = \{f \in C^\infty, \|f\|_\infty \leq 1, \|f'\|_\infty \leq \sqrt{\frac{2}{\pi}}, \|f''\|_\infty \leq 2,$$

we then have $f_g \in \mathcal{F}$. Therefore, we find that Equation (3) holds with this choice of $\mathcal{F}$.

Now, we will discuss various techniques for applying Stein's method.

[30] Indeed, Section 18.1 is simply the special case $g = \mathbb{I}(\cdot \leq a)$.

18.2. **Techniques for applying Stein's Method.** The main step remaining in applying Stein's method is to prove an approximate Stein's identity. Two main techniques are the *leave one out method* and the *exchangeable pairs method*. We will give an example to demonstrate the main idea of the Leave-One-Out (LOO) technique, and then discuss the exchangeable pairs method, which is a bit more complicated, in more detail.

Example 18.2 (Leave-One-Out Technique for quantitative CLTs)**.** We will demonstrate the leave-one-out technique via proving a quantitative CLT bound. Recall the setup we have

$$W_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}}$$

as the usual sample mean, where X_i are independent (not necessarily IID!) random variables where $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$, $\mathbb{E}X_i^4 < \infty$, and let $Z \sim \mathcal{N}(0,1)$. We are interested in bounding $d_W(W, \mathcal{N}(0,1)$.

The intuition is as follows. First we can write $W_n f(W_n)$ as $\sum_i X_i f(W_n)$. Now, the difficulty is that X_i has some impact on W_n, so analyzing the expectation of this quantity is difficult. However, since each $\frac{X_i}{\sqrt{n}}$ does not contribute significantly to the sum, we can define $W_n^{-i} = W_n - \frac{X_i}{\sqrt{n}}$, and just replace W_n with W_n^{-i}. The main technical analysis then relies on Taylor expansion to bound the difference.

In particular, we have

$$\begin{aligned}
\mathbb{E}W_n f(W_n) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}[X_i f(W_n)] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}[X_i f(W_n^{-i})] + \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}[X_i (f(W_n) - f(W_n^{-i}))] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}[X_i (f(W_n) - f(W_n^{-i}) - (W_n - W_n^{-i})(f'(W_n^{-i}))] \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{E}[X_i (W_n - W_n^{-i}) f'(W_n^{-i})] \\
&= \tau_1 + \tau_2.
\end{aligned}$$

Now, the first term via a second-order Taylor expansion is bounded above by

$$\begin{aligned}
|\tau_1| &\le \frac{\|f''\|_\infty}{2\sqrt{n}} \sum_{i=1}^n \mathbb{E}|X_i (W_n - W_n^{-i})^2| \\
&\le \frac{1}{n^{\frac{3}{2}}} \sum_{i=1}^n \mathbb{E}|X_i|^3.
\end{aligned}$$

Moreover, notice that $X_i(W_n - W_n^{-i}) = \frac{X_i^2}{\sqrt{n}}$ is independent of $f'(W_n^{-i})$, hence we can write

$$\begin{aligned}|\mathbb{E}f'(W_n) - \tau_2| &= |Ef'(W_n) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}f'(W_n^{-i})| \\ &\leq \frac{1}{n}\sum_{i=1}^n |\mathbb{E}f'(W_n) - \mathbb{E}f'(W_n^{-i}) \\ &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}|(W_n - W_n^{-1})\|f''\|_\infty| \\ &\leq \frac{2}{n^{\frac{3}{2}}}\sum_{i=1}^n \mathbb{E}|X_i|.\end{aligned}$$

Combining everything, we find that

$$\begin{aligned}|\mathbb{E}[f'(W_n) - W_n f(W_n)]| &= |\mathbb{E}f'(W_n) - \tau_2 - \tau_1| \\ &\leq \frac{1}{n^{\frac{3}{2}}}\sum_{i=1}^n \mathbb{E}|X_i|^3 + \frac{2}{n^{\frac{3}{2}}}\sum_{i=1}^n \mathbb{E}|X_i|.\end{aligned}$$

Combined with earlier observations, this gives some quantitative CLT bound of the form

$$d_W(W_n, Z) \leq \frac{1}{n^{\frac{3}{2}}}\sum_{i=1}^n \mathbb{E}|X_i|^3 + \frac{2}{n^{\frac{3}{2}}}\sum_{i=1}^n \mathbb{E}|X_i|,$$

giving us a nice quantitative upper bound resembling the Berry-Esseen CLT.

Remark 18.3. This method generalizes nicely; for instance, we can write $W_n = \sum a_i X_i$; if a_i satisfies nice conditions, then we obtain a similar result.

We now move onto discussing the exchangeable pairs method.

18.3. **Exchangeable Pairs Method.** First, we need a few definitions.

Definition 18.4 (Exchangeable RVs)**.** We say a pair of random variables (W_i, W_i') is *exchangeable* if

$$(W_i, W_i') \sim (W_i', W_i).$$

Definition 18.5 (α-Stein Pair)**.** An exchangeable pair (W_i, W_i') is an *α-Stein pair* for some $\alpha \in (0, 1]$ if we have

$$\mathbb{E}[W' \mid W] = (1 - \alpha)W.$$

Theorem 18.6. [Exchangeable Pair Bound] Suppose (W, W') is an α-Stein exchangeable pair and $\mathbb{E}W^2 = 1$. Then

$$\begin{aligned}d_W(W, Z) &\leq \frac{\sqrt{\mathrm{Var}\left(\mathbb{E}\left[(W' - W)^2 \mid W\right]\right)}}{\sqrt{2\pi\alpha}} + \frac{\mathbb{E}|W - W'|^3}{3\alpha} \\ &= \tau_1 + \tau_2\end{aligned}$$

We will prove Theorem 18.6 in the next lecture after providing an example application to the classical CLT bound via moments.

Example 18.7 (Classical CLT)**.** The setup is the exact same as in the Leave-One-Out example; we aim to find an alternative proof. Let $(X_1', \cdots, X_n')$ be an independent copy of $(X_1, \ldots, X_n)$. Then we can define

$$W_n' = W_n - \frac{X_I}{\sqrt{n}} + \frac{X_I'}{\sqrt{n}}, \quad I \sim \text{Unif}[n].$$

Then clearly (W_n', W_n) is an exchangeable pair. Moreover, it is a $\frac{1}{n}$-Stein pair:

$$\begin{aligned}
\mathbb{E}(W_n' \mid W_n) &= \mathbb{E}\left(W_n - \frac{X_I}{\sqrt{n}} + \frac{X_I'}{\sqrt{n}} \mid W_n\right) \\
&= W_n - \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i - X_i' \mid W_n) \\
&= W_n - \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i \mid W_n) \\
&= W_n - \frac{1}{n} W_n.
\end{aligned}$$

Applying Theorem 18.6 with $\alpha = \frac{1}{n}$, we know that

$$d_W(W, Z) \le \tau_1 + \tau_2,$$

hence it suffices to control τ_1 and τ_2. The latter term is easier to bound; we have

$$\tau_2 = \frac{n}{3} \cdot \frac{1}{n^{\frac{3}{2}}} \mathbb{E}|X_I - X_I'|^3 = \frac{1}{3n^{\frac{3}{2}}} \sum_{i=1}^n \mathbb{E}|X_i - X_i'|^3 \le \frac{8}{3\sqrt{n}} \sum_{i=1}^n \mathbb{E}|X_i|^3,$$

where we have used the bound $|X - Y|^3 \le 8(|X|^3 + |Y|^3)$.

For the former term, note first that

$$\mathbb{E}\left[(W_n' - W_n)^2 \mid W_n\right] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}(X_i - X_i')^2 \mid W_n).$$

Therefore,

$$\begin{aligned}
\tau_1 &= \sqrt{\frac{n}{2\pi}} \sqrt{\text{Var}\mathbb{E}[(W_n - W_n')^2 \mid W_n]} \\
&= \sqrt{\frac{n}{2\pi}} \sqrt{\frac{1}{n^2} \text{Var}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}((X_i - X_i')^2 \mid W_n)\right)} \\
&= \sqrt{\frac{1}{2\pi n^3}} \text{Var}\left(\sum_{i=1}^n \mathbb{E}((X_i - X_i')^2 \mid W_n)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\frac{1}{2\pi n^3}} \mathrm{Var}\left(\sum_{i=1}^n [\mathbb{E}(X_i^2 \mid W_n) + 1 + 0] \right) \\
&= \sqrt{\frac{1}{2\pi n^3}} \mathrm{Var}\left(\sum_{i=1}^n \mathbb{E}(X_i^2 \mid W_n) \right) \\
&= \sqrt{\frac{1}{2\pi n^3}} \mathrm{Var}\left(\mathbb{E}(\sum_{i=1}^n X_i^2 \mid W_n) \right) \\
&\leq \sqrt{\frac{1}{2\pi n^3}} \mathrm{Var}\left(\frac{\sum_{i=1}^n X_i^2}{n} \right) \\
&\leq \sqrt{\frac{1}{2\pi n^3}} \sum_{i=1}^n \mathbb{E}|X_i|^4.
\end{aligned}
$$

Here, we have used the previous statements along with independence of X_i' and W_n (to justify line 4), and the Law of Total Variance (which justifies the second-to-last inequality). Combining everything, we find that

$$
d_W(W_n, Z) \leq \frac{1}{\sqrt{2\pi} n^{\frac{3}{2}}} \sqrt{\sum_{i=1}^n \mathbb{E}X_i^4} + \frac{8}{3n^{\frac{3}{2}}} \sum_{i=1}^n \mathbb{E}|X_i|^3.
$$

This completes the proof of a classical quantitative CLT bound.

19. Lecture 19: November 11

In this lecture, we first complete the discussion of the exchangeable pair method by proving Theorem 18.6.

Proof of Theorem 18.6. In line with Stein's method, let $f \in \mathcal{F}$ be an appropriate test function. Set

$$
F(w) = \int_0^w f(t) \mathrm{d}t.
$$

We aim to analyze $\mathbb{E}[f'(W) - Wf(W)]$. By exchangeability, $\mathbb{E}[F(W') - F(W)] = 0$. Taylor expanding $F(W')$ near W gives

$$
\mathbb{E}[(W' - W)f(W) + \frac{1}{2}(W' - W)^2 f'(W) + \frac{1}{6}(W' - W)^3 f''(W^*)] = 0
$$

for some $W^* \in [W, W']$. By the α-Stein property, we find that

$$
\mathbb{E}[(W' - W)f(W)] = \mathbb{E}[f(W)\mathbb{E}[W' - W \mid W]] = -\alpha \mathbb{E}[Wf(W)].
$$

It follows that

$$
\mathbb{E}[Wf(W)] = \mathbb{E}\left[\frac{(W' - W)^2 f'(W)}{2\alpha} + \frac{(W' - W)^3 f''(W^*)}{6\alpha} \right].
$$

Therefore, we have the bound

$$|\mathbb{E}[f'(W)] - \mathbb{E}[Wf(W)]| \le \|f'\|_\infty \mathbb{E}\left|1 - \frac{\mathbb{E}[(W'-W)^2 \mid W]}{2\alpha}\right| + \|f''\|_\infty \frac{\mathbb{E}|W'-W|^3}{6\alpha}$$
$$\le \sqrt{\frac{2}{\pi}}\mathbb{E}\left|1 - \frac{\mathbb{E}[(W'-W)^2 \mid W]}{2\alpha}\right| + \frac{\mathbb{E}|W'-W|^3}{3\alpha}.$$

This already resembles the desired inequality. We mainly need to analyze the first term; note that

$$\mathbb{E}[(W'-W)^2] = \mathbb{E}W'^2 + \mathbb{E}W^2 - 2\mathbb{E}[WW'] = 2\mathbb{E}W^2 - 2\mathbb{E}[W\mathbb{E}[W' \mid W]] = 2\alpha.$$

It follows that the first term above is

$$\mathbb{E}\left|1 - \frac{\mathbb{E}[(W'-W)^2 \mid W)]}{2\alpha}\right| = \frac{1}{2\alpha}\mathbb{E}\left|\mathbb{E}[(W'-W)^2] - \mathbb{E}[(W'-W)^2 \mid W)]\right|$$
$$\le \frac{1}{2\alpha}\sqrt{\mathrm{Var}(\mathbb{E}[(W'-W)^2 \mid W)]},$$

where the final inequality follows from Jensen's. Combining this with Lemma 18.1 and the result of Stein's method completes the proof. □

We now have a nice bound on the Wasserstein distance between W and a standard normal, provided that we can find W' such that (W, W') is an α-Stein pair. We have already seen one sample application; however, in general, it is not clear how to find such W'. We thus have the natural question:

Question 19.1. How can we effectively construct exchangeable pairs in general?

Here is one general, high-level idea. Suppose that W is defined on the probability space $\mathbb{P} = (\Omega, \mathcal{F}, \mu)$. Suppose we have some *Markov chain* $X_0, X_1, \ldots$ on Ω which is stationary and reversible with respect to μ. Then $W = W(X_0)$ and $W' = W(X_1)$ is an exchangeable pair!

Example 19.2 (Classical CLT, Revisited)**.** The example W' used in the classical CLT is motivated by the following Markov chain. First, take $(X_1, \ldots, X_n, X_1', \ldots, X_n') = (Y_1, \ldots, Y_{2n})$ as in the proof. Then, pick $I \sim \mathrm{Unif}([n])$. At a given step of the Markov chain, swap Y_I with Y_{n+I}. Then the (W, W') used in the example correspond to can be derived by taking one step of this Markov chain.

This completes our discussion of Stein's method and its various realizations.

20. Lecture 20: November 13

20.1. **Concentration via Exchangeable Pairs.** We now shift towards discussing *concentration* via exchangeable pairs. For a reference, see Sourav Chatterjee's PhD thesis [Cha05].

The main result is the following.

Theorem 20.1 (Exchangeable Pairs Concentration Inequality)**.** Let (X, X') be an exchangeable pair. Let $F : \mathbb{R}^2 \to \mathbb{R}$ be an *anti-symmetric* function; that is, $F(x, y) = -F(y, x)$. Let

$$f(x) = \mathbb{E}[F(X, X') \mid X = x]$$

be our function of interest and

$$v(x) = \frac{1}{2}\mathbb{E}\left[|(f(X) - f(X'))F(x, x')| \mid X = x\right]$$

be a *variance proxy* term. Then:

(1) We have $\mathbb{E}f(X) = 0$ and $\mathrm{Var}(f(X)) \le \mathbb{E}[v(X)]$.
(2) Suppose that $\mathbb{E}\left[e^{\theta f(X)}\,|F(X,X')|\right] < \infty$ for all $\theta > 0$. If there are constants B, C such that $v(x) \le Bx + c$, then

$$\mathbb{P}\left(|f(X)| > t\right) \le 2\exp\left(-\frac{t^2}{2(Bt+C)}\right).$$

Proof. We first show the first item. Note that

$$\mathbb{E}[f(X)] = \mathbb{E}[F(X,X')] = \mathbb{E}[F(X',X)] = -\mathbb{E}[F(X,X')],$$

where we have used exchangeability and the anti-symmetry of F. Thus $\mathbb{E}f(X) = 0$. Now, note that

$$\mathrm{Var}(f(X)) = \mathbb{E}[f(X)^2] = \mathbb{E}[f(X)F(X,X')] = \mathbb{E}[f(X')F(X',X)] = -\mathbb{E}[f(X)F(X,X')],$$

again via exchangeability and anti-symmetry. Thus

$$\mathrm{Var}(f(X)) = \frac{1}{2}\mathbb{E}[(f(X') - f(X))F(X,X')] \le \mathbb{E}[v(X)].$$

This proves both parts of the first item.

Now we show the second item, namely the tail bound. As usual, we will employ a Chernoff-type bound; the main difficulty is analyzing the MGF of $f(X)$. We now do this.

Let $m(\theta) = \mathbb{E}[e^{\theta f(X)}]$ be the MGF of $f(X)$. Then note that

$$m'(\theta) = \mathbb{E}[f(X)e^{\theta f(X)}] = \mathbb{E}[f(X,X')e^{\theta f(X)}] = \frac{1}{2}\mathbb{E}[F(X,X')(e^{\theta f(X)} - e^{\theta f(X')})],$$

where the second equality follows from exchangeability and anti-symmetry. The idea is now to bound m' using the variance proxy, which it somewhat resembles, then bound m accordingly.

Invoking the basic inequality

$$|e^x - e^y| \le \frac{1}{2}|x-y|(e^x + e^y)^{31},$$

we see that

$$\begin{aligned}
|m'(\theta)| &\le \frac{1}{2}\mathbb{E}[|F(X,X')||e^{\theta f(X)} - e^{\theta f(X')}] \\
&\le \frac{|\theta|}{4}\mathbb{E}[|F(X,X')(f(X) - f(X'))|(e^{\theta f(X)} + e^{\theta f(X')})] \\
&= \frac{|\theta|}{2}\mathbb{E}[|F(X,X')(f(X) - f(X'))|e^{\theta f(X)}] \\
&= \frac{|\theta|}{2}\mathbb{E}[e^{\theta f(X)}\mathbb{E}[|F(X,X')(f(X) - f(X'))| \mid X]] \\
&= |\theta|\mathbb{E}[v(X)e^{\theta f(X)}],
\end{aligned}$$

where we have applied the aforementioned inequality in the second line, exchangeability and antisymmetry in the third, and the Law of Total Expectation in the fourth.

[31] We can see this as follows. Let $x = y + c$ where WLOG $c \ge 0$; the inequality reduces to $e^c - 1 \le \frac{1}{2}c(e^c+1)$. Using the Taylor expansion of e^x, the coefficients of c^k on the LHS are $\frac{1}{k!}$, while the coefficients of c^k on the RHS are 1 for $k = 1$ and $\frac{1}{2(k-1)!}$ for $k \ge 2$. since $\frac{1}{2(k-1)!} \ge \frac{1}{k!}$ for $k \ge 2$, the result follows.

Now, note that $m'(\theta) \geq 0$ as well, since $m''(\theta) = \mathbb{E}[f(X)^2 e^{\theta f(X)}] \geq 0$ and $m(0) = 1$ and $m'(0) = \mathbb{E}f(X) = 0$. Moreover, given the variance proxy bound $v(x) \leq Bf(x) + C$, we find from the above inequality that

$$|m'(\theta)| \leq |\theta|[Bm'(\theta) + Cm(\theta)].$$

Therefore, for $0 < \theta < \frac{1}{B}$, we find that

$$m'(\theta) \leq \frac{C\theta m(\theta)}{1 - B\theta} \implies \frac{\mathrm{d}}{\mathrm{d}\theta} \log m(\theta) = \frac{m'(\theta)}{m(\theta)} \leq \frac{C\theta}{1 - B\theta}.$$

Thus, we have

$$\log m(\theta) \leq \int_0^\theta \frac{Ct}{1 - Bt} \mathrm{d}t \leq \frac{1}{1 - B\theta} \cdot \int_0^\theta Ct \mathrm{d}t \leq \frac{1}{1 - B\theta} \cdot \frac{C\theta^2}{2}.$$

Applying the Chernoff bound alluded to earlier, we find that

$$\mathbb{P}(f(X) > t) \leq \frac{\mathbb{E}e^{f(X)}}{e^{\theta t}} \leq \exp\left(-\theta t + \frac{C\theta^2}{2(1 - B\theta)}\right).$$

Minimizing the RHS, we can take $\theta = \theta^* = \frac{t}{C + Bt}$, from which we attain

$$\mathbb{P}(f(X) > t) \leq \exp\left(-\frac{t^2}{2(Bt + C)}\right).$$

The other tail ($\mathbb{P}(f(X) < -t)$) is exactly analogous. This completes the proof of the theorem. □

We now discuss an example application, namely to concentration results in the classical Curie-Weiss model in statistical physics.

20.2. **Concentration in the Curie-Weiss Model.** We first describe the setup. Let $\sigma \in \{\pm 1\}^n$ be an elemenet of the standard hypercube, and let

$$p(\sigma) = \frac{1}{Z_\beta} \exp\left(\frac{\beta}{n} \sum_{i<j} \sigma_i \sigma_j\right)$$

be a probability distribution over this cube. Here, β is a 'temperature' term, and Z_β is the normalization term, also known as the 'partition function'.

Remark 20.2. The history of this model was with magnets, where $\sigma_i \sigma_j$ can be considered as some 'magnetic interaction' (sometimes also some 'spin interaction'). There has been lots of work in such models (Curie-Weiss, as well as the more general Gibbs distribution). For instance, even Z_β is intractable and there are interesting algorithms to estimate this normalization term.

We are interested in the *magnetization*, denoted by $m(\sigma) = \frac{1}{n} \sum_{i=1}^n \sigma_i$. Note that if $\beta = 0$, then $m(\sigma) \approx 0$ by the Law of Large Numbers.

It turns out that the magnetization is concentrated near the fixed points of $\tanh(\beta x)$. In particular, we have the following result.

Theorem 20.3 (Curie-Weiss Magnetization Concentration)**.** Let $m(\sigma) = \frac{1}{n} \sum_{i=1}^n \sigma_i$. Then for all $t \geq 0$, we have the inequality

$$\mathbb{P}\left[|m(\sigma) - \tanh(\beta m(\sigma)| \geq \frac{\beta}{n} + t\right] \leq 2\exp(-\frac{nt^2}{2(1 + \beta)}).$$

Remark 20.4. In particular, we find that if $\mathcal{M} = \{m \in [-1,1] : m = \tanh(\beta m)\}$, then $m(\sigma)$ is with high probability close to an element of $\mathcal{M}$.

In particular, we have an interesting phase transition: when $\beta < 1$, $\mathcal{M} = \{0\}$, but when $\beta > 1$, $\mathcal{M} = \{0, \pm m^*\}$ for a some $m^* > 0$. In particular, we find that $m(\sigma) \approx 0$ for $\beta < 1$, but for $\beta > 1$ we have no such guarantees. In fact, it has been shown that $m(\sigma) \approx \pm m^*$ and is not close to 0.

21. Lecture 21: November 18

We will continue discussing concentration for the Curie-Weiss model, followed by nonasymptotic random matrix theory.

21.1. Curie-Weiss Model, Cont'd.

We begin by proving Theorem 20.3, which was stated in the previous lecture.

Proof. First consider an exchangeable pair (σ, σ') attained by taking one-step of the Gibbs sampling algorithm.[32] That is, we sample $I \sim \text{Unif}([n])$, and replace σ_I by $\sigma_I' \sim \sigma_I \mid (\sigma_{-I})$.

Write $F(\sigma, \sigma') = n(m(\sigma) - m(\sigma'))$; note that F is anti-symmetric. The point is now to apply Theorem 20.1. First, note that we have

$$\begin{aligned}
\mathbb{E}[F(\sigma,\sigma') \mid \sigma] &= \mathbb{E}[n(m(\sigma) - m(\sigma')) \mid \sigma] \\
&= \mathbb{E}[\sigma_I - \sigma_I' \mid \sigma] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\sigma_i - \sigma_i' \mid \sigma, I = i] \\
&= m(\sigma) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\sigma_i' \mid I = i, \sigma] \\
&= m(\sigma) - \frac{1}{n}\sum_{i=1}^n \tanh(\beta m_i(\sigma)) \\
&= f(\sigma),
\end{aligned}$$

where we have used that

$$\mathbb{E}[\sigma_i' \mid I = i, \sigma] = \frac{\exp(\beta m_i(\sigma)) - \exp(-\beta m_i(\sigma))}{\exp(\beta m_i(\sigma)) + \exp(-\beta m_i(\sigma))} = \tanh(\beta m_i(\sigma)).$$

By Theorem 20.1, we have

$$\mathbb{P}[|f(\sigma)| \geq t] \leq 2\exp\left(-\frac{t^2}{2Bt + c}\right)$$

where B, C are constants such that the linear function $Bx + C$ bounds the variance proxy v.

We now analyze v. Note that $|F| \leq 2$ because $|m(\sigma) - m(\sigma')| \leq \frac{1}{n}$ as (σ, σ') differ in at most one coordinate. We can thus write

$$\begin{aligned}
v(\sigma) &= \frac{1}{2}\mathbb{E}\left[|f(\sigma) - f(\sigma')|\,|F(\sigma,\sigma')| \mid \sigma\right] \\
&\leq \mathbb{E}\left[|f(\sigma) - f(\sigma')| \mid \sigma\right].
\end{aligned}$$

[32] This is in line with the intuition of choosing exchangeable pairs based on Markov chains.

Now, the term inside the conditional expectation resembles a Lipschitz-type expression. To bound this, note that tanh is 1-Lipschitz.[33] Now, expanding f and using the above Lipschitzness property, we have that

$$\begin{aligned}|f(\sigma)-f(\sigma')| &\le |m(\sigma)-m(\sigma')| + \left|\frac{1}{n}\sum_{i=1}^n \tanh(\beta m_i(\sigma)) - \frac{1}{n}\sum_{i=1}^n \tanh(\beta m_i(\sigma'))\right| \\ &\le \frac{2}{n} + \frac{1}{n}\sum_{i=1}^n |\tanh(\beta m_i(\sigma)) - \tanh(\beta m_i(\sigma'))| \\ &\le \frac{2}{n}(1+\beta).\end{aligned}$$

Therefore, we can bound the variance proxy by $v(\sigma) \le Bt + C$ for $B = 0$, $C = \frac{2}{n}(1+\beta)$. We thus find by plugging in the formula for f that

$$\mathbb{P}\left[\left|m(\sigma) - \frac{1}{n}\sum_{i=1}^n \tanh(\beta m_i(\sigma))\right| \ge t\right] \le 2\exp\left(-\frac{nt^2}{2(1+\beta)}\right).$$

Now, we have that

$$\left|\tanh(\beta m(\sigma)) - \frac{1}{n}\sum_{i=1}^n \tanh(\beta m_i(\sigma))\right| = \frac{1}{n}\left|\sum_{i=1}^n (\tanh(\beta m(\sigma)) - \tanh(\beta m_i(\sigma))\right| \le \frac{\beta}{n},$$

where the final inequality follows again from 1-Lipschitzness of tanh. Combining this with the previous tail bound completes the proof.

□

This completes our discussion of the exchangeable pairs method.

Remark 21.1. The two main steps in this method are (1) find the exchangeable pair, and (2) find and antisymmetric function. The step (1) is often easy, because we can just take one step of a Gibbs sampling algorithm. The second part is usually more difficult. The above example shows that for linear statistics, some sort of difference can be a good proxy. In general, however, one requires some amount of ingenuity to find a suitable function.

21.2. **Nonasymptotic Random Matrix Theory.** We will now move to discussing nonasymptotic random matrix theory (RMT). A standard reference is [Ver11]. Another reference for the immediately upcoming result is [Ver18], Ch. 4.4.

An initial setup is as follows: we have a matrix $A \in \mathbb{R}^{m\times n}$ with iid entries $A_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$. Recall that we have in this setting from Gaussian comparison the inequality

$$\mathbb{E}\|A\| \le \sqrt{m} + \sqrt{n}.$$

Of course, for application, we would like to generalize this to non-gaussian entries:

Question 21.2. What happens if the A_{ij} are not Gaussian?

The natural generalization is for general subgaussian random variables. In particular, we have the following theorem.

[33] In particular, $|\tanh x - \tanh y| = \operatorname{sech}^2(x^*) \le 1$ by the mean value theorem.

Theorem 21.3 (Operator Norm of Subgaussian RM)**.** Let $A \in \mathbb{R}^{m\times n}$ with independent, centered subgaussian entries. Then

$$\mathbb{P}\left[\|A\| \le CK(\sqrt{m}+\sqrt{n}+t)\right] \ge 1-2\exp(-t^2)$$

for $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$ for some universal constant $C > 0$.[34]

Proof. Recall that in the Gaussian setting, we have employed the operator norm expression

$$\|A\| = \sup_{x\in\mathcal{S}^{n-1}, y\in\mathcal{S}^{m-1}} \langle Ax, y\rangle.$$

The main difficulty now is that we can no longer use Gaussian comparison directly. However, recall that we have a general perspective on bounding the suprema of Gaussian processes via ϵ-nets and metrics. In these results, the Gaussian dependence is weak; therefore, one can hope that a similar net argument will help us bound the operator norm in this setting.

In fact, this will work. Let $\mathcal{N}, \mathcal{M}$ be ϵ-nets for $\mathcal{S}^{n-1}, \mathcal{S}^{m-1}$, respectively. The main idea is now to write an inequality of the form

$$\sup_{x\in\mathcal{N}, y\in\mathcal{M}} \langle Ax, y\rangle \le \|A\| \le \frac{1}{1-2\epsilon}\cdot \sup_{x\in\mathcal{N}, y\in\mathcal{M}} \langle Ax, y\rangle. \tag{4}$$

In other words, bounding $\|A\|$ is essentially equivalent to bounding this *discretized* 'subgaussian process' $\langle Ax, y\rangle$. We can then apply a union bound, remarking that the ϵ-nets are of roughly exponential order. It therefore suffices to analyze $\mathbb{P}[\langle Ax, y\rangle] > t]$. But one can check that the inner product is a sum of centered subgaussian RVs with bounded Orlicz norm, which is another centered subgaussian RV with bounded Orlicz norm; one can then use properties of subgaussian RVs to bound the above probability by a negative exponential.

Now, let us move onto the formal proof. Fix $\epsilon = \frac{1}{4}$; by Lemma 13.1, there exists ϵ-nets $\mathcal{N}, \mathcal{M}$ of $\mathcal{S}^{n-1}$, $\mathcal{S}^{m-1}$ with $|\mathcal{N}| \le 9^n$, $|\mathcal{N}| \le 9^m$. Now, we will justify Equation (4).

The LHS of Equation (4) is immediate because $\mathcal{N}\times\mathcal{M} \subseteq S^{n-1}\times S^{m-1}$. For the RHS, fix $x, y \in \mathcal{S}^{n-1}, \mathcal{S}^{m-1}$ and take $x_0 \in \mathcal{N}$, $y_0 \in \mathcal{M}$ such that $\|x-x_0\|_2, \|y-y_0\|_2 \le \epsilon$. It then follows that

$$\langle Ax, y\rangle - \langle Ax_0, y_0\rangle = \langle Ax, y-y_0\rangle + \langle A(x-x_0), y_0\rangle \le 2\epsilon\|A\|.$$

It follows by choosing x, y such that $\langle Ax, y\rangle = \sup\langle Ax, y\rangle$[35]

$$\sup\langle Ax_0, y_0\rangle \ge (1-2\epsilon)\|A\|,$$

proving the RHS. This proves Equation (4).

Now, we apply the aforementioned union bound. We find that

$$\mathbb{P}\left(\max_{x\in\mathcal{N}, y\in\mathcal{M}} \langle Ax, y\rangle \ge t\right) \le \sum_{x\in\mathcal{N}, y\in\mathcal{M}} \mathbb{P}\left(\langle Ax, y\rangle \ge t\right).$$

For each individual term, let us note that

$$\langle Ax, y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}x_i y_j$$

[34]Here, the ψ_2 norm is the Orlicz norm of a subgaussian random variable, given by $\inf\{\sigma : E[e^{\lambda x}] \le e^{\frac{\lambda^2\sigma^2}{2}}\forall\lambda\}$.

[35]Such an optimizer exists by compactness of $\mathcal{S}^{n-1}\times\mathcal{S}^{m-1}$

is a sum of independent subgaussian RVs, so itself is subgaussian[36] with Orlicz norm

$$\|\langle Ax, y\rangle\|_{\psi_2}^2 \leq C_0 \sum_{i=1}^{n}\sum_{j=1}^{m} \|A_{ij}x_i y_j\|_{\psi_2}^2 \leq C_0 K^2 \sum_{i=1}^{n}\sum_{j=1}^{m} x_i^2 y_j^2 = C_0 K^2$$

for some constant C_0. It follows from the subgaussian tail bound that

$$\mathbb{P}\left(\langle Ax, y\rangle \geq t\right) \leq 2\exp\left(-\frac{cu^2}{K^2}\right)$$

for some constant c. Combining with the previous union bound, we find that

$$\mathbb{P}\left(\max_{x\in\mathcal{N}, y\in\mathcal{M}} \langle Ax, y\rangle \geq t\right) \leq 9^{n+m}\cdot 2\exp\left(-\frac{cu^2}{K^2}\right).$$

Recalling our parameter regime, replace t with $C_1K(\sqrt{n}+\sqrt{m}+t)$, which is the quantity we want to analyze, where C_1 is a new, sufficient large constant such that

$$\frac{c(\sqrt{n}+\sqrt{m}+t)^2}{K^2} \geq 4(n+m)+t^2.$$

Then we find that

$$\mathbb{P}\left(\max_{x\in\mathcal{N}, y\in\mathcal{M}} \langle A, y\rangle \geq C_1K(\sqrt{n}+\sqrt{m}+t)\right) \leq 2\exp(-t^2).$$

It follows from Equation (4) that

$$\mathbb{P}\left(\|A\| \geq 2C_1K(\sqrt{n}+\sqrt{m}+t)\right) \leq 2\exp(-t^2).$$

Letting $C = 2C_1$ be the constant in the theorem statement, we find exactly the desired result. This completes the proof. □

Remark 21.4. There are many more results in this line of study; for instance, one can improve constants above, or relax the conditions. We will not go into further details here, but the interested reader can consult [Ver11].

22. Lecture 22: November 20

We will now discuss an application of the above result, namely applications of spectral theory to the problem of *community detection*. For a reference, see [Ver18], Ch. 4.5.

22.1. **Community Detection.** The setup is as follows. Take a graph $G = ([n], E)$. One can imagine that there would be subsets $S \subseteq G$ with 'high connectivity', for instance measured by the number of edges; we can try to imagine these as 'communities'. We are then interested in determining these 'communities'.[37]

In order to understand and analyze algorithms for community detection, we first need an appropriate model. This will be the *stochastic block model*[38]

[36]Here, we use the fact that if X_i are independent and subgaussian with Orlicz norms K_i, then $\sum X_i$ is subgaussian with Orlicz norm at most $\sqrt{\sum X_i^2}$.

[37]One can imagine 'community detection' as some type of clustering problem for graphs.

[38]See Holland-Laskey-Leinhardt in the 80s and then Dyer-Frieze in the 90s.

Definition 22.1 (Stochastic Block Model)**.** The simplest version is a variant of the Erdos-Renyi random graph. Let $[n] = S \cup S^c$ for $|S| = \frac{n}{2}$.[39] Then we consider random graphs G such that

$$\mathbb{P}[\{i,j\} \in E] = \begin{cases} p & (i,j) \in S \text{ or } (i,j) \notin S \\ q & \text{else.} \end{cases}$$

Assume that p, q are roughly fixed, and of order $O(1)$. Intuitively, if $p \gg q$, we can expect more density of edges within the communities, and if $p \ll q$, we can expect more density of edges between the communities.

The main question is the following:

Question 22.2. How do we recover S from G?

One idea is via *spectral clustering*. The main point is to utilize tools in spectral graph theory, which intuitively give us useful information about the connectivity of a graph.[40]

In fact, let us note the following. Let A denote the adjacency matrix of G. We can, in the spirit of this course, write

$$A = \mathbb{E}A + (A - \mathbb{E}A).$$

It turns out the expectation term contains all the information about the groups. In particular, supposing that $S = \{1, \dots, j\}$, we have

$$\mathbb{E}A = \begin{pmatrix} p\mathbf{J} & q\mathbf{J} \\ q\mathbf{J}' & p\mathbf{J} \end{pmatrix}$$ [41]

Here $\mathbf{J}$ is the all ones matrix of appropriate size. It is easy to see that $\mathbb{E}A$ is a rank-two matrix with nonzero eigenvalues given by $\lambda_1 = n\left(\frac{p+q}{2}\right)$, $\lambda_2 = n\left(\frac{p-q}{2}\right)$, and associated eigenvectors $\mathbf{u}_1 = \mathbf{1}$, $\mathbf{u}_2 = [\overbrace{1, \dots, 1}^{\frac{n}{2}}, \overbrace{-1, \dots, -1}^{\frac{n}{2}}]^\intercal$. Note that if $S \neq [\frac{n}{2}]$ then the entries of $\mathbf{u}_2$ will be permuted accordingly.

A natural algorithm to consider is therefore the following. If the noise of A is not large, then $\mathbf{u}_2$ essentially gives us an indicator of whether a vertex is in S or S^c. Of course, in the problem we are not given access to $\mathbf{u}_2$. But we can consider the eigenvector corresponding to the second-largest eigenvalue, and hope that it is 'close enough' to $\mathbf{u}_2$.

Formally, we therefore have the following algorithm:

(1) Calculate the adjacency matrix A and its second-largest eigenvector $\mathbf{v}_2$.
(2) Cluster based on the signs of the eigenvector entries.

Now, we can ask, *does this algorithm work?* It turns out the answer is yes; in particular, we can state the following result.

Theorem 22.3 (Lower-Bound Guarantee on the Accuracy of Classification)**.** With high probability $1 - \exp(-n)$, we will misclassify only $\mathcal{O}(1)$ entries.

To prove this, we will appeal to the Davis-Kahan theorem.

[39] Of course n is usually large so the possible rounding error is irrelevant.

[40] Another analogue is, for instance, spectral methods in mixing times of Markov chains, where the spectral gap dictates fast-mixing time.

[41] Note the block structure of the matrix. This gives justification to the name 'block model'.

Theorem 22.4 (Davis-Kahan)**.** Let S, T be symmetric matrices of matching dimensions. Fix i and assume that the i-th largest eigenvalue of S, $\lambda_i(S)$, is well-separated from the rest of the spectrum, in the sense that

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| > \delta.$$

Then the angle

$$\theta = \cos^{-1}\left(\frac{\langle v_i(S), v_i(T)\rangle}{\|v_i(S)\|\|v_i(T)\|}\right)$$

between the i-th eigenvectors of S and T satisfies

$$|\sin\theta| \leq \frac{2\delta}{\|S-T\|}.$$

This is a type of 'perturbation result', in that perturbing S slightly (in a rigorous sense defined by the operator norm) to T does not significantly change θ.

In order to use this result, we need the following lemma.

Lemma 22.5. We have the inequality $|\|A\| - \mathbb{E}\|A\|| \leq C\sqrt{n}$ with probability $1 - \exp(-n)$.

Proof. This follows because A has iid Bernoulli entries, which are subgaussian; we can then apply Theorem 21.3 directly. □

This essentially shows that the 'noise' term is small with high probability. We are now ready to prove Theorem 22.3.

Proof of Theorem 22.3. Intuitively, the point is that Lemma 22.5 bounds the noise term; then applying Davis-Kahan will show that the second-largest eigenvector of A is close to the 'true' second-largest eigenvector $\mathbb{E}A$, which gives us the true classification.

Formally, let $S = \mathbb{E}A$ and $T = A$, we know that $|\|S\| - \|T\|| \leq C\sqrt{n}$ with high probability. Applying the Davis-Kahan theorem with the parameter setting

$$\delta = \min(\lambda_2(S), \lambda_1(S) - \lambda_2(S)) = n\min\left(\frac{p-q}{2}, q\right) = \mu n,$$

we find that

$$\sin(\theta_2) \leq \frac{2\|S-T\|}{\delta} \leq 2\frac{C\sqrt{n}}{\delta} = \frac{2C}{\mu\sqrt{n}}.$$

Recall that $\theta_2 = \angle(\mathbf{u}_2, \mathbf{v}_2)$, where $\mathbf{u}_2$, $\mathbf{v}_2$ are the eigenvectors corresponding to $\lambda_2(S)$, $\lambda_2(T)$. Recalling that $\mathbf{u}_2 \in \{\pm 1\}^n$ has norm $\sqrt{n}$, take $\mathbf{v}_2$ to have norm $\sqrt{n}$ as well. It follows that there exists $\theta \in \{\pm 1\}$ such that

$$\|u_2(A) - \theta u_2(\mathbb{E}A)\|_2 \leq \frac{C'}{\mu},$$

we have two eigenvectors of length $\sqrt{n}$ with the sine of their angle on the order of $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

Now, our classification algorithm is incorrect for a given coordinate i only if $|u_2(A)_i - \theta u_2(\mathbb{E}A)|^2 \geq 1$. Combining everything, we conclude that our algorithm misclassifies $\mathcal{O}(1)$ entries with probability $1 - \exp(-n)$. This completes the proof. □

Remark 22.6. This is a very nice result, and essentially completely solves the problem in this case. More modern work, however, has been inspired by the observation that in practice, our adjacency matrices are much more sparse. Therefore, a more realistic assumption does not fix p, q, but rather takes $p = p_n, q = q_n$ with $p_n, q_n \overset{n\to\infty}{\longrightarrow} 0$.

The algorithms above now break down in the worst case. For instance, if $p_n, q_n = \frac{c}{n}$, then the noise term $A - \mathbb{E}A$ has centered entries with variance on the order of $\mathcal{O}\left(\frac{1}{n}\right)$. Now, if the entries were Gaussian, then the operator norm $\|A - \mathbb{E}A\| = \mathcal{O}(1)$. Unfortunately, in this case, it turns out that

$$\|A - \mathbb{E}A\| = \mathcal{O}\left(\sqrt{\frac{\log n}{\log\log n}}\right).$$

This type of phenomenon happens essentially due to the presence of a few high-degree vertices, which turn out to dominate. Unfortunately, failure to bound the noise term appropriately now forces the method to break down.

Intriguingly, it turns out that spectral clustering methods still work! However, the key insight is to choose a *different* matrix altogether, namely the so-called non-backtracking (Hashimoto) matrix, $\mathcal{M}$. These revised spectral methods were proposed around a decade ago (c. 2011), and there have since been formal guarantees proven for them.

23. Conclusion

This concludes the main lecture content in the course! As a brief summary, we have covered the following main topics:

- Concentration
- Gaussian Suprema and Comparison
- Universality
- Stein's Method, Exchangeable Pairs Method
- Nonasymptotic Random Matrix Theory

There are many more interesting ideas and extensions of each of the individual topics we have covered. For instance, in concentration, the recent and more abstract idea of 'superconcentration' (by Chatterjee), or even more classical results such as Poincare and log-Sobolev inequalities are highly intriguing. Recent applications of Stein's method has been to Poisson and even general non-Gaussian distributions. Ultimately, we have covered a survey of both classical and modern ideas from this beautiful field; we encourage the interested reader to explore the vast expanses beyond.

References

[BLM13] Boucheron, Lugosi, and Massart, 'Concentration Inequalities: A Nonasymptotic Theory of Independence'.

[Cha05] Chatterjee, 'Concentration Inequalities with Exchangeable Pairs.' Link.

[DSL24] Dudeja, Sen, and Lu, 'Universality of approximate message passing with semirandom matrices'. Link

[HS23] Qiyang Han and Yandi Shen, 'Universality of Regularized Regression Estimators in High Dimensions'. Link

[Lal] Steve Lalley, 'Concentration Inequalities'. Link.

[MM18] Leo Miolane and Andrea Montanari, 'The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning'. Link

[TOH15] Thrampoulidis, Oymak, Hassibi, 'Regularized Linear Regression: A Precise Analysis of the Estimation Error.' Link
[H18] Ramon Van Handel, 'Probability in High Dimension'. Link.
[Ver18] Roman Vershynin, 'High-Dimensional Probability'. Link.
[Ver11] Roman Vershynin, 'Non-Asymptotic Random Matrix Theory'. Link.
[Ros11] Nathan Ross, 'Fundamentals of Stein's Method'. Link.