

STAT 9910-303: Random Matrix Theory

Instructor: Jiaoyang Huang

October 12, 2026

Contents

1	Introduction	5
2	Eigenvalues of Random matrices	7
2.1	Lecture 1: Semicircle Distribution	7
2.2	Lecture 2: Extreme Eigenvalues	12
2.3	Lecture 3: Marchenko-Pastur Distribution	18
2.4	Lecture 4: Marchenko-Pastur Distribution Continued	22
3	Non-asymptotic Analysis	25
3.1	Lecture 5: Largest and Smallest Singular Values	26
3.2	Lecture 6: Randomized Algorithms	31
3.3	Lecture 7: Matrix Concentration Inequality	37
4	Spiked Matrix Models	41
4.1	Lecture 7: Spiked Matrix Model I	41
4.2	Lecture 8: Spiked Matrix Model II	44
4.3	Lecture 9: Spiked Tensor Model	49
5	Statistical Learning	53
5.1	Introduction to Deep Learning	54
5.1.1	Overview	54
5.1.2	Transfer Learning: Theoretical Development	55
5.1.3	Optimization of Over-Parametrized Neural Networks	58
5.2	Lecture 11: Random Feature Model	59
5.3	Lecture 12.1: Approximate message passing	65
5.4	Lecture 12.2: Ridge Regression	69
5.5	Lecture 13: Mean Field Limit of Two-Layer Neural Networks	73

Chapter 1

Introduction

In this class, we will cover the following four aspects of random matrix theory:

Universality of eigenvalue distributions: Random matrices have been introduced by physicist Wigner in the 1950s to model the Hamiltonians of heavy nuclei. The success of random matrices in modeling physical systems lies in the universality phenomenon of their eigenvalue and eigenvector statistics. Under mild conditions, the empirical eigenvalue distributions of Wigner matrices and sample covariance matrices converge to the *Semi-circle distribution* and *Marchenko-Pastur distribution*. We will discuss different approaches to prove this universality phenomenon.

Non-asymptotic analysis of random matrices: Sample covariance matrices play a central role in statistics. For high-dimensional data, when the dimension is comparable to the sample size, it falls into the regime of random matrix theory. We will discuss tools for the analysis of the extreme singular values of random matrices with possibly dependent rows or columns. Several applications will be covered, particularly for the problem of estimating covariance matrices in statistics and for probabilistic constructions of measurement matrices in compressed sensing.

Signal detection in spiked random matrix models: Many statistical models are in the form of “signal” + “noise”, and the signal-to-noise ratio measures the strength of the signal relative to the noise. For the spiked random matrix model, the signal is a deterministic low rank matrix and noise is some random matrix, and the goal is to detect/recover the signal from the noisy observation. For the standard dense priors discussed below, there is a critical signal-to-noise ratio for consistent detection (both testing errors tending to zero). Below this threshold, consistent detection is information-theoretically impossible; above this threshold, simple spectral algorithms detect the signal. These thresholds need not coincide for other priors. We will discuss this phenomenon for spiked Wigner matrices and sample covariance matrices.

Deep neural networks: Deep neural networks are built out of matrices and nonlinear activation functions. The weights of deep neural networks are initialized to random numbers, and then trained using stochastic gradient descent. The randomly initialized deep neural networks are compositions of random matrices and nonlinear activation functions. We will discuss the properties of randomly initialized deep neural networks, and the neural tangent kernel associated with deep neural networks. Under suitable initialization and scaling, in the infinite-width limit the gradient-flow training dynamics on fixed time intervals are described by the neural tangent kernel; when this flow converges, its predictions are related to kernel regression.

Besides the above materials, we will also have several guest lectures on the applications of

random matrix theory to high dimensional statistics, computer science and bio-statistics.

Acknowledgements

These lecture notes are based on STAT 9910-303, a topics course on random matrix theory and its applications taught at the University of Pennsylvania. I thank the following students for scribing the lectures and contributing to these notes: Yangxinyu Xie, Anirban Chatterjee, Bingyu Zhang, Abhinandan Dalal, Zhaojun Zhang, Lu Li, Joseph Deutsch, Wei Fan, Zirui Fan, Zi Wang, Evangelos Chatzipantazis, Qipeng Zhan, Jinho Bok, Ziang Niu, Changhao Ge, Ryan Jeong, Brian Lee, Mahdi Sabbaghi, Buxin Su, and Xinmeng Huang.

Chapter 2

Eigenvalues of Random matrices

2.1 Lecture 1: Semicircle Distribution

In this lecture, we prove the convergence of the empirical distribution of the eigenvalues of the Wigner matrix via the moment method.

Theorem 2.1.1. *Let*

$$H = \frac{1}{\sqrt{n}} \begin{bmatrix} h_{11} & \dots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \dots & h_{nn} \end{bmatrix}$$

be a real Wigner matrix, i.e. for $i, j = 1, \dots, n$, $h_{ij} = h_{ji}$ and $\{h_{ij}\}_{i \leq j}$ are i.i.d. with $\mathbb{E}[h_{ij}] = 0$, $\mathbb{E}[|h_{ij}|^2] = 1$. Moreover, for each $k \geq 1$, assume $\mathbb{E}[|h_{ij}|^k] \leq B_k$ uniformly in n . Denote the eigenvalues of H as $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ and the empirical eigenvalue distribution $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i}$. Then

$$\mu_n \xrightarrow{D} \rho_{sc}, \quad \text{in probability}$$

That is, μ_n converges to the semicircle law weakly in probability, where

$$\rho_{sc}(x) = \frac{1}{2\pi} \sqrt{4 - x^2} \mathbb{1}_{\{|x| \leq 2\}}.$$

Since H is symmetric, we can write the Wigner matrix as

$$H = O \Lambda O^T$$

where O is orthogonal and $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_n\}$. Observe we can write the trace of H as

$$\text{Tr}(H) = \text{Tr}(O \Lambda O^T) = \sum_{i=1}^n \lambda_i$$

$$\text{Tr}(H^2) = \text{Tr}(O \Lambda^2 O^T) = \sum_{i=1}^n \lambda_i^2$$

$$\text{Tr}(H^k) = \text{Tr}(O \Lambda^k O^T) = \sum_{i=1}^n \lambda_i^k = n \int x^k d\mu_n$$

Meanwhile, we have

$$\mathrm{Tr}(H^k) = \sum_{i_1, \dots, i_k} H_{i_1 i_2} \cdot H_{i_2 i_3} \cdot \dots \cdot H_{i_k i_1} = \left(\frac{1}{\sqrt{n}}\right)^k \sum_{i_1, \dots, i_k} h_{i_1 i_2} \cdot h_{i_2 i_3} \cdot \dots \cdot h_{i_k i_1}$$

To prove Theorem 2.1.1, we aim to show that for $k \geq 1$,

$$\int x^k d\mu_n = \frac{1}{n^{k/2+1}} \sum_{i_1, \dots, i_k} h_{i_1 i_2} \cdot h_{i_2 i_3} \cdot \dots \cdot h_{i_k i_1} \approx \int x^k \rho_{sc}(x) dx$$

Notice that

$$\int x^k \rho_{sc}(x) dx = \begin{cases} 0 & \text{for odd } k \\ C_{k/2} & \text{for even } k \end{cases}$$

where

$$C_k = \frac{\binom{2k}{k}}{k+1}$$

is the **Catalan number**. The Catalan number C_k is a sequence of natural numbers that occur in various counting problems. For example, the Catalan number counts

- the number of simple random walks from $(0, 0)$ to $(2k, 0)$, whose trajectories stay at or above 0;
- the number of non-crossing partitions of $\{1, 2, \dots, 2k\}$ in which every block is of size 2.
- the number of planar rooted trees with k edges and $k+1$ vertices. This particular example will come up in our proof later.

We need two auxiliary lemmas: for fixed $k \geq 1$,

Lemma 2.1.2. $\mathbb{E}[\int x^k d\mu_n] \rightarrow \int x^k \rho_{sc}(x) dx$ as $n \rightarrow \infty$.

Lemma 2.1.3. For $\varepsilon > 0$, $\mathbb{P}(|\int x^k d\mu_n - \mathbb{E}[\int x^k d\mu_n]| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$.

Combining Lemma 2.1.2 and Lemma 2.1.3, we have $\mathbb{P}(|\int x^k d\mu_n - \int x^k \rho_{sc}(x) dx| > \varepsilon) \rightarrow 0$ and thus $\mathbb{P}(|\int f(x) d\mu_n - \int f(x) \rho_{sc}(x) dx| > \varepsilon) \rightarrow 0$ for any polynomial f . A tail argument is needed before using polynomial approximation. From any subsequence we can choose a further subsequence along which all moments converge almost surely. Along each such realization, bounded second moments give tightness, and bounded higher even moments give uniform integrability of every fixed power. Consequently any weak subsequential limit has the semicircle moments. These moments determine a unique measure: since $C_k \leq 4^k$, Markov's inequality applied to all even moments forces its support to lie in $[-2, 2]$, where polynomial approximation determines the measure. Thus every such limit is the semicircle law. The subsequence criterion for convergence in probability proves Theorem 2.1.1.

Proof of Lemma 2.1.2. It suffices to show that

$$\frac{1}{n^{k/2+1}} \sum_{i_1, \dots, i_k} \mathbb{E}[h_{i_1 i_2} \cdot h_{i_2 i_3} \cdot \dots \cdot h_{i_k i_1}] \rightarrow \int x^k \rho_{sc}(x) dx$$

For each term $\mathbb{E}[h_{i_1 i_2} \cdot h_{i_2 i_3} \cdots h_{i_k i_1}]$, consider the closed walk $(i_1, \dots, i_k, i_1)$. Its connected graph $\mathcal{G} = (V, E)$ has distinct vertices V and distinct undirected edges E (loops are allowed). Let m_e be the number of traversals of e . In the sums below, $\mathcal{G}$ denotes the entire walk up to relabeling, not just its underlying graph: label vertices $1, 2, \dots$ in order of first appearance. Then $\Pi(\mathcal{G}) = \prod_{e \in E} \mathbb{E}[h_e^{m_e}]$ depends only on this canonical walk, and the number of labelings is $(n)_{|V|} = n(n-1) \cdots (n-|V|+1)$. Thus, for fixed k ,

$$\frac{1}{n^{k/2+1}} \sum_{i_1, \dots, i_k} \mathbb{E}[h_{i_1 i_2} \cdots h_{i_k i_1}] = \frac{1}{n^{k/2+1}} \sum_{\mathcal{G}} (n)_{|V|} \Pi(\mathcal{G}).$$

Let's make a few observations:

- $\Pi(\mathcal{G}) = 0$ unless the multiplicity of every edge is at least 2.
- Notice

$$\Pi(\mathcal{G}) = \mathbb{E}[h_{i_1 i_2} \cdot h_{i_2 i_3} \cdots h_{i_k i_1}] = \mathbb{E}[h_{e_1}^{m_1} \cdot h_{e_2}^{m_2} \cdots h_{e_r}^{m_r}]$$

By Hölder's inequality,

$$|\mathbb{E}[h_{e_1}^{m_1} \cdot h_{e_2}^{m_2} \cdots h_{e_r}^{m_r}]| \leq \mathbb{E}[|h_{e_1}|^k]^{\frac{m_1}{k}} \cdot \mathbb{E}[|h_{e_2}|^k]^{\frac{m_2}{k}} \cdots \mathbb{E}[|h_{e_r}|^k]^{\frac{m_r}{k}} \leq B_k$$

- If the multiplicity of every edge is at least 2, then we have

$$|E| \leq \frac{k}{2}, \quad |V| \leq \frac{k}{2} + 1$$

with $|V| = \frac{k}{2} + 1$ if and only if $\mathcal{G}$ is a tree, each of whose edges has multiplicity 2.

Consider the following two cases (only walks with nonzero $\Pi(\mathcal{G})$ need to be included):

1. $|V| \leq k/2 + 1/2$. The absolute value of their total contribution is bounded by

$$\frac{1}{n^{k/2+1}} \sum_{\mathcal{G}: |V| \leq k/2+1/2} (n)_{|V|} |\Pi(\mathcal{G})| \leq \frac{B_k k^k}{\sqrt{n}} \longrightarrow 0.$$

Here there are at most k^k canonical words of length k .

2. $|V| = k/2 + 1$. Then k is even, the graph is a tree, each edge is traversed twice, and $\Pi(\mathcal{G}) = 1$. Starting at the root, the walk orders the children of each vertex by their first visits, and is exactly the contour walk of the resulting plane rooted tree. Conversely, each plane rooted tree gives one canonical walk. Their total contribution is therefore

$$\frac{(n)_{k/2+1}}{n^{k/2+1}} C_{k/2} = (1 + \mathcal{O}_k(n^{-1})) C_{k/2}.$$

For odd k only the first case occurs. Adding the contributions proves the lemma. $\square$

Proof of Lemma 2.1.3. It suffices to show that

$$\mathbb{E}\left[\left|\int x^k d\mu_n - \mathbb{E}\left[\int x^k d\mu_n\right]\right|^2\right] \rightarrow 0$$

Now,

$$\begin{aligned}
& \mathbb{E} \left[\left| \int x^k d\mu_n - \mathbb{E} \left[\int x^k d\mu_n \right] \right|^2 \right] \\
&= \mathbb{E} \left[\left(\int x^k d\mu_n \right)^2 \right] - \mathbb{E} \left[\int x^k d\mu_n \right]^2 \\
&= \frac{1}{n^{k+2}} \left[\sum_{i_1, \dots, i_k, i'_1, \dots, i'_k} \mathbb{E} [h_{i_1 i_2} \cdots h_{i_k i_1} \cdot h_{i'_1 i'_2} \cdots h_{i'_k i'_1}] \right. \\
&\quad \left. - \sum_{i_1, \dots, i_k, i'_1, \dots, i'_k} \mathbb{E} [h_{i_1 i_2} \cdots h_{i_k i_1}] \cdot \mathbb{E} [h_{i'_1 i'_2} \cdots h_{i'_k i'_1}] \right]
\end{aligned} \tag{2.1.1}$$

Let $I = (i_1, \dots, i_k)$ and $I' = (i'_1, \dots, i'_k)$ denote the two closed walks. Combine the two sums in (2.1.1) into a sum of covariances. If the walks share no undirected edge, the covariance is exactly zero by independence, even if they share a vertex. Otherwise their union is connected. A nonzero covariance requires every edge in the union to appear at least twice in total, so the union has at most k edges.

In fact it has at most k vertices. If the union contains a cycle (including a loop), this follows from $|V| \leq |E| \leq k$. If it is a tree, every closed walk traverses each of its used edges an even number of times. A shared edge is therefore traversed at least four times in total, so $|E| \leq k - 1$ and again $|V| \leq k$. There are finitely many canonical pairs for fixed k , each has at most n^k labelings, and each covariance is bounded in absolute value by a constant depending only on B_{2k} . Consequently

$$\text{Var} \left(\int x^k d\mu_n \right) = \mathcal{O}_k(n^{-2}).$$

Chebyshev's inequality proves the lemma. In addition, summability and the Borel–Cantelli lemma imply almost-sure convergence of every fixed moment. The preceding tightness argument then also gives almost-sure weak convergence of μ_n . $\square$

Remark 2.1.4. By generalizing the proof we have above, we can relax two conditions of Theorem 2.1.1:

1. $\mathbb{E}[|h_{ij}|^2] = 1$ can be replaced by

$$c \leq \mathbb{E}[|h_{ij}|^2] \leq C, \frac{1}{n} \sum_{j=1}^n \mathbb{E}|h_{ij}|^2 = 1$$

for constants $0 < c < C$, allowing independent, not necessarily identically distributed, upper-triangular entries with uniform moment bounds. It is the normalized variance profile $(n^{-1} \mathbb{E}|h_{ij}|^2)_{ij}$ that is stochastic, not $\mathbb{E}H$.

2. For a fixed entry distribution, the higher-moment assumptions can be dropped, retaining mean zero and variance one. Start with the centered truncation

$$\tilde{h}_{ij} = h_{ij} \mathbb{1}_{\{|h_{ij}| \leq M\}} - \mathbb{E}[h_{ij} \mathbb{1}_{\{|h_{ij}| \leq M\}}],$$

and divide by its standard deviation, which tends to one as $M \rightarrow \infty$. For fixed M the bounded-entry proof applies. The Hoffman–Wielandt inequality bounds the mean squared displacement of eigenvalues by $n^{-1}\|H - \tilde{H}\|_F^2$, whose expectation tends to zero as $M \rightarrow \infty$, uniformly in n . This transfers weak convergence in probability to the original matrix. For entry distributions that vary with n , an additional uniform-integrability or Lindeberg condition is needed.

2.2 Lecture 2: Extreme Eigenvalues

In this lecture, we prove the convergence of extreme values.

Theorem 2.2.1. *Let*

$$H = \frac{1}{\sqrt{n}} \begin{bmatrix} h_{11} & \dots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \dots & h_{nn} \end{bmatrix}$$

be a real Wigner matrix, i.e. for $i, j = 1, \dots, n$, $h_{ij} = h_{ji}$ and $\{h_{ij}\}_{i \leq j}$ are i.i.d. with $\mathbb{E}[h_{ij}] = 0$, $\mathbb{E}[|h_{ij}|^2] = 1$, and entries are bounded, i.e., $|h_{ij}| \leq B$. Denote the eigenvalues of H as $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, then $\lambda_1 \rightarrow 2$ and $\lambda_n \rightarrow -2$ in probability and almost surely (as n goes to $+\infty$).

Here we are going to prove two parts:

1. $\mathbb{P}(\lambda_1 \leq 2 - \epsilon) \rightarrow 0$.
2. $\mathbb{P}(\lambda_1 \geq 2 + \epsilon) \rightarrow 0$.

Proof of 1. If $\lambda_1 \leq 2 - \epsilon$, take f as:

$$f(x) = \begin{cases} 0, & x \leq 2 - \epsilon, \\ \frac{2}{\epsilon}(x - 2 + \epsilon), & 2 - \epsilon < x < 2 - \epsilon/2, \\ 1, & x \geq 2 - \epsilon/2, \end{cases}$$

then we have $\int f(x) \rho_{sc}(x) dx = C_\epsilon > 0$.

We also have $\int f d\mu_n = 0$ if $\lambda_1 \leq 2 - \epsilon$, then

$$\mathbb{P}(\lambda_1 \leq 2 - \epsilon) \leq \mathbb{P}\left(\int f d\mu_n = 0\right).$$

Since

$$\mathbb{P}\left(\left|\int f d\mu_n - \int f \rho_{sc}(x) dx\right| > \delta\right) \rightarrow 0 \quad \text{and} \quad \int f \rho_{sc}(x) dx \geq 2\delta,$$

we have

$$\mathbb{P}\left(\left|\int f d\mu_n\right| < \delta\right) \rightarrow 0,$$

which implies that $\mathbb{P}(\lambda_1 \leq 2 - \epsilon) \rightarrow 0$. □

Proof of 2. We can write the trace of H^{2k} as

$$\text{Tr}(H^{2k}) = \lambda_1^{2k} + \lambda_2^{2k} + \dots + \lambda_n^{2k} \geq \lambda_1^{2k},$$

then

$$\begin{aligned} \max\{\lambda_1^{2k}, \lambda_n^{2k}\} &\leq \text{Tr}(H^{2k}) \leq n \max\{\lambda_1^{2k}, \lambda_n^{2k}\}, \\ \max\{|\lambda_1|, |\lambda_n|\} &\leq (\text{Tr}(H^{2k}))^{1/2k} \leq n^{1/2k} \max\{|\lambda_1|, |\lambda_n|\}. \end{aligned}$$

(IDEA: Take $k \gg \ln n$, i.e., $(\ln n)^2$, we can have $n^{1/2k} = e^{\ln n / 2k} \approx e^0 = 1$.)

We need the following Lemma to finish the proof:

Lemma 2.2.2. Take $k = \lceil (\log n)^2 \rceil$. For any fixed $\epsilon > 0$ and all sufficiently large n , we have

$$\mathbb{E}[\text{Tr}(H^{2k})] \leq \left(2 + \frac{\epsilon}{2}\right)^{2k}.$$

Applying Lemma 2.2.2 yields

$$\mathbb{P}(\lambda_1 \geq 2 + \epsilon) \leq \mathbb{P}(\lambda_1^{2k} \geq (2 + \epsilon)^{2k}) \leq \frac{\mathbb{E}[\lambda_1^{2k}]}{(2 + \epsilon)^{2k}} \leq \frac{(2 + \epsilon/2)^{2k}}{(2 + \epsilon)^{2k}} \rightarrow 0.$$

□

For this choice of k , the same estimate for $\|H\|$ is summable in n , so the Borel–Cantelli lemma gives $\limsup \|H\| \leq 2$ almost surely. The almost-sure semicircle law established in Lecture 1 gives $\liminf \lambda_1 \geq 2$. Applying the same argument to $-H$ proves $\lambda_n \rightarrow -2$ almost surely as well.

Next we are going to prove Lemma 2.2.2.

Proof of Lemma 2.2.2. Firstly we know

$$\mathbb{E}[\text{Tr}(H^{2k})] = \frac{1}{n^k} \sum_{i_1, \dots, i_{2k}} \mathbb{E}[h_{i_1 i_2} \cdot h_{i_2 i_3} \cdots h_{i_{2k} i_1}],$$

where $\mathcal{G}$ denotes a canonical closed walk as in Lecture 1, and only nonzero contributions are retained; writing its graph as (V, E) , $1 \leq |V| \leq k + 1$ and $\mathbb{E}[h_{i_1 i_2} \cdots h_{i_{2k} i_1}] = \Pi(\mathcal{G})$. The number of terms $(i_1, i_2, \dots, i_{2k}) \sim \mathcal{G} \leq n^{|V|}$, then

$$\mathbb{E}[\text{Tr}(H^{2k})] \leq \frac{1}{n^k} \sum_{\mathcal{G}} n^{|V|} |\Pi(\mathcal{G})| = \sum_{t=1}^{k+1} \frac{1}{n^{k-t}} \sum_{\mathcal{G}, |V|=t} |\Pi(\mathcal{G})|.$$

Denote $E = E_2 \cup E_{>2}$, where E_2 are edges with multiplicity 2 and where $E_{>2}$ are edges with multiplicity larger than 2. Let m_e be the multiplicity of e , then it is clear that

$$2k = \sum_{e \in E} m_e = \sum_{e \in E_2} 2 + \sum_{e \in E_{>2}} m_e.$$

Also let $|E_2| = l$ and $|E| = |E_2| + |E_{>2}| \geq t - 1$, we claim:

Claim 2.2.3. $|E_{>2}| \leq 2(k - t + 1)$ and $k - l \leq 3(k - t + 1)$.

Then we have

$$|\Pi(\mathcal{G})| = \left| \prod_{e \in E} \mathbb{E}[h_e^{m_e}] \right| \leq \prod_{e \in E_2} \mathbb{E}[h_e^2] \prod_{e \in E_{>2}} \mathbb{E}[|h_e|^{m_e}] \leq B^{\sum_{e \in E_{>2}} m_e} = B^{2k-2l} \leq B^{6(k-t+1)}.$$

We have another claim:

Claim 2.2.4. The number of canonical closed walks with t vertices, all of whose edges have multiplicity at least two, is bounded by

$$\binom{2k}{2t-2} t^{4(k-t+1)} C_{t-1},$$

where $C_{t-1} = \frac{1}{t} \binom{2t-2}{t-1} \leq 2^{2t} = 4^t$.

The second claim is the standard Füredi–Komlós walk-counting bound (*The eigenvalues of random symmetric matrices*, Combinatorica 1 (1981), 233–241); unlike the fixed- k count in Lecture 1, it controls the dependence on k . It is a combinatorial input to this high-moment argument. The walk (including its traversal order), rather than only its underlying graph, must be counted. With these two claims, we plug back to the first equation:

$$\begin{aligned}
\mathbb{E}[\text{Tr}(H^{2k})] &\leq \sum_{t=1}^{k+1} \frac{1}{n^{k-t}} \sum_{\mathcal{G}, |V|=t} |\Pi(\mathcal{G})| \\
&\leq \sum_{t=1}^{k+1} \frac{1}{n^{k-t}} B^{6(k-t+1)} \sum_{\mathcal{G}, |V|=t} 1 \\
&\leq \sum_{t=1}^{k+1} \frac{1}{n^{k-t}} B^{6(k-t+1)} (2k)^{2k-2t+2} (t^2)^{2(k-t+1)} 4^t \\
&\leq n \sum_{t=1}^{k+1} \frac{B^{6(k-t+1)}}{n^{k-t+1}} (2kt^2)^{2(k-t+1)} 4^t \\
&= n \sum_{t=1}^{k+1} \left(\frac{B^6 (2kt^2)^2}{n} \right)^{k+1-t} 4^t \quad (t \leq k+1 \ll n^{1/6}) \\
&\leq n(k+1)4^{k+1} \\
&\leq (2 + \epsilon/2)^{2k}.
\end{aligned}$$

Here $B \geq 1$ is fixed, $B^6(2k(k+1)^2)^2/n = o(1)$, and $\log(n(k+1)4) = o(k)$, which justify the last two inequalities for our choice of k . $\square$

Proof of Claim 2.2.3. We have

$$\begin{aligned}
2k &= \sum_{e \in E_2} 2 + \sum_{e \in E_{>2}} m_e \\
&\geq 2|E_2| + 3|E_{>2}| \\
&= 2(|E_2| + |E_{>2}|) + |E_{>2}| \\
&\geq 2(t-1) + |E_{>2}|,
\end{aligned}$$

then it follows that

$$|E_{>2}| \leq 2k - 2t + 2.$$

Also from the second line, we get

$$2k \geq 2l + 3(t-1-l),$$

thus

$$k - l \leq 3(k - t + 1).$$

$\square$

Remark 2.2.5. For a fixed entry distribution, the finite-fourth-moment hypothesis $\mathbb{E}[|h_{ij}|^4] < \infty$ suffices for the almost-sure conclusion (the Bai–Yin theorem). An infinite fourth moment does not

by itself imply a power-law tail or polynomially diverging eigenvalues. As an example, if

$$\mathbb{P}(|h_{ij}| > t) \sim ct^{-\alpha}, \quad 2 < \alpha < 4,$$

then the maximum among the order- n^2 independent entries is of scale $n^{2/\alpha}$. Since $\|H\| \geq \max_{ij} |h_{ij}|/\sqrt{n}$, these entries force the norm to diverge on scale $n^{2/\alpha-1/2}$.

In the above discussion, in particular, in Theorem 2.2.1 we have shown convergence of extreme eigenvalues of a Wigner matrix $\mathbf{H}$, which can be thought of as a version of the law of large numbers for the extreme eigenvalues of $\mathbf{H}$. In the following, we discuss the concentration of extreme eigenvalues around their means. These are concentration bounds, not a central limit theorem.

Before going into the optimal concentration of eigenvalues of $\mathbf{H}$, let us first look at a more straightforward concentration result that uses tools from Gaussian concentration.

Theorem 2.2.6 (Gaussian concentration inequality). *Let $F : \mathbb{R}^n \rightarrow \mathbb{R}$ be a Lipschitz function, that is,*

$$|F(x) - F(y)| \leq L\|x - y\|_2 \text{ for all } x, y \in \mathbb{R}^n$$

and for some constant $L > 0$. Then for $\mathbf{X} = (X_1, X_2, \dots, X_n) \sim N(\mathbf{0}, \mathbf{I}_{n \times n})$,

$$\mathbb{P}(|F(\mathbf{X}) - \mathbb{E}F(\mathbf{X})| > t) \leq C \exp\left(-\frac{c}{L^2}t^2\right)$$

where the constants, C and c are universal constants.

Now consider a $n \times n$ symmetric matrix $\mathbf{A} = ((a_{ij}))_{i,j=1,\dots,n}$ with eigenvalues $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots \geq \lambda_n(\mathbf{A})$. For $1 \leq i \leq n$, by definition, $\lambda_i(\mathbf{A})$ is a function of $\{a_{ij} : 1 \leq i \leq j \leq n\}$. For two $n \times n$ symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, by Weyl's inequality,

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|_{2 \rightarrow 2} \leq \|\mathbf{A} - \mathbf{B}\|_F \quad (2.2.1)$$

where $\|\cdot\|_{2 \rightarrow 2}$ is the spectral norm and $\|\cdot\|_F$ is the Frobenius norm. By (2.2.1) it can be seen that $\lambda_i(\mathbf{A})$ is a Lipschitz function of $\{a_{ij} : 1 \leq i \leq j \leq n\}$. Thus, for $\mathbf{H} = n^{-1/2}(h_{ij})$ with independent $h_{ij} \sim N(0, 1)$ for $i \leq j$ and $h_{ji} = h_{ij}$, the Lipschitz constant in the independent entries is at most $\sqrt{2/n}$. Using Theorem 2.2.6 we get,

$$\mathbb{P}(|\lambda_i(\mathbf{H}) - \mathbb{E}\lambda_i(\mathbf{H})| > t) \leq C \exp(-cnt^2)$$

Having established a concentration result for eigenvalues of the Wigner matrix $\mathbf{H}$, we present an informal version of the optimal concentration result in the following.

Theorem 2.2.7 (Eigenvalue rigidity I). *Consider a Wigner matrix $\mathbf{H}$ satisfying the assumptions of Theorem 2.2.1. For every fixed $\varepsilon, D > 0$, for all sufficiently large n , with probability at least $1 - n^{-D}$, simultaneously for all intervals $I \subset \mathbb{R}$,*

$$\left| \sum_{i=1}^n \mathbb{1}_{\{\lambda_i(\mathbf{H}) \in I\}} - n \int_I \rho_{sc}(t) dt \right| \leq n^\varepsilon.$$

This is a standard high-probability formulation of the local semicircle law and rigidity. Sharper logarithmic versions require corresponding sharper estimates; they are not proved here.

Now, suppose $Y_1, Y_2, \dots, Y_n \stackrel{i.i.d.}{\sim} \rho_{sc}$. Then the number of points falling in an interval $I \subseteq [-2, 2]$ is given by $\sum_{i=1}^n \mathbf{1}\{Y_i \in I\}$ having variance,

$$\text{Var} \left[\sum_{i=1}^n \mathbf{1}\{Y_i \in I\} \right] = np(1-p)$$

where $p = \int_I \rho_{sc}(t) dt$. For a fixed interval with $0 < p < 1$, we have $np(1-p) \asymp n$ and thus $\sum_{i=1}^n \mathbf{1}\{Y_i \in I\}$ has fluctuations of order $O(\sqrt{n})$. Thus independent points have fluctuations of order $\sqrt{n}$, much larger than n^ε for $\varepsilon < 1/2$. This comparison illustrates the stronger rigidity of eigenvalues; by itself it does not establish the optimal rate.

Note that Theorem 2.2.7 estimates the number of eigenvalues present in an interval between $[-2, 2]$. In the following, we show that Theorem 2.2.7 can be used to estimate the eigenvalue. Consider the upper-tail (i/n) quantiles of the semicircle distribution, that is, $2 = \gamma_0 > \gamma_1 > \dots > \gamma_n = -2$, defined by

$$\frac{i}{n} = \int_{\gamma_i}^2 \rho_{sc}(t) dt \quad (2.2.2)$$

Then using the above relation and the local behaviour of ρ_{sc} it is immediate that,

$$\frac{i}{n} \asymp (2 - \gamma_i)^{3/2} \implies 2 - \gamma_i \asymp (i/n)^{2/3}, \quad 1 \leq i \leq n/2$$

Now consider the interval $[\gamma_{i+1}, \gamma_i]$. The area under the curve ρ_{sc} in that interval is $\frac{1}{n}$, and hence roughly speaking the length of the interval is given by,

$$\gamma_i - \gamma_{i+1} \approx \frac{1}{n\rho_{sc}(\gamma_i)} \asymp \frac{1}{n\sqrt{2 - \gamma_i}} \approx \frac{1}{n^{2/3}i^{1/3}}$$

The spacing estimate suggests a corresponding statement for individual eigenvalues. Control of possible eigenvalues outside $[-2, 2]$ is also required; it does not follow from counts restricted to intervals inside $[-2, 2]$ alone.

Theorem 2.2.8 (Eigenvalue rigidity II). *For every fixed $\varepsilon, D > 0$, for all sufficiently large n , with probability at least $1 - n^{-D}$,*

$$|\lambda_i(\mathbf{H}) - \gamma_i| \leq \frac{n^\varepsilon}{n^{2/3} \min\{i, n+1-i\}^{1/3}}, \quad 1 \leq i \leq n.$$

This is the standard rigidity theorem, quoted here without proof. In the bulk, a sharper statement is available: for fixed $0 < \kappa < 1/2$ and $D > 0$, there is $C = C(\kappa, D)$ such that

$$\mathbb{P} \left(\max_{\kappa n \leq i \leq (1-\kappa)n} |\lambda_i - \gamma_i| > \frac{C \log n}{n} \right) \leq Cn^{-D}.$$

This logarithmic bulk bound is Theorem 1.8(ii) of Bourgade, Lopatto, and Zeitouni, Optimal rigidity and maximum of the characteristic polynomial of Wigner matrices (arXiv:2312.13335).

Having discussed the concentration of eigenvalues, it is now natural to ask about the limiting law of the same. If $\mathbf{H}$ is a GOE/GUE then we have the following limiting behavior of the largest eigenvalue of $\mathbf{H}$.

$$n^{2/3}(\lambda_1(\mathbf{H}) - 2) \xrightarrow{D} \text{TW}_\beta \text{ (Tracy-Widom distribution)}$$

where we have $\beta = 1$ if $\mathbf{H}$ is a GOE and $\beta = 2$ if $\mathbf{H}$ is GUE. The same result holds true for Wigner matrices as long as $\mathbb{E}|h_{ij}|^4 < \infty$.

2.3 Lecture 3: Marchenko-Pastur Distribution

We begin this lecture by defining and exploring properties of the **Stieltjes transform**.

Definition 2.3.1 (Stieltjes transform). Given a probability measure μ on $\mathbb{R}$, define its Stieltjes transform $S_\mu : \mathbb{C} \setminus \text{supp}(\mu) \rightarrow \mathbb{C}$ by

$$S_\mu(z) = \int \frac{d\mu(x)}{x - z}. \quad (2.3.1)$$

Example 2.3.2. For the semi-circle distribution with density given by

$$\rho_{sc}(x) = \frac{1}{2\pi} \sqrt{4 - x^2} \mathbb{1}(|x| < 2) \quad (2.3.2)$$

the Stieltjes transform is given by

$$S_{\rho_{sc}}(z) = \frac{-z + \sqrt{z^2 - 4}}{2}, \quad z \in \mathbb{C} \setminus [-2, 2]$$

Here the square root is the analytic branch on $\mathbb{C} \setminus [-2, 2]$ satisfying $\sqrt{z^2 - 4} \sim z$ at infinity, so that $S_{\rho_{sc}}(z) \sim -1/z$.

Properties of Stieltjes transform

- **Stieltjes transform as a generating function for moments**

$$\frac{1}{x - z} = \frac{1}{z} \cdot \frac{1}{\frac{x}{z} - 1} = -\frac{1}{z} \sum_{i=0}^{\infty} \left(\frac{x}{z}\right)^i = -\sum_{i=0}^{\infty} \frac{x^i}{z^{i+1}} \text{ for } |z| > |x|.$$

Thus, if μ is supported in $[-R, R]$, the series converges absolutely for $|z| > R$, and we have

$$S_\mu(z) = -\sum_{i=0}^{\infty} \frac{\int x^i d\mu}{z^{i+1}} = -\sum_{i=0}^{\infty} \frac{\mathbb{E}_{X \sim \mu}[X^i]}{z^{i+1}}.$$

- $\text{Im}(z) > 0$ iff $\text{Im}[S_\mu(z)] > 0$.

To see this,

$$\text{Im}[S_\mu(z)] = \text{Im} \left[\int \frac{d\mu(x)}{x - z} \right] = \int \text{Im} \left[\frac{1}{x - z} \right] d\mu(x) = \int \frac{\text{Im}(z)}{|x - z|^2} d\mu(x)$$

- One can recover μ from $S_\mu(z)$.

For $z = E + i\eta$, $\eta > 0$,

$$\frac{1}{\pi} \Im S_\mu(E + i\eta) = \int \frac{\eta}{\pi((x - E)^2 + \eta^2)} d\mu(x).$$

This is the density of μ convolved with the centered Cauchy law of scale η . Consequently

$$\frac{1}{\pi} \Im S_\mu(E + i\eta) dE \implies \mu \quad (\eta \downarrow 0).$$

This statement concerns measures; it does not assert pointwise convergence to a density when μ has atoms.

- For probability measures, $\{\mu_n\}_{n=1}^\infty \rightarrow \mu$ weakly iff $S_{\mu_n}(z) \rightarrow S_\mu(z)$ for each $z \in \mathbb{C}^+$.

Example 2.3.3 (Stieltjes transform continuity theorem). μ_n is a sequence of random probability measures and μ is a deterministic probability measure. Then

$$\mu_n \rightarrow \mu \text{ weakly in probability/almost surely} \iff S_n(z) \xrightarrow[\text{in p/a.s.}]{} S_\mu(z) \quad \forall z \in \mathbb{C}^+$$

For the almost-sure statement it suffices to first impose convergence on a countable dense set of $z \in \mathbb{C}^+$: the bound $|S_\mu(z)| \leq (\Im z)^{-1}$ and analyticity then extend the limit to all z on one event of probability one.

□

Next, consider a symmetric matrix $A = O\Lambda O^T$ where O is an orthogonal matrix, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ such that $\lambda_1 \geq \dots \geq \lambda_n$. Then,

$$(A - z)^{-1} = (O(\Lambda - z)O^T)^{-1} = O(\Lambda - z)^{-1}O^T$$

and thus,

$$\frac{1}{n} \text{Tr}(A - z)^{-1} = \frac{1}{n} \text{Tr}(\Lambda - z)^{-1} = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i - z}$$

We know by Theorem 2.1.1 that if H is a Wigner matrix of order n , then the empirical distribution of its eigenvalues converges to $\rho_{sc}(x)$ in probability as $n \rightarrow \infty$. Now, we also have,

$$S_n(z) = \frac{1}{n} \text{Tr}(H - z)^{-1} \rightarrow S(z) = \frac{-z + \sqrt{z^2 - 4}}{2}.$$

Clearly $S(z)$ satisfies $S^2(z) + zS(z) + 1 = 0$. Show that $S_n^2(z) + zS_n(z) + 1 = o_n(1)$.

Now, we shall deal with the empirical distribution of the eigenvalues of the sample covariance matrix. Suppose $X = n^{-1/2}[X_1, \dots, X_n]$, where the coordinates of the vectors $X_i \in \mathbb{R}^p$ are i.i.d. real random variables from a fixed distribution with mean zero and variance one. The matrix $XX^T = n^{-1} \sum_{i=1}^n X_i X_i^T \in \mathbb{R}^{p \times p}$ has eigenvalues $\lambda_1 \geq \dots \geq \lambda_p$.

In the classical setting, as p is fixed when $n \rightarrow \infty$, we have, $\frac{1}{n} \sum_{i=1}^n X_i X_i^T \rightarrow I_p$, and hence $\frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i} \rightarrow \delta_1$.

Next we shall consider the high dimensional regime $\frac{p}{n} \rightarrow \lambda \in (0, \infty)$.

Theorem 2.3.4 (Marchenko-Pastur theorem). *Under the preceding assumptions, when $p/n \rightarrow \lambda \in (0, \infty)$, the empirical measure $p^{-1} \sum_{i=1}^p \delta_{\lambda_i}$ converges weakly in probability to μ_{MP}^λ . Define $\lambda_\pm = (1 \pm \sqrt{\lambda})^2$ and $\rho_{MP}(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi\lambda x} \mathbb{1}_{\{\lambda_- < x < \lambda_+\}}$. Then μ_{MP}^λ is defined as follows*

(i) $0 < \lambda < 1$. Then $d\mu_{MP}(x) = \rho_{MP}(x) dx$.

(ii) $\lambda = 1$, then $\lambda_+ = 4$, $\lambda_- = 0$, and thus

$$\frac{d\mu_{MP}}{dx}(x) = \frac{1}{2\pi} \sqrt{\frac{4-x}{x}} \mathbb{1}_{\{0 < x < 4\}}$$

(iii) $\lambda > 1$. Then $\mu_{MP} = (1 - \lambda^{-1})\delta_0 + \rho_{MP}(x) dx$. The continuous part already has mass $1/\lambda$.

Remark 2.3.5. It should be evident that when $\lambda > 1$, there should exist a mass at 0 for the limiting distribution. This is because, when $\lambda > 1$, we have $p > n$, and there are at least $p - n$ zero eigenvalues. Their forced proportion tends to $1 - 1/\lambda$, which is the atom in the limit.

Next, we will recap some linear algebra facts and use them to prove Marchenko-Pastur Theorem.

Linear algebra facts

Consider $X \in \mathbb{R}^{p \times n}$ where $p \leq n$.

- Singular value decomposition (SVD). X has SVD of the form $X = USV^T$, where $U \in \mathbb{R}^{p \times p}$, $S \in \mathbb{R}^{p \times n}$, $V \in \mathbb{R}^{n \times n}$. Then $XX^T = U(SS^T)U^T$, $X^T X = V(S^T S)V^T$, which implies that the eigenvalues of XX^T , denoted by $\{\lambda_1, \dots, \lambda_p\}$, equal $\{s_1^2, \dots, s_p^2\}$. $X^T X$ has these p eigenvalues and $n - p$ additional zeros.

- Consider $Y = \begin{bmatrix} 0 & X^T \\ X & 0 \end{bmatrix} \in \mathbb{R}^{(p+n) \times (p+n)}$. It has at least $(n - p)$ -many zero eigenvalues, its remaining eigenvalues, counted with multiplicity, are $\pm s_1, \dots, \pm s_p$ (some s_i may also be zero). From $X = USV^T$, we have $Xv_i = s_i u_i$, $X^T u_i = s_i v_i$. It follows $Y \begin{bmatrix} v_i \\ u_i \end{bmatrix} = s_i \begin{bmatrix} v_i \\ u_i \end{bmatrix}$ and $Y \begin{bmatrix} v_i \\ -u_i \end{bmatrix} = -s_i \begin{bmatrix} v_i \\ -u_i \end{bmatrix}$.

- Schur complement formula.

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right]^{-1} = \left[\begin{array}{cc} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{array} \right]$$

Assuming D and $A - BD^{-1}C$ are invertible, this formula computes the inverse of a large matrix by inverting the smaller matrices D and $A - BD^{-1}C$ instead.

- Cauchy interlacing. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$. $A^{(i)} \in \mathbb{R}^{(n-1) \times (n-1)}$ is from A by removing i -th row and i -th column. Then $A^{(i)}$ is symmetric with eigenvalues $\mu_1 \geq \dots \geq \mu_{n-1}$, which interlace with eigenvalues of A : $\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \dots \geq \mu_{n-1} \geq \lambda_n$.
- Rank-1 perturbation. A is a symmetric matrix with eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$. Consider the rank-1 perturbation: $A + vv^T$ with eigenvalues $\mu_1 \geq \dots \geq \mu_n$. Then they are interlaced: $\mu_1 \geq \lambda_1 \geq \mu_2 \geq \lambda_2 \geq \dots \geq \mu_n \geq \lambda_n$.

Now, we will use the above facts in linear algebra to prove Marchenko-Pastur Theorem.

Proof of Theorem 2.3.4 for sample covariance matrix. Instead of XX^T , we study $Y = \begin{bmatrix} 0 & X^T \\ X & 0 \end{bmatrix}$. For $z \in \mathbb{C}^+$ let $G(z) = (Y - zI)^{-1}$. Schur complements give

$$G(z) = \begin{bmatrix} (X^T X / z - zI_n)^{-1} & * \\ * & (X X^T / z - zI_p)^{-1} \end{bmatrix}.$$

Write $x_i = X_i/\sqrt{n}$ for the i -th column of X , and let $X^{(i)}$ be X with that column removed. Let $Y^{(i)}$ be Y with its i -th row and column removed, where $1 \leq i \leq n$ indexes the first block. For $i = 1$,

$$Y - zI = \begin{bmatrix} -z & 0 & x_1^T \\ 0 & -zI_{n-1} & (X^{(1)})^T \\ x_1 & X^{(1)} & -zI_p \end{bmatrix}.$$

Hence, writing $G^{(1)} = (Y^{(1)} - zI)^{-1}$,

$$\begin{aligned} G_{11}(z) &= \left(-z - [0, x_1^T] G^{(1)} \begin{bmatrix} 0 \\ x_1 \end{bmatrix} \right)^{-1} \\ &= (-z - x_1^T A_1(z) x_1)^{-1}, \\ A_1(z) &= \left(X^{(1)} (X^{(1)})^T / z - zI_p \right)^{-1}. \end{aligned}$$

The key point is that x_1 and $A_1(z)$ are independent, since $A_1(z)$ is computed from the other columns $x_2, \dots, x_n$. This is the “leave-one-out” argument.

We will continue this proof in the next lecture. □

2.4 Lecture 4: Marchenko-Pastur Distribution Continued

Recalling our set-up, we have $X = n^{-1/2}(X_{ij})_{i=1,\dots,p,j=1,\dots,n}$ with the X_{ij} iid with mean 0 and variance 1. The sample covariance matrix is given by the $p \times p$ matrix XX^T , with eigenvalues denoted by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. For this section we initially assume $p \leq n$ and $p/n \rightarrow \lambda \in (0, 1]$. For the elementary fourth-moment concentration proof below, assume also $\mathbb{E}|X_{ij}|^8 \leq M_8 < \infty$. The finite-variance extension is explained after the proof.

The Stieltjes transform of the empirical eigenvalue distribution is

$$\int \frac{1}{x-w} d\mu_n = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - w} = \frac{1}{p} \text{Tr}(XX^T - w)^{-1}.$$

We also have

$$\frac{1}{p} \text{Tr}(X^T X - w)^{-1} = \frac{1}{p} \left(\sum_{i=1}^p \frac{1}{\lambda_i - w} - \frac{n-p}{w} \right)$$

because $X^T X$ has the p eigenvalues of XX^T together with $n-p$ additional zeros. Last lecture we used Schur complements to find

$$(Y-z)_{11}^{-1} = \left(-z - x_1^\top \left(\frac{X^{(1)}(X^{(1)})^\top}{z} - z \right)^{-1} x_1 \right)^{-1}$$

which we need to evaluate the trace which gives us the Stieltjes transform of μ_n . If we can find $(Y-z)_{11}^{-1}$ we can find $(Y-z)_{jj}^{-1}$ for $j > 1$ so we will focus on this case. Our move is to use the independence of $X^{(1)}$ and x_1 to allow us to use concentration inequalities.

Lemma 2.4.1. *Let $A \in \mathbb{C}^{p \times p}$ be deterministic and symmetric, and let $X = p^{-1/2}(X_1, \dots, X_p)^T$, where the real coordinates X_i are independent, centered, of variance one, with $\mathbb{E}|X_i|^8 \leq M_8$. If $\|A\| \leq C$, then*

$$\mathbb{P} \left(\left| X^T A X - \frac{1}{p} \text{Tr} A \right| > \varepsilon \right) \leq \frac{C M_8 C^4}{p^2 \varepsilon^4}.$$

The statement also holds conditionally when A is random and independent of X .

Proof of Lemma 2.4.1. Independence and unit variance give $\mathbb{E}[X^T A X] = p^{-1} \text{Tr} A$. Write

$$pX^T A X - \text{Tr} A = \sum_i A_{ii}(X_i^2 - 1) + 2 \sum_{i < j} A_{ij} X_i X_j.$$

Expanding the fourth absolute moment, independence eliminates terms with a coordinate occurring only once. The remaining terms are bounded by the eighth-moment assumption and Cauchy-Schwarz, giving

$$\mathbb{E}|pX^T A X - \text{Tr} A|^4 \leq C_{M_8} \|A\|_F^4 \leq C_{M_8} C^4 p^2.$$

For complex A , apply the real estimate to its real and imaginary parts and use $|u+iv|^4 \leq 2(u^4+v^4)$. This is also the $q=4$ quadratic-form bound in Lemma B.26 of Bai and Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*. Dividing by p^4 and applying Markov's inequality proves the claim. $\square$

Applying this result conditionally on $X^{(i)}$, and recalling $x_i = n^{-1/2}X_i$, we have

$$x_i^T A_i(z) x_i = \frac{1}{n} \operatorname{Tr} A_i(z) + o_{\mathbb{P}}(1).$$

Fix z in the first quadrant and put $w = z^2 \in \mathbb{C}^+$. The matrix $A_i(z)$ is a block of a Hermitian resolvent, so $\|A_i(z)\| \leq (\Im z)^{-1}$. The error is uniform over $i \leq n$: for example, the union bound at tolerance $n^{-1/8}$ has failure probability $\mathcal{O}(n^{-1/2})$ when p/n stays bounded away from zero and infinity. Thus

$$G_{ii}(z) = \left(-z - \frac{1}{n} \operatorname{Tr} A_i(z) + o_{\mathbb{P}}(1) \right)^{-1}, \quad i \leq n,$$

uniformly in i . The imaginary part of $z + x_i^T A_i(z) x_i$ is at least $\Im z$, so taking inverses does not amplify these fixed- z errors uncontrollably.

Next we want to write this in terms of X rather than $X^{(i)}$.

Claim 2.4.2. *For fixed $w = z^2 \in \mathbb{C}^+$,*

$$\begin{aligned} \frac{1}{p} \operatorname{Tr} A_i(z) &= \frac{z}{p} \operatorname{Tr} \left(X^{(i)} (X^{(i)})^T - w I_p \right)^{-1} \\ &= \frac{z}{p} \operatorname{Tr} (X X^T - w I_p)^{-1} + \mathcal{O} \left(\frac{|z|}{p \Im w} \right). \end{aligned}$$

Indeed, $X X^T = X^{(i)} (X^{(i)})^T + x_i x_i^T$. Rank-one interlacing bounds the difference of the eigenvalue counting functions by one. Integration by parts against $(x - w)^{-1}$ gives a trace difference bounded by $\int_{\mathbb{R}} |x - w|^{-2} dx = \pi / \Im w$, proving the claim.

Denote $c_n = p/n$ and $S_n(w) = p^{-1} \operatorname{Tr} (X X^T - w I_p)^{-1}$. Then

$$G_{ii}(z) = \frac{1}{-z - z c_n S_n(z^2) + o_{\mathbb{P}}(1)}, \quad i \leq n.$$

On the other hand, the first diagonal block of G gives the exact identity

$$\frac{1}{n} \sum_{i=1}^n G_{ii}(z) = \frac{z}{n} \operatorname{Tr} (X^T X - z^2 I_n)^{-1} = c_n z S_n(z^2) + \frac{c_n - 1}{z}.$$

Averaging the preceding approximation, multiplying by z , and setting $w = z^2$ yields

$$c_n w S_n(w) + c_n - 1 = \frac{1}{-1 - c_n S_n(w) + o_{\mathbb{P}}(1)}.$$

Since $|S_n(w)| \leq (\Im w)^{-1}$, we obtain

$$c_n w S_n(w)^2 + (w + c_n - 1) S_n(w) + 1 = o_{\mathbb{P}}(1).$$

This equation alone must not be used to assume convergence. The resolvent bound makes the family locally bounded and hence normal in $\mathbb{C}^+$. Also $p^{-1} \operatorname{Tr} X X^T = (np)^{-1} \sum_{ij} X_{ij}^2 \rightarrow 1$ in probability, which gives tightness of the empirical measures. Along any subsequence, extract a further subsequence on which these bounds and the equation on a countable dense set hold almost

surely. Every limiting Stieltjes transform is analytic, is the transform of a probability measure on $[0, \infty)$, and solves

$$\lambda w S(w)^2 + (w + \lambda - 1)S(w) + 1 = 0.$$

The probability-measure normalization $S(w) \sim -1/w$ at infinity selects the unique branch

$$S(w) = \frac{-(w + \lambda - 1) + \sqrt{(w - \lambda_-)(w - \lambda_+)}}{2\lambda w}, \quad \sqrt{(w - \lambda_-)(w - \lambda_+)} \sim w.$$

Consequently $S_n(w) \rightarrow S(w)$ in probability. Stieltjes inversion gives

$$\lim_{\eta \downarrow 0} \frac{1}{\pi} \Im S(x + i\eta) = \frac{\sqrt{(x - \lambda_-)(\lambda_+ - x)}}{2\pi \lambda x}, \quad \lambda_- < x < \lambda_+.$$

For $\lambda \leq 1$ this has total mass one. The same trace identities remain valid when $p > n$ (now the extra zeros are in XX^T), so the same argument proves the result for $\lambda > 1$; its atom at zero is $\lim_{\eta \downarrow 0} (-i\eta)S(i\eta) = 1 - 1/\lambda$.

Finally, to retain only the stated finite-variance assumption, center and truncate the entries and then normalize their variance as in Lecture 1. The inequality $\sum_j (s_j(X) - s_j(\tilde{X}))^2 \leq \|X - \tilde{X}\|_F^2$, followed by Cauchy–Schwarz and the bounded mean squared singular values, makes the normalized sum of $|s_j(X)^2 - s_j(\tilde{X})^2|$ small in expectation as the truncation level tends to infinity. This transfers weak convergence from the truncated matrices to the original matrices.

As a final detail, the preceding proof uses independent whitened coordinates, not merely the covariance identity $\text{cov}[X_i] = I$. In practice we might have $\text{cov}[X_i] = \Sigma$. In the following paragraph, let $X = [X_1, \dots, X_n]$ and $Z = [Z_1, \dots, Z_n]$ denote *unnormalized* data matrices, with $X_i = \Sigma^{1/2} Z_i$ and the coordinates of Z_i i.i.d., centered, and of variance one. Then

$$\frac{1}{n} XX^T = \frac{1}{n} \Sigma^{1/2} ZZ^T \Sigma^{1/2}.$$

Isotropy alone, without suitable assumptions on the coordinate dependence, is not enough for the Marchenko–Pastur conclusion.

A similar moment method lets us study the empirical eigenvalue distribution of XX^T/n . Denote the eigenvalues of Σ by $\sigma_1 \geq \dots \geq \sigma_p$, and those of XX^T/n by $\lambda_1 \geq \dots \geq \lambda_p$.

Theorem 2.4.3. *Suppose $p/n \rightarrow \lambda \in (0, \infty)$, $\Sigma = \Sigma_p$ is deterministic and positive semidefinite, $\sup_p \|\Sigma_p\| < \infty$, and its empirical measure converges weakly to ν . Under the preceding assumptions on Z ,*

$$\frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i} \longrightarrow \nu \boxtimes \mu_{MP}^\lambda$$

weakly in probability, where $\boxtimes$ denotes free multiplicative convolution.

Chapter 3

Non-asymptotic Analysis

We begin by considering the problem of solving $Ax = b$ for $A \in \mathbb{R}^{n \times d}$. If this is overconstrained then there may not exist a solution so instead we will seek the x that is as close to solving the equation as possible; that is, we solve $\operatorname{argmin}_x \|Ax - b\|$ for some norm. For the rest of this discussion we will use the L_2 norm but this could be done with a different norm if desired. Using the L_1 norm would lead to a more robust solution but it is mathematically harder to work with.

Taking derivatives of the squared Euclidean objective, which has the same minimizers, we obtain the normal equations $A^T Ax = A^T b$ which govern the critical points of this optimization. If A has full column rank then $\hat{x} = (A^T A)^{-1} A^T b$.

Definition 3.0.1 (Condition number). For a matrix A , we define its condition number to be $\frac{S_{\max}(A)}{S_{\min}(A)}$ where $S_j(A)$ denotes the j th singular value of A .

We have two main questions now:

1. What can we say about the condition number of A if A is a random matrix?
2. The time complexity of solving the normal equations is $O(nd^2)$. Can we find some faster algorithm, perhaps if we are willing to sacrifice some accuracy? To be more precise, can we find an x with $\|Ax - b\|_p \leq (1 + \varepsilon) \|Ax^* - b\|_p$ ($p = 1, 2$) for some small $\varepsilon > 0$, and x^* denoting the true solution?

The latter question leads us to *sketching*. Let $S \in \mathbb{R}^{r \times n}$ for $r \ll n$ (since $n \gg d$ is the dominating term in the time complexity) and compute $\tilde{A} := SA$ and $\tilde{b} := Sb$. We can now solve

$$\min \|SAx - Sb\|_2 = \min \|\tilde{A}x - \tilde{b}\|_2$$

in $O(rd^2)$ time. We want to understand when the solution to this problem will be close to x^* .

The key idea that if S is an *isometry*, so it preserves the distance from b to the column space of A , denoted $\operatorname{col}(A)$, then we will get a good result.

If $(1 - \varepsilon)\|y\|_2 \leq \|Sy\|_2 \leq (1 + \varepsilon)\|y\|_2$ for $y \in \operatorname{span}(\operatorname{col}[A, b])$, then, as $Ax - b \in \operatorname{span}(\operatorname{col}[A, b])$, we will have

$$(1 - \varepsilon)\|Ax - b\|_2 \leq \|SAx - Sb\|_2 \leq (1 + \varepsilon)\|Ax - b\|_2$$

for any x .

If $\tilde{x}$ minimizes $\|SAx - Sb\|_2$ then we have

$$\begin{aligned} \|A\tilde{x} - b\|_2 &\leq \frac{1}{1-\varepsilon} \|SA\tilde{x} - Sb\|_2 \\ &\leq \frac{1}{1-\varepsilon} \|SAx^* - Sb\|_2 \\ &\leq \frac{1+\varepsilon}{1-\varepsilon} \|Ax^* - b\|_2 \end{aligned}$$

so we pay a price of order ε .

To find such an S , we can note that $(1-\varepsilon)\|y\|_2 \leq \|Sy\|_2 \leq (1+\varepsilon)\|y\|_2$ is equivalent to a condition on the singular values of S , but only over y in the column space of A and b which is a weaker requirement. So let $U \in \mathbb{R}^{n \times q}$, $q = \text{rank}[A, b] \leq d+1$, have orthonormal columns spanning this column space (which can be produced via Gram-Schmidt). Then $y = Uv$ for some v and $\|y\|_2 = \|v\|_2$ by the isometry of U .

This means $(1-\varepsilon)\|v\|_2 \leq \|SUv\|_2 \leq (1+\varepsilon)\|v\|_2$ and now v is unconstrained, so we can see that if we find an S such that with probability $1-\delta$

$$1-\varepsilon \leq S_{\min}(SU) \leq S_{\max}(SU) \leq 1+\varepsilon$$

then our required condition holds.

Theorem 3.0.2. *Let $A \in \mathbb{R}^{p \times n}$, $p \leq n$, be a random matrix with $A = [A_1, A_2, \dots, A_n]$ and A_i are independent sub-Gaussian vectors. Assume $\|A_i\|_{\psi_2} \leq k$, $\mathbb{E}[A_i] = 0$, and $\text{cov}[A_i] = I$. Then for $t \geq 0$, with probability at least $1 - 2e^{-c_k t^2}$, where $c_k, C_k > 0$ depend only on k , $\sqrt{n} - C_k \sqrt{p} - t \leq S_{\min}(A) \leq S_{\max}(A) \leq \sqrt{n} + C_k \sqrt{p} + t$*

Note that unless $n \gg p$ the lower bound here will be trivial. For i.i.d. centered unit-variance entries, the corresponding extreme-eigenvalue theorem gives $S_{\min}(A) \approx \sqrt{n} - \sqrt{p}$, $S_{\max}(A) \approx \sqrt{n} + \sqrt{p}$ with probability $1 - o(1)$. The theorem above gives better probability bounds.

Theorem 3.0.3. *Let $A \in \mathbb{R}^{p \times n}$ have independent centered entries with variance one and $\sup_{ij} \|A_{ij}\|_{\psi_2} \leq K < \infty$. If $\frac{p}{n} \rightarrow \lambda < 1$, then with probability $1 - e^{-c_n}$, $c_\lambda \sqrt{n} \leq S_{\min}(A) \leq S_{\max}(A) \leq C_\lambda \sqrt{n}$, with $0 < c_\lambda < C_\lambda$ depending on λ and the uniform sub-Gaussian bound.*

Note that this gives a nontrivial lower bound even if $n \sim p$, and the singular values are bounded above and below by positive constant multiples of $\sqrt{n}$. They are relatively close to $\sqrt{n}$ when $p/n \rightarrow 0$.

We will continue with the proofs of the two theorems next time.

3.1 Lecture 5: Largest and Smallest Singular Values

Theorem 3.1.1. *Let $A \in \mathbb{R}^{p \times n}$, $p \leq n$, be a random matrix with $A = [A_1, A_2, \dots, A_n]$ and A_i are independent sub-Gaussian vectors. Assume $\|A_i\|_{\psi_2} \leq k$, $\mathbb{E}[A_i] = 0$, and $\text{cov}[A_i] = I$. Then for $t \geq 0$, with probability at least $1 - 2e^{-c_k t^2}$, where $c_k, C_k > 0$ depend only on k , $\sqrt{n} - C_k \sqrt{p} - t \leq S_{\min}(A) \leq S_{\max}(A) \leq \sqrt{n} + C_k \sqrt{p} + t$*

Definition and properties of sub-Gaussian random variables and sub-Gaussian vectors:

Definition 3.1.2 (Sub-Gaussian random variable). A real random variable X is sub-Gaussian if the first two conditions below hold for some positive constants. For centered X , the third is equivalent as well:

- $\mathbb{P}(|X| > t) \leq 2e^{-\frac{t^2}{k_1^2}}$
- $\mathbb{E}\left(e^{\frac{X^2}{k_2^2}}\right) \leq 2$
- If $\mathbb{E}(X) = 0$, then $\mathbb{E}(e^{\lambda X}) \leq e^{k_3^2 \lambda^2}$

where $k_1, k_2, k_3 > 0$ are constants; the tail bound is for all $t \geq 0$, and the Gaussian moment-generating bound is for all $\lambda \in \mathbb{R}$.

With the second condition, we define the sub-Gaussian norm.

Definition 3.1.3 (Sub-Gaussian norm). Let $\|X\|_{\psi_2}$ denote the sub-Gaussian norm of X defined by

$$\|X\|_{\psi_2} = \inf \left\{ t > 0 : \mathbb{E}\left(e^{\frac{X^2}{t^2}}\right) \leq 2 \right\} \quad (3.1.1)$$

Lemma 3.1.4 (Sum of independent sub-Gaussian random variables). *If $X_1, X_2, \dots, X_n$ are independent sub-Gaussian random variables with $\mathbb{E}(X_i) = 0$, then $\sum_{i=1}^n X_i$ is a sub-Gaussian random variable and*

$$\left\| \sum_{i=1}^n X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^n \|X_i\|_{\psi_2}^2 \quad (3.1.2)$$

where C is a constant.

We will prove this property using the third condition.

Proof of Lemma 3.1.4.

$$\begin{aligned} \mathbb{E}\left(e^{\lambda \sum_{i=1}^n X_i}\right) &= \prod_{i=1}^n \mathbb{E}\left(e^{\lambda X_i}\right) \\ &\leq \prod_{i=1}^n e^{C\lambda^2 \|X_i\|_{\psi_2}^2} && \text{(by condition 3)} \\ &= e^{C\lambda^2 \sum_{i=1}^n \|X_i\|_{\psi_2}^2} \end{aligned}$$

where C is constant. Then by condition 3, $\sum_{i=1}^n X_i$ is sub-Gaussian and the square of the sub-Gaussian norm is bounded by $C \sum_{i=1}^n \|X_i\|_{\psi_2}^2$. $\square$

Instead of using condition 3, we get the concentration inequality for sum of independent sub-Gaussian random variables if we use condition 1.

Lemma 3.1.5 (concentration inequality for sum of independent sub-Gaussian random variables). *If $X_1, X_2, \dots, X_n$ are independent sub-Gaussian random variables with $\mathbb{E}(X_i) = 0$, then*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{Ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right) \quad (3.1.3)$$

Definition 3.1.6 (sub-Gaussian random vector). A random vector $\mathbf{X} \in \mathbb{R}^n$ is a sub-Gaussian random vector if any linear combination of its coordinates is sub-Gaussian, and the sub-Gaussian norm is defined as

$$\|\mathbf{X}\|_{\psi_2} = \sup_{\|a\|_2=1} \|\langle \mathbf{X}, a \rangle\|_{\psi_2} \quad (3.1.4)$$

Example 3.1.7. If a random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$, where $X_1, \dots, X_n$ are independent centered sub-Gaussian random variables, and $\|X_i\|_{\psi_2} \leq k$, k is constant, then

$$\|\mathbf{X}\|_{\psi_2} = \sup_{\|a\|=1} \left\| \sum_{i=1}^n a_i X_i \right\|_{\psi_2}$$

Take the square on both sides,

$$\begin{aligned} \|\mathbf{X}\|_{\psi_2}^2 &= \sup_{\|a\|=1} \left\| \sum_{i=1}^n a_i X_i \right\|_{\psi_2}^2 \\ &\leq \sup_{\|a\|=1} C \sum_{i=1}^n \|a_i X_i\|_{\psi_2}^2 \quad (\text{by Lemma 3.1.4}) \\ &= \sup_{\|a\|=1} C \sum_{i=1}^n a_i^2 \|X_i\|_{\psi_2}^2 \\ &\leq Ck^2 \end{aligned}$$

So the sub-Gaussian norm of $\mathbf{X}$ is bounded by Ck , where C is an absolute constant.

Another related concept is sub-Exponential.

Definition 3.1.8 (sub-Exponential random variable). A real random variable X is sub-exponential if the first two conditions below hold for some positive constants. For centered X , the third is equivalent as well:

- $\mathbb{P}(|X| > t) \leq 2e^{-\frac{t}{k_1}}$
- $\mathbb{E}\left(e^{\frac{|X|}{k_2}}\right) \leq 2$
- If $\mathbb{E}(X) = 0$, then $\mathbb{E}(e^{\lambda X}) \leq e^{k_3^2 \lambda^2}$, for $|\lambda| \leq \frac{1}{k_3}$

where $k_1, k_2, k_3 > 0$ are constants, and the tail bound is for all $t \geq 0$.

The sub-exponential norm is the infimum of the positive constants k_2 for which condition 2 holds.

Definition 3.1.9 (sub-Exponential norm). $\|X\|_{\psi_1}$ denotes the sub-Exponential norm of X , which is defined by

$$\|X\|_{\psi_1} = \inf \left\{ t > 0 : \mathbb{E}\left(e^{\frac{|X|}{t}}\right) \leq 2 \right\} \quad (3.1.5)$$

The product of two sub-Gaussian random variables is sub-exponential; independence is not needed.

Claim 3.1.10. *If X, Y are sub-Gaussian random variables, then XY is sub-exponential and the sub-Exponential norm is bounded:*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2} \quad (3.1.6)$$

Indeed, $|XY|/(ab) \leq X^2/(2a^2) + Y^2/(2b^2)$ and Cauchy-Schwarz imply $\mathbb{E}e^{|XY|/(ab)} \leq 2$ for $a > \|X\|_{\psi_2}$ and $b > \|Y\|_{\psi_2}$; take infima.

Similarly, there is a concentration inequality for the sum of independent sub-Exponential random variables, which is also known as the Bernstein inequality.

Theorem 3.1.11 (Bernstein inequality). *If $X_1, \dots, X_n$ are independent sub-Exponential random variables with $\mathbb{E}(X_i) = 0$, then*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > t\right) \leq 2 \exp\left(-C \min\left\{\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{1 \leq i \leq n} \|X_i\|_{\psi_1}}\right\}\right) \quad (3.1.7)$$

To prove theorem 3.0.2, we still need the concept of ϵ -net.

Definition 3.1.12 (ϵ -net). An ϵ -net, denoted by N_ϵ , is a collection of points on S^{n-1} that for any $x \in S^{n-1}$, there is $y \in N_\epsilon$ such that $\|x - y\| \leq \epsilon$.

Claim 3.1.13. *There is N_ϵ such as $|N_\epsilon| \leq (1 + \frac{2}{\epsilon})^n$*

Proof of Claim 3.1.13. Greedy algorithm to construct the ϵ -net: pick one point, and keep adding new points that the distance is greater than or equal to ϵ with the previous points, until no more points could be added. With the Greedy algorithm, we get a maximal ϵ -separated set N_ϵ such that for $x, y \in N_\epsilon$, $\|x - y\|_2 \geq \epsilon$. For $x \in N_\epsilon$, put a ball with center x and radius $\frac{\epsilon}{2}$. Recall that the volume of a radius r ball is

$$\text{vol}(B_r) = r^n \text{vol}(B_1)$$

Sum of volumes of all radius $\frac{\epsilon}{2}$ balls are bounded by the volume of the ball with radius $1 + \frac{\epsilon}{2}$, then

$$|N_\epsilon| \left(\frac{\epsilon}{2}\right)^n \text{vol}(B_1) \leq \left(1 + \frac{\epsilon}{2}\right)^n \text{vol}(B_1)$$

$\text{vol}(B_1)$ cancels out, which gives

$$|N_\epsilon| \leq \left(1 + \frac{2}{\epsilon}\right)^n$$

□

Claim 3.1.14. $\max_{x \in N_\epsilon} \|Ax\| \leq \|A\| \leq \frac{1}{1-\epsilon} \max_{x \in N_\epsilon} \|Ax\|$

Proof of Claim 3.1.14. $\|A\| = \sup_{x \in S^{n-1}} \|Ax\|$ and then the lower bound follows trivially. For the upper bound, suppose $\|A\| = \|Ax^*\|$ with $x^* \in S^{n-1}$, and assume $0 < \epsilon < 1$. By the definition of ϵ -net, there is $y \in N_\epsilon$ such that $\|x^* - y\| \leq \epsilon$. Then

$$\begin{aligned} \|A\| &= \|Ax^*\| \\ &= \|Ay + A(x^* - y)\| \\ &\leq \|Ay\| + \|A(x^* - y)\| \\ &\leq \|Ay\| + \epsilon \|A\| \end{aligned}$$

Then, $(1 - \epsilon)\|\mathbf{A}\| \leq \|\mathbf{A}y\|$ and $\|\mathbf{A}\| \leq \frac{1}{1-\epsilon}\|\mathbf{A}y\|$. $\square$

Claim 3.1.15. *If $\mathbf{A}$ is symmetric and $0 < \epsilon < 1/2$, $\|\mathbf{A}\| = \sup_{x \in S^{n-1}} |\langle x, \mathbf{A}x \rangle| \leq \frac{1}{1-2\epsilon} \max_{x \in N_\epsilon} |\langle x, \mathbf{A}x \rangle|$*

Proof of Claim 3.1.15. Choose $x^* \in S^{n-1}$ with $|\langle x^*, \mathbf{A}x^* \rangle| = \|\mathbf{A}\|$, and choose $y \in N_\epsilon$ with $\|x^* - y\| \leq \epsilon$. Then

$$\begin{aligned} |\langle x^*, \mathbf{A}x^* \rangle - \langle y, \mathbf{A}y \rangle| &\leq |\langle x^* - y, \mathbf{A}x^* \rangle| + |\langle y, \mathbf{A}(x^* - y) \rangle| \\ &\leq 2\epsilon\|\mathbf{A}\|. \end{aligned}$$

Thus $(1 - 2\epsilon)\|\mathbf{A}\| \leq |\langle y, \mathbf{A}y \rangle|$, proving the claim. $\square$

Next we will prove Theorem 3.0.2, which is stated at the end of last lecture.

Proof of Theorem 3.0.2. Because A is $p \times n$ with $p \leq n$, its p singular values are controlled by A^T on $\mathbb{R}^p$:

$$S_{\max}(A) = \sup_{x \in S^{p-1}} \|A^T x\|, \quad S_{\min}(A) = \inf_{x \in S^{p-1}} \|A^T x\|.$$

Let $N_{1/4} \subset S^{p-1}$ be a $1/4$ -net of size at most 9^p . Then

$$\|AA^T - nI_p\| \leq 2 \max_{x \in N_{1/4}} |\langle x, (AA^T - nI_p)x \rangle|.$$

For fixed $x \in S^{p-1}$, set $z_i = A_i^T x$. The z_i are independent, $\mathbb{E}z_i = 0$, $\mathbb{E}z_i^2 = x^T \mathbb{E}[A_i A_i^T] x = 1$, and $\|z_i\|_{\psi_2} \leq k$. Hence $\|z_i^2 - 1\|_{\psi_1} \leq Ck^2$, and Bernstein's inequality gives

$$\mathbb{P}\left(\left|\sum_{i=1}^n (z_i^2 - 1)\right| \geq nu\right) \leq 2 \exp\left[-cn \min\left\{\frac{u^2}{k^4}, \frac{u}{k^2}\right\}\right].$$

Take $\delta = C_k \sqrt{p/n} + t/\sqrt{n}$ and $u = \frac{1}{2} \max\{\delta, \delta^2\}$. With C_k sufficiently large, a union bound and the net estimate imply

$$\mathbb{P}(\|AA^T/n - I_p\| > \max\{\delta, \delta^2\}) \leq 2e^{-c_k t^2}.$$

Here $\min\{u^2/k^4, u/k^2\} \geq c_k \delta^2$, and the $C_k^2 p$ part of $n\delta^2$ absorbs $p \log 9$. On the complementary event, $|S_i(A)^2/n - 1| \leq \max\{\delta, \delta^2\}$. If $\delta \leq 1$, taking square roots gives $|S_i(A)/\sqrt{n} - 1| \leq \delta$; if $\delta > 1$, the lower bound is trivial and $\sqrt{1 + \delta^2} \leq 1 + \delta$ gives the upper bound. This proves the theorem. $\square$

Theorem 3.1.16. *Let $\mathbf{A} \in \mathbb{R}^{p \times n}$ have independent centered entries with variance one and $\sup_{ij} \|A_{ij}\|_{\psi_2} \leq K$. Suppose $p/n \leq \lambda < 1$, where λ is fixed. Then with probability $1 - e^{-cn}$,*

$$c_\lambda \sqrt{n} \leq S_{\min}(\mathbf{A}) \leq S_{\max}(\mathbf{A}) \leq C_\lambda \sqrt{n}$$

where c_λ and C_λ are constants depending only on λ and K .

This theorem provides better control for the smallest singular value when $p < n$. We will start the proof of this theorem in the next lecture.

3.2 Lecture 6: Randomized Algorithms

Theorem 3.2.1. *Let A be a $p \times n$ matrix with independent centered entries of variance one and $\sup_{i,j} \|A_{ij}\|_{\psi_2} \leq K$. If $p/n \leq \lambda < 1$, where λ is fixed, then with probability at least $1 - e^{-cn}$,*

$$c_\lambda \sqrt{n} \leq S_{\min}(A) \leq S_{\max}(A) \leq C_\lambda \sqrt{n},$$

where $c, c_\lambda, C_\lambda > 0$ depend only on λ, K . The probability statement is for all sufficiently large n ; constants may be adjusted to cover smaller dimensions.

- **The ε -net**

Lemma 3.2.2. *For any $0 < \varepsilon < 1$, there exists an ε -net $\mathcal{N}_\varepsilon$ of S^{p-1} such that*

$$|\mathcal{N}_\varepsilon| \leq (1 + 2/\varepsilon)^p.$$

Proof. The maximal separated-set and volume argument from the preceding lecture proves this assertion; no assumption $n \gg p$ is needed for the net itself. $\square$

For a fixed unit vector X , write $V_i = A_i^T X$, where A_i is the i -th column of A . Then $\mathbb{E}V_i^2 = 1$ and $\mathbb{E}V_i^4 \leq C_K$. By the Paley–Zygmund inequality,

$$\mathbb{P}(V_i^2 \geq 1/2) \geq \frac{1}{4C_K} =: q_K > 0.$$

The indicators are independent, so a scalar Chernoff bound shows that, for constants $a_K, b_K > 0$,

$$\mathbb{P}(\|A^T X\|_2 \leq a_K \sqrt{n}) = \mathbb{P}\left(\sum_{i=1}^n V_i^2 \leq a_K^2 n\right) \leq e^{-b_K n}.$$

A union bound over a fixed-resolution net proves the lower singular-value bound when p/n is sufficiently small. To treat every fixed $\lambda < 1$, we need a better estimate on most vectors.

- **The η -net**

Let C denote the compressible set defined below, and put $m = \min\{p, \lfloor \varepsilon^{-2} \rfloor\}$. The unit vectors supported on at most m coordinates have an η -net $\mathcal{M}_\eta$ of size

$$|\mathcal{M}_\eta| \leq \binom{p}{m} (1 + 2/\eta)^m.$$

For each support, use the sphere-net bound in that coordinate subspace, and then take the union over supports. A unit vector within η of an m -sparse vector of norm at most one is within 2η of an m -sparse unit vector, by normalizing that vector. Therefore $\mathcal{M}_\eta$ is a 3η -net for C .

The upper bound from the previous lecture gives an event $\mathcal{E} = \{\|A\| \leq C_K \sqrt{n}\}$ with $\mathbb{P}(\mathcal{E}^c) \leq e^{-c_K n}$, after adjusting constants. Fix $\eta < a_K/(6C_K)$. On $\mathcal{E}$, if some $X \in C$ has $\|A^T X\| \leq a_K \sqrt{n}/2$, then some $Y \in \mathcal{M}_\eta$ has $\|A^T Y\| \leq a_K \sqrt{n}$. Thus

$$\mathbb{P}\left(\inf_{X \in C} \|A^T X\| \leq a_K \sqrt{n}/2\right) \leq e^{-c_K n} + \binom{p}{m} (1 + 2/\eta)^m e^{-b_K n} \leq e^{-c'_K n}$$

for sufficiently large n , when ε, η are fixed. Indeed, m is then fixed and the logarithm of the net size is $\mathcal{O}_{\varepsilon, \eta}(\log n) = o(n)$.

• **Compressible and Incompressible Vectors**

We decompose S^{p-1} into two types of vectors. First fix $\eta > 0$ as above; later choose $\varepsilon \ll \eta$.

Definition 3.2.3. A vector $X \in S^{p-1}$ is called compressible if it is within Euclidean distance η of a vector of norm at most one supported on at most $m = \min\{p, \lfloor \varepsilon^{-2} \rfloor\}$ coordinates. It is incompressible otherwise. Denote these sets by C and I , respectively.

In particular, if $\sum_{|X_i| \leq \varepsilon} |X_i|^2 \leq \eta^2$, then X is compressible: define

$$\tilde{X}_i = \begin{cases} X_i, & |X_i| > \varepsilon, \\ 0, & |X_i| \leq \varepsilon. \end{cases}$$

Then $\|\tilde{X}\| \leq 1$, $\|X - \tilde{X}\|^2 = \sum_{|X_i| \leq \varepsilon} |X_i|^2 \leq \eta^2$, and $\tilde{X}$ is supported on at most m coordinates. Consequently every $X \in I$ satisfies

$$\sum_{|X_i| \leq \varepsilon} |X_i|^2 > \eta^2.$$

The converse of this last implication is not needed. A maximal ε -separated subset of I gives a net $\mathcal{N}_\varepsilon(I) \subset I$ with $|\mathcal{N}_\varepsilon(I)| \leq (1 + 2/\varepsilon)^p$.

We will show the improved estimate, for fixed $X \in I$ and $0 < \tau < 1$,

$$\mathbb{P}(\|A^T X\| \leq 2C_K \varepsilon \sqrt{n}) \leq C_{\eta, \tau, K}^n \varepsilon^{(1-\tau)n}.$$

How this estimate completes the theorem. On $\mathcal{E}$, approximating by $\mathcal{N}_\varepsilon(I)$ gives

$$\begin{aligned} & \mathbb{P}\left(\inf_{X \in I} \|A^T X\| \leq C_K \varepsilon \sqrt{n}, \mathcal{E}\right) \\ & \leq (1 + 2/\varepsilon)^p C_{\eta, \tau, K}^n \varepsilon^{(1-\tau)n} \leq (3C_{\eta, \tau, K})^n \varepsilon^{(1-\tau)n-p}. \end{aligned}$$

Choose $\tau = (1 - \lambda)/2$, so $(1 - \tau)n - p \geq (1 - \lambda)n/2$. Choosing $\varepsilon > 0$ sufficiently small, depending only on λ, K, η , makes the last bound exponentially small. Combining this with $\mathbb{P}(\mathcal{E}^c)$ and the compressible-vector estimate proves the lower bound. The upper bound is already included in $\mathcal{E}$. $\square$

We next establish the fixed-vector estimate used in this union bound.

Claim 3.2.4. For $X \in I$ and any fixed $L > 0$,

$$\sup_{u \in \mathbb{R}} \mathbb{P}(|\langle A_i, X \rangle - u| \leq L\varepsilon) \leq C_{\eta, L, K} \varepsilon.$$

- Let $V_i = \langle A_i, X \rangle$. If $\sum_{i=1}^n V_i^2 \leq 4C_K^2 \varepsilon^2 n$, then at least $\lceil (1 - \tau)n \rceil$ indices satisfy $|V_i| \leq 2C_K \varepsilon / \sqrt{\tau}$. Independence and Claim 3.2.4 therefore imply, after taking ε small enough,

$$\begin{aligned} \mathbb{P}\left(\sum_i V_i^2 \leq 4C_K^2 \varepsilon^2 n\right) & \leq 2^n (C_{\eta, \tau, K} \varepsilon)^{\lceil (1-\tau)n \rceil} \\ & \leq \tilde{C}_{\eta, \tau, K}^n \varepsilon^{(1-\tau)n}. \end{aligned}$$

- This is precisely the estimate required by the union bound over $\mathcal{N}_\varepsilon(I)$. The factor ε^{-p} from the net is defeated by $\varepsilon^{(1-\tau)n}$, because $p/n < 1 - \tau$.

Proof of Claim 3.2.4. Let $A_i = (a_1, \dots, a_p)^T$ and $X = (x_1, \dots, x_p)^T$. The central limit heuristic suggests that sums with many small coefficients have Gaussian small-ball probabilities. A quantitative statement, uniform in the center of the interval, is required here.

- ◊ For independent centered variables y_j with variances σ_j^2 , the Lyapunov condition

$$\frac{\sum_j \mathbb{E}|y_j|^{2+\delta}}{(\sum_j \sigma_j^2)^{1+\delta/2}} \rightarrow 0$$

for some $\delta > 0$ is sufficient for the normalized sum to converge to $N(0, 1)$. It implies the Lindeberg condition; these conditions are not identical.

- ◊ The Berry–Esseen theorem, with $v^2 = \sum_j \sigma_j^2 > 0$, gives

$$\sup_t \left| \mathbb{P} \left(v^{-1} \sum_j y_j \leq t \right) - \Phi(t) \right| \leq C \frac{\sum_j \mathbb{E}|y_j|^3}{v^3}.$$

Since the standard Gaussian density is bounded, it follows for $h > 0$ that

$$\sup_u \mathbb{P} \left(\left| \sum_j y_j - u \right| \leq h \right) \leq C \frac{h}{v} + C \frac{\sum_j \mathbb{E}|y_j|^3}{v^3}.$$

Here Φ is the standard Gaussian distribution function.

Decompose the sum as

$$\sum_{j=1}^p a_j x_j = \sum_{|x_j| \leq \varepsilon} a_j x_j + \sum_{|x_j| > \varepsilon} a_j x_j.$$

Condition on all variables in the second sum. This sum is now a fixed shift, and the first sum has variance $v^2 = \sum_{|x_j| \leq \varepsilon} x_j^2 > \eta^2$. Also

$$\sum_{|x_j| \leq \varepsilon} \mathbb{E}|a_j x_j|^3 \leq C_K \varepsilon \sum_{|x_j| \leq \varepsilon} x_j^2 = C_K \varepsilon v^2.$$

The preceding Berry–Esseen estimate, applied with $h = L\varepsilon$, is at most $C_K(L+1)\varepsilon/v \leq C_{\eta, L, K}\varepsilon$, uniformly in the conditioned shift. Averaging over that shift proves the claim. $\square$

• Sub-Gaussian Random Variables

Let's come back to the questions discussed at the beginning: What is the condition number of a random matrix?

- For independent centered unit-variance sub-Gaussian entries with a uniform sub-Gaussian bound, and $p/n \leq \lambda < 1$, we have $K(A) = S_{\max}(A)/S_{\min}(A) = \mathcal{O}_{\lambda, K}(1)$ with probability at least $1 - e^{-cn}$.

- If $p = n$, then $S_{\max}(A) \lesssim \sqrt{n}$ with exponentially high probability, but the smallest singular value is typically of order $n^{-1/2}$, not order $\sqrt{n}$.

Theorem 3.2.5. *Let $A \in \mathbb{R}^{n \times n}$ have independent centered unit-variance entries with $\sup_{ij} \|A_{ij}\|_{\psi_2} \leq K$. Then, for $\varepsilon \geq 0$,*

$$\mathbb{P}(S_{\min}(A) \leq \varepsilon n^{-1/2}) \leq C_K \varepsilon + e^{-c_K n}.$$

This is the square-matrix least-singular-value theorem; its proof requires an additional small-ball argument beyond the rectangular proof above.

Remark 3.2.6 (3.2.5). The scale $n^{-1/2}$ is suggested by the Marchenko–Pastur density when $p = n$:

$$\frac{1}{2\pi} \sqrt{\frac{4-x}{x}} \sim \frac{1}{\pi\sqrt{x}} \quad (x \downarrow 0).$$

Heuristically, the first eigenvalue of AA^T/n occurs when $n \int_0^x \rho_{MP}(t) dt$ is of order one, namely $x \asymp n^{-2}$; taking square roots gives $S_{\min}(A) \asymp n^{-1/2}$. The weak Marchenko–Pastur law alone does not prove this microscopic conclusion. The theorem and the upper bound imply $K(A) = \mathcal{O}_{\mathbb{P}}(n)$, meaning tightness of $K(A)/n$, not an exponentially high-probability bound $K(A) \leq Cn$ for a fixed C .

- For the sketching algorithm, now let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. We wish to solve $\operatorname{argmin}_x \|Ax - b\|_2$.
 1. Sample a random matrix $S \in \mathbb{R}^{r \times n}$, with $r \ll n$.
 2. Compute SA and Sb .
 3. Solve the smaller problem $(\star)$:

$$\tilde{x} \in \operatorname{argmin}_x \|(SA)x - Sb\|_2.$$

For sufficiently small r , step (3) is cheap. Let $U \in \mathbb{R}^{n \times q}$ have orthonormal columns spanning $\operatorname{col}[A, b]$, where $q \leq d + 1$. If S is an almost-isometry on this space, so that

$$(1 - \varepsilon)\|v\| \leq \|Sv\| \leq (1 + \varepsilon)\|v\|, \quad v \in \operatorname{col}[A, b],$$

then, writing $x^* \in \operatorname{argmin}_x \|Ax - b\|_2$,

$$\begin{aligned} \|A\tilde{x} - b\|_2 &\leq \frac{1}{1 - \varepsilon} \|SA\tilde{x} - Sb\|_2 \\ &\leq \frac{1}{1 - \varepsilon} \|SAx^* - Sb\|_2 \leq \frac{1 + \varepsilon}{1 - \varepsilon} \|Ax^* - b\|_2. \end{aligned}$$

Equivalently, the sufficient condition $\|(SU)^T(SU) - I_q\| \leq \delta < 1$ controls the squared norms and gives a residual factor $\sqrt{(1 + \delta)/(1 - \delta)} = 1 + \mathcal{O}(\delta)$ for small δ .

- Take $S = r^{-1/2}\tilde{S}$, where $\tilde{S} \in \mathbb{R}^{r \times n}$ has independent centered unit-variance entries with uniformly bounded sub-Gaussian norm. The rows of $\tilde{S}U$ are independent isotropic sub-Gaussian vectors. For $r \geq C_K(d + 1)/\delta^2$, the preceding singular-value theorem, applied to $(\tilde{S}U)^T$, implies

$$\|(SU)^T(SU) - I_q\| \leq \delta$$

with probability at least $1 - 2e^{-c_K(d+1)}$, after adjusting C_K . No additional division by r is present: it is already included in the normalization of S .

Theorem 3.2.7. *For the preceding sub-Gaussian sketch, $0 < \delta < 1/2$, and $r \geq C_K(d+1)/\delta^2$, the solution of $(\star)$ satisfies*

$$\|A\tilde{x} - b\|_2 \leq (1 + C\delta)\|Ax^* - b\|_2$$

with probability at least $1 - 2e^{-c_K(d+1)}$.

Remark 3.2.8. – A dense least-squares solve on the sketched problem costs $\mathcal{O}(rd^2) = \mathcal{O}(d^3/\delta^2)$ when $r \asymp d/\delta^2$.

– A dense solve of the original problem costs $\mathcal{O}(nd^2)$, assuming $n \geq d$.

There is a problem: dense multiplication to compute SA costs $\mathcal{O}(rnd)$, and computing Sb costs $\mathcal{O}(rn)$. Thus the sketch is not automatically faster overall. There are different approaches:

- Take S sparse.
- Take S with a fast structured transform, such as a randomized Fourier or Hadamard transform, so that SA is easier to compute.

• Sparse Randomized decomposition

We first define the one-nonzero-per-column sketch. Independently for each $i \in [n]$, choose $h(i)$ uniformly in $[r]$ and a Rademacher sign σ_i , independent of all hashes, and set

$$S_{ki} = \sigma_i \mathbb{1}_{\{h(i)=k\}}.$$

Let $U \in \mathbb{R}^{n \times d}$ have orthonormal columns. For least squares, replace d by $q \leq d+1$ and use a basis for $\text{col}[A, b]$.

Theorem 3.2.9. *If $r \geq 2d^2/(\varepsilon\delta^2)$, then $\|(SU)^T(SU) - I_d\| \leq \delta$ with probability at least $1 - \varepsilon$. Forming SA and Sb costs $\mathcal{O}(\text{nnz}(A) + n)$, and the subsequent dense solve costs $\mathcal{O}(rd^2)$, where $\text{nnz}(A)$ is the number of nonzero entries of A .*

Remark 3.2.10 (3.2.9). – Forming the sketch takes input-sparsity time, not $\mathcal{O}(n)$ for an arbitrary dense $n \times d$ matrix.

- This elementary argument requires order d^2 rows at fixed accuracy and failure probability, rather than order d for a dense sub-Gaussian sketch.

Let's set up the proof. Set $X = (SU)^T(SU) - I_d$. This matrix is symmetric, so

$$\|X\|^2 \leq \|X\|_F^2 = \text{Tr}(X^2).$$

Proof of Theorem 3.2.9. By Markov's inequality,

$$\mathbb{P}(\|X\| \geq \delta) \leq \delta^{-2} \mathbb{E}\|X\|_F^2.$$

To compute the entries, write U_a for column a of U , and put $X_i(k) = \mathbb{1}_{\{h(i)=k\}}$. Then

$$\begin{aligned} ((SU)^T(SU))_{ab} &= \sum_k \sum_{i,j} \sigma_i \sigma_j X_i(k) X_j(k) U_a(i) U_b(j) \\ &= \delta_{ab} + \sum_k \sum_{i \neq j} \sigma_i \sigma_j X_i(k) X_j(k) U_a(i) U_b(j), \end{aligned}$$

since $\sum_k X_i(k) = 1$ and $\langle U_a, U_b \rangle = \delta_{ab}$. The following claim gives $\mathbb{E}\|X\|_F^2 \leq 2d^2/r$, hence the required probability bound. $\square$

Claim 3.2.11. *For all a, b , $\mathbb{E}X_{ab}^2 \leq 2/r$. Consequently $\mathbb{E} \operatorname{Tr}(X^2) \leq 2d^2/r$. Higher moments $\mathbb{E} \operatorname{Tr}(X^{2l})$ can yield sharper estimates.*

Proof of Claim 3.2.11. We can write

$$X_{ab} = \sum_{i \neq j} \sigma_i \sigma_j \mathbb{1}_{\{h(i)=h(j)\}} U_a(i) U_b(j).$$

After squaring, the expectation over the signs is zero unless the two ordered pairs are $(i, j), (i, j)$ or $(i, j), (j, i)$. In either case the collision probability is $1/r$.

– Case I: the contribution is

$$\frac{1}{r} \sum_{i \neq j} U_a(i)^2 U_b(j)^2 = \frac{1}{r} \left(1 - \sum_i U_a(i)^2 U_b(i)^2 \right).$$

– Case II: the contribution is

$$\frac{1}{r} \sum_{i \neq j} U_a(i) U_b(i) U_a(j) U_b(j) = \frac{1}{r} \left(\delta_{ab} - \sum_i U_a(i)^2 U_b(i)^2 \right).$$

Adding these contributions gives

$$\mathbb{E}X_{ab}^2 = \frac{1}{r} \left(1 + \delta_{ab} - 2 \sum_i U_a(i)^2 U_b(i)^2 \right) \leq \frac{2}{r}.$$

Summing over a, b proves the claim. $\square$

• Matrix Inequality

More generally, a sparse embedding can have t nonzeros per column, each equal to an independent sign divided by $\sqrt{t}$. One useful construction divides the r rows into t equal blocks and, independently in each block and column, chooses one uniformly random row and a sign. Thus $S_a(i) = \sigma_{ai} X_a(i)/\sqrt{t}$, where X_a has exactly one entry equal to one in each block and all others zero; in particular $\|X_a/\sqrt{t}\| = 1$.

Theorem 3.2.12 (Sparse subspace embedding; Cohen). *For fixed $U \in \mathbb{R}^{n \times d}$ with $U^T U = I_d$, and $0 < \varepsilon, \delta < 1/2$, there are choices*

$$r = \mathcal{O} \left(\frac{d \log(d/\delta)}{\varepsilon^2} \right), \quad t = \mathcal{O} \left(\frac{\log(d/\delta)}{\varepsilon} \right)$$

for the preceding sparse construction such that $\|(SU)^T(SU) - I_d\| \leq \varepsilon$ with probability at least $1 - \delta$. The constants and integer roundings are chosen sufficiently large, with t dividing r .

This sharper result uses matrix trace inequalities; see M. Cohen, *Nearly Tight Oblivious Subspace Embeddings by Trace Inequalities* (SODA 2016). The sparsity parameter here is t throughout; a further tradeoff parameter would require a separately specified construction and theorem. We will continue this topic in the next class.

3.3 Lecture 7: Matrix Concentration Inequality

Reminder:

- Suppose $U \in \mathbb{R}^{n \times (d+1)}$ is an orthonormal basis for the span of the columns of $[A \ b]$.
- For the sketching algorithm we want to construct a matrix $S \in \mathbb{R}^{r \times n}$ such that with probability at least $1 - O(1)$ (over the draw of S) it holds:

$$\|(SU)^T(SU) - I\| \leq \delta$$

Or equivalently,

$$\| \underbrace{U^T(S^T S)U}_X - I \| \leq \delta$$

Last time we bounded this norm using the inequality:

$$\|X\|^{2l} \leq \text{Tr } X^{2l}$$

(we proved for $l = 1$ but we can extend it for higher-order moments as we did in the previous lectures).

In this lecture, we will prove another way to control this norm by bounding the exponential function of the eigenvalues using, for $\theta > 0$,

$$e^{\theta \lambda_{\max}(X)} \leq \text{Tr } e^{\theta X}$$

First, observe that $X \in \mathbb{R}^{(d+1) \times (d+1)}$ is a symmetric matrix. We know that every symmetric matrix $A \in \mathbb{R}^{d \times d}$, is diagonalizable, with real eigenvalues and a full set of orthogonal eigenvectors. The eigenvalue decomposition of A can be written as:

$$A = U \Lambda U^*,$$

where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ is its diagonal matrix of eigenvalues and U the orthogonal matrix of its eigenvectors.

If $f : \mathbb{R} \rightarrow \mathbb{R}$ is a polynomial function (or more generally has a power series convergent on a disk containing the spectrum) then we can define (by overloading the notation) the function $f : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d \times d}$ by substituting x with a matrix $A \in \mathbb{R}^{d \times d}$. Then, in our case,

$$f(A) = U f(\Lambda) U^*$$

In particular, for $f(x) = e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k$ we define the matrix exponential $f(X) = e^X = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$ and in our case this specializes to:

$$f(\theta A) = U \text{diag}(e^{\theta \lambda_1}, \dots, e^{\theta \lambda_d}) U^*$$

Also, we define a partial ordering in the set of symmetric matrices. In particular, we have: $A \preceq B$ iff $B - A$ is symmetric positive semi-definite matrix. Then it holds that

- if $A \preceq B$ a.s. then $\mathbb{E}[A] \preceq \mathbb{E}[B]$.
- if $f \leq g$ on $\mathbb{R}$ then $f(A) \preceq g(A)$.

Theorem 3.3.1 (Matrix Chernoff Bound (Tropp)). *Let $A_1, \dots, A_n$ be independent random positive semidefinite $d \times d$ matrices with $\mathbb{E} \sum_i A_i = I_d$ and $\|A_i\| \leq a$ almost surely, where $a > 0$. For $0 < \epsilon < 1$, with probability at least $1 - 2d \exp(-\epsilon^2/(3a))$,*

$$\left\| \sum_{i=1}^n A_i - I_d \right\| \leq \epsilon.$$

Remark: What makes the matrix bounds more difficult than the scalar case is the non-commutativity of matrix multiplication, which in turn leads to $e^{X+Y} \neq e^X e^Y$.

Application 1: Before the proof, we give an application. Consider independent $X_i \sim \mathcal{N}(0, I_p)$ with sample covariance $n^{-1} \sum_i X_i X_i^T$. For $A_i = n^{-1} X_i X_i^T$, we have $\|A_i\| = \|X_i\|^2/n$ and $\mathbb{E} \sum_i A_i = I_p$. However, Gaussian vectors are unbounded, so the almost-sure hypothesis $\|A_i\| \leq a$ cannot be imposed directly.

Here is the required truncation. Fix $0 < q < 1$ and take $L = C(p + \log(n/q))$ with C sufficiently large. Gaussian norm tails imply

$$\mathbb{P}(\max_i \|X_i\|^2 > L) \leq q/2, \quad a_L := p^{-1} \mathbb{E}[\|X_1\|^2 \mathbb{1}_{\{\|X_1\|^2 \leq L\}}] \geq 1 - q/(2n).$$

The matrices $B_i = (na_L)^{-1} X_i X_i^T \mathbb{1}_{\{\|X_i\|^2 \leq L\}}$ remain independent, satisfy $\mathbb{E} \sum_i B_i = I_p$ by rotational invariance, and obey $\|B_i\| \leq L/(na_L)$. Applying Chernoff and then restoring the discarded event shows, with probability at least $1 - q$, that

$$\left\| \frac{1}{n} \sum_i X_i X_i^T - I_p \right\| \leq C \sqrt{\frac{(p + \log(n/q)) \log(4p/q)}{n}} + \frac{q}{2n},$$

provided the displayed square-root term is less than a sufficiently small constant (otherwise this small-deviation application gives no useful bound). **Remark:** When p dominates $\log(n/q)$, this general matrix bound loses a square-root logarithmic factor relative to the sharp Gaussian covariance scale $\sqrt{p/n}$; it is not a sharp Marchenko–Pastur edge estimate.

Application 2: Another application is in our sketching algorithm. By denoting the rows of U by u_i i.e.:

$$U = \begin{bmatrix} -u_1 - \\ -u_2 - \\ \vdots \\ -u_n - \end{bmatrix}$$

we can write:

$$U^T (S^T S) U = \sum_{ij} \underbrace{(S^T S)_{ij}}_{A_{ij}} u_i^T u_j$$

which can be understood as a sum of n^2 random matrices. They are not independent, and the off-diagonal summands are generally neither symmetric nor positive semidefinite. Thus this expansion is not directly an application of Matrix Chernoff. Sparse-embedding proofs require additional decoupling and matrix moment or trace estimates, together with control of the diagonal terms.

To prove the theorem we will use the Lemma:

Lemma 3.3.2. (*Lieb Theorem*) If $H \in \mathbb{R}^{d \times d}$ is a symmetric matrix, then the map

$$\mathbb{R}^{d \times d} \ni A \mapsto \text{Tr } e^{H + \log A}$$

is a concave function on the space of positive definite matrices.

What we need from Lieb Theorem is the fact that since the map $f(A)$ is concave, it holds that: $\mathbb{E}[f(A)] \leq f(\mathbb{E}[A])$. Now, for a random symmetric matrix X , the matrix e^X is positive definite; assume the required exponential moments are finite. What we need is to bound the trace of this exponential matrix. We have:

$$\mathbb{E}_X[\text{Tr } e^{H+X}] = \mathbb{E}_X[\text{Tr } e^{H + \log e^X}] \leq \text{Tr } \exp(H + \log \mathbb{E}[e^X])$$

Proof. We will prove that: $\mathbb{P}[\lambda_{\max}(\sum_{i=1}^n A_i) \geq 1 + \epsilon] \leq d \exp(-\epsilon^2/(3a))$. We also give the lower-tail bound at the end of the proof.

$$\mathbb{P}[\lambda_{\max}(\sum_{i=1}^n A_i) \geq t] \leq \mathbb{P}[\text{Tr } e^{\theta \sum_{i=1}^n A_i} \geq e^{\theta t}] \leq e^{-\theta t} \mathbb{E}[\text{Tr } e^{\theta \sum_{i=1}^n A_i}]$$

For the upper tail take $\theta > 0$. Conditioning on $A_1, \dots, A_{n-1}$, independence and the preceding concavity bound, with $H = \theta \sum_{i=1}^{n-1} A_i$, $X = \theta A_n$ give:

$$\mathbb{E}[\text{Tr } e^{\theta \sum_{i=1}^n A_i}] \leq \mathbb{E}[\text{Tr } e^{H + \log \mathbb{E}[e^{\theta A_n}]}]$$

and applying the inequality recursively to the remaining terms in the sum $\sum_{i=1}^{n-1} A_i$ we get the upper bound:

$$\leq \mathbb{E} \left[\text{Tr } \exp \sum_{i=1}^n \log \mathbb{E}[e^{\theta A_i}] \right] = \text{Tr } \exp \sum_{i=1}^n \log \mathbb{E}[e^{\theta A_i}]$$

Since, each A_i is symmetric and positive semidefinite with $\|A_i\| \leq a$ we have: $0 \preceq A_i \preceq aI$. Moreover, the function $f(x) = e^{\theta x}$ is convex which implies that for all $x \in [0, a]$ we have:

$$f(x) \leq f(0) + \frac{f(a) - f(0)}{a} x$$

From the properties of the partial order above we get:

$$e^{\theta A_i} \preceq I + \frac{e^{\theta a} - 1}{a} A_i$$

which implies:

$$\mathbb{E}[e^{\theta A_i}] \preceq I + \frac{e^{\theta a} - 1}{a} \mathbb{E}[A_i] \preceq \exp \left(\frac{e^{\theta a} - 1}{a} \mathbb{E}[A_i] \right)$$

where in the last inequality we used $1 + x \leq e^x$. To use this inequality, note that the matrix logarithm is operator monotone on positive definite matrices, and $B \mapsto \text{Tr } e^B$ is monotone in

the positive-semidefinite order (by eigenvalue monotonicity). The matrix exponential itself is not operator monotone. Therefore:

$$\begin{aligned} \text{Tr} \exp \sum_{i=1}^n \log \mathbb{E}[e^{\theta A_i}] &\leq \text{Tr} \exp \sum_{i=1}^n \log \exp \left(\frac{e^{\theta a} - 1}{a} \mathbb{E}[A_i] \right) = \text{Tr} \exp \sum_{i=1}^n \left(\frac{e^{\theta a} - 1}{a} \mathbb{E}[A_i] \right) \\ &= \text{Tr} \exp \frac{e^{\theta a} - 1}{a} \left(\sum_{i=1}^n \mathbb{E}[A_i] \right) = \text{Tr} \exp \frac{e^{\theta a} - 1}{a} I_d = d \exp \frac{e^{\theta a} - 1}{a} \end{aligned}$$

By putting this bound on the expectation back to the total bound we get:

$$\mathbb{P}[\lambda_{\max}(\sum_{i=1}^n A_i) \geq t] \leq e^{-\theta t} (d \exp \frac{e^{\theta a} - 1}{a}) = d \exp \left(\frac{e^{\theta a} - 1}{a} - \theta t \right)$$

The upper-tail inequality above holds for every $\theta > 0$. To make it as tight as possible we get that θ that minimizes the bound. If we take the derivative equal to zero this happens for θ such that $e^{\theta a} = t$ from which we get:

$$d \exp \left(-\frac{t \ln t - t + 1}{a} \right)$$

For $t = 1 + \epsilon$ and $0 < \epsilon < 1$, $(1 + \epsilon) \log(1 + \epsilon) - \epsilon \geq \epsilon^2/3$, giving the upper-tail bound $de^{-\epsilon^2/(3a)}$. For the lower tail take $\theta < 0$. If $\lambda_{\min}(\sum_i A_i) \leq 1 - \epsilon$, then $\text{Tr} e^{\theta \sum_i A_i} \geq e^{\theta(1-\epsilon)}$. The same convex chord bound on $[0, a]$ and the same trace-mgf argument apply for negative θ . Taking $\theta = a^{-1} \log(1 - \epsilon)$ gives

$$\mathbb{P} \left(\lambda_{\min} \left(\sum_i A_i \right) \leq 1 - \epsilon \right) \leq d \exp \left[-\frac{\epsilon + (1 - \epsilon) \log(1 - \epsilon)}{a} \right] \leq de^{-\epsilon^2/(2a)}.$$

The union bound proves the stated two-sided estimate. □

Chapter 4

Spiked Matrix Models

4.1 Lecture 7: Spiked Matrix Model I

Definition 4.1.1. The spiked matrix model is defined as follows:

$$Y = \beta \mu \mu^T + X \quad (4.1.1)$$

where μ is a deterministic unit vector, independent of the noise, and $\beta \geq 0$ is sometimes called the signal-to-noise ratio. In the theorem below, take X to be the normalized Gaussian orthogonal ensemble: its independent upper-triangular entries have $X_{ii} \sim N(0, 2/n)$ and $X_{ij} \sim N(0, 1/n)$ for $i < j$.

Theorem 4.1.2 (BBP phase transition). *Denote the largest eigenvalue of Y by λ , and take β fixed as $n \rightarrow \infty$. There is a phase transition:*

- If $\beta < 1$, then $\lambda \rightarrow 2$ and $n^{2/3}(\lambda - 2)$ converges in distribution to the GOE Tracy–Widom law (no separated outlier).
- If $\beta > 1$, then $\lambda \rightarrow \beta + 1/\beta$, with Gaussian fluctuations on scale $n^{-1/2}$.

Denote $\tilde{\mu}$ the normalized eigenvector associated with λ .

- If $\beta < 1$, then $|\langle \mu, \tilde{\mu} \rangle|^2 \rightarrow 0$
- If $\beta > 1$, then $|\langle \mu, \tilde{\mu} \rangle|^2 \rightarrow 1 - 1/\beta^2$

The case $\beta = 1$ has a critical deformed edge law and is not covered by the two fluctuation statements above.

The argument below proves the limiting locations and eigenvector overlaps; the Tracy–Widom and Gaussian fluctuation statements are quoted as part of the BBP theorem. The location and overlap limits extend to suitable non-Gaussian Wigner matrices, but Gaussian outlier fluctuations need not hold for an arbitrary localized spike and non-Gaussian entries.

Proof. Because X is a Wigner matrix, without the signal, eigenvalues of X are in $[-2 - \epsilon, 2 + \epsilon]$ with high probability $1 - \frac{1}{n^D}$.

The eigenvalues of Y are roots of the characteristic polynomial

$$\det(z - Y) = \det(z - \beta\mu\mu^T - X) = 0 \quad (*)$$

Detecting outlier is equivalent to ask, are there $z \notin [-2 - \epsilon, 2 + \epsilon]$ such that $(*)$ holds.

For $z \notin [-2 - \epsilon, 2 + \epsilon]$, $z - X$ is invertible,

$$\begin{aligned} \det(z - \beta\mu\mu^T - X) &= \det((z - X) - \beta(z - X)(z - X)^{-1}\mu\mu^T) \\ &= \det(z - X) \det(I - \beta(z - X)^{-1}\mu\mu^T) \\ &= 0 \\ \iff \det(I - \beta(z - X)^{-1}\mu\mu^T) &= \det(I + \beta(X - z)^{-1}\mu\mu^T) = 0 \end{aligned} \quad (4.1.2) \quad (**)$$

Fact: $\det(I_n + AB^T) = \det(I_r + B^T A)$ where $A, B \in \mathbb{R}^{n \times r}$

Proof of the fact:

- If $n = r$ and $\det A \neq 0$.

$$\det(I_n + AB^T) \det(A) = \det(A + AB^T A) = \det(A) \det(I_n + B^T A)$$

If $\det A = 0$, take a sequence $A_k \rightarrow A$ entrywise with $\det A_k \neq 0$

$$\det(I_n + AB^T) = \lim_{k \rightarrow +\infty} \det(I_n + A_k B^T) = \lim_{k \rightarrow +\infty} \det(I_n + B^T A_k) = \det(I_n + B^T A)$$

- If $r < n$.

$$\tilde{A}_{n \times n} = (A \quad \vec{0}), \quad \tilde{B}_{n \times n} = (B \mid \vec{0})$$

$$AB^T = \tilde{A}\tilde{B}^T, \quad \tilde{B}^T \tilde{A} = \begin{bmatrix} B^T A & 0 \\ 0 & 0 \end{bmatrix}$$

$$\det(I_n + AB^T) = \det(I_n + \tilde{A}\tilde{B}^T) = \det(I_n + \tilde{B}^T \tilde{A}) = \det(I_r + B^T A)$$

The case $r > n$ follows by interchanging the two sides. We can rewrite $(**)$ as

$$\det(I + \beta(X - z)^{-1}\mu\mu^T) = 1 + \beta\mu^T(X - z)^{-1}\mu = 0$$

For $\beta > 0$, locating a separated outlier therefore becomes the question:

$$\mu^T(X - z)^{-1}\mu = -\frac{1}{\beta}. \quad (***)$$

$(X - z)^{-1} = G(z)$ is the Green's function of the matrix.

$$\frac{1}{n} \text{Tr} G(z) \rightarrow m_{sc}(z) = \int \frac{\rho_{sc}(x)}{x - z} dx = \frac{-z + \sqrt{z^2 - 4}}{2}$$

in probability, locally uniformly away from $[-2, 2]$.

Theorem 4.1.3. *For any fixed deterministic unit vector μ , and fixed $\epsilon > 0$,*

$$\mu^T (X - zI)^{-1} \mu = m_{sc}(z) + o_{\mathbb{P}}(1)$$

locally uniformly for $z \notin [-2 - \epsilon, 2 + \epsilon]$.

Proof in the Gaussian case. The GOE is invariant under orthogonal conjugation: $X \stackrel{d}{=} V X V^T$ for every deterministic orthogonal V . We must also establish concentration; invariance alone is not enough. Conditional on the eigenvalues, the eigenvectors form a Haar orthogonal basis independent of those eigenvalues. Thus the coordinates of μ in this basis have the law of a uniform $q \in S^{n-1}$, and

$$\mu^T G(z) \mu \stackrel{d}{=} \sum_{j=1}^n \frac{q_j^2}{\lambda_j - z}.$$

Using $\mathbb{E} q_j^2 = 1/n$ and $\mathbb{E}(q_i^2 q_j^2) = (1 + 2\mathbb{1}_{\{i=j\}})/(n(n+2))$, the conditional mean equals $n^{-1} \text{Tr } G(z)$ and, when $\text{dist}(z, \{\lambda_j\}) \geq a > 0$, its conditional variance is at most $2/((n+2)a^2)$. The semicircle law and $\|X\| \rightarrow 2$ then give the claimed pointwise convergence in probability. A finite grid and the bound $\|G'(z)\| \leq a^{-2}$ give local uniform convergence.

Equivalently, the diagonal case $\mu = e_i$ extends to any deterministic μ by taking $V e_i = \mu$:

$$\begin{aligned} \langle e_i, (X - zI)^{-1} e_i \rangle &\stackrel{d}{=} \langle e_i, (V^T X V - zI)^{-1} e_i \rangle \\ &= \langle e_i, V^T (X - zI)^{-1} V e_i \rangle = \langle \mu, (X - zI)^{-1} \mu \rangle. \end{aligned}$$

□

□

4.2 Lecture 8: Spiked Matrix Model II

(proof continued). Write $u = \mu$ and $\tilde{u} = \tilde{\mu}$. On $z \in (2, \infty)$, the range of $m_{sc}(z) = (-z + \sqrt{z^2 - 4})/2$ is $(-1, 0)$. Thus the equation

$$u^*(X - zI)^{-1}u = -\frac{1}{\beta} + o(1)$$

has a solution that stays a fixed positive distance above 2 with probability tending to one if and only if $\beta > 1$. Indeed, $u^T(X - zI)^{-1}u$ is increasing on $z > \lambda_1(X)$; local uniform convergence and a sign change around the limiting root give existence and uniqueness of the separated root. For $\beta < 1$ there is no such root; interlacing and $\lambda_1(X) \rightarrow 2$ then imply $\lambda_1(Y) \rightarrow 2$. A finite- n root just above $\lambda_1(X)$ can still exist below threshold, so it is the separation from the bulk, not merely existence of a root, that distinguishes the two regimes. Also, solving

$$\frac{-\lambda + \sqrt{\lambda^2 - 4}}{2} = -\frac{1}{\beta} + o(1)$$

in λ we have $\lambda = \beta + 1/\beta + o(1)$, as desired.

For the eigenvector overlap, first consider $\beta > 1$. For this case, with $\lambda = \lambda_1(Y)$ (top eigenvalue of Y) by definition we have

$$\begin{aligned} \lambda \tilde{u} &= (X + \beta u u^*) \tilde{u} \\ \Rightarrow \tilde{u} &= \beta(\lambda I - X)^{-1} u u^* \tilde{u} = \frac{(\lambda I - X)^{-1} u}{\|(\lambda I - X)^{-1} u\|} \end{aligned}$$

after choosing the sign of $\tilde{u}$, since it is proportional to $(\lambda I - X)^{-1}u$ and has unit norm. Thus

$$\langle u, \tilde{u} \rangle = \frac{u^*(\lambda I - X)^{-1}u}{\|(\lambda I - X)^{-1}u\|}$$

where $u^*(\lambda I - X)^{-1}u = -m_{sc}(\lambda) + o(1) = 1/\beta + o(1)$ from the results above and $\|(\lambda I - X)^{-1}u\|^2 = u^*(\lambda I - X)^{-2}u$. Note that $\frac{\partial}{\partial z}(A - zI)^{-1} = (A - zI)^{-2}$ since

$$\begin{aligned} \frac{\partial}{\partial z}(A - zI)^{-1} &= \lim_{t \rightarrow 0} \frac{(A - (z+t)I)^{-1} - (A - zI)^{-1}}{t} \\ &= \lim_{t \rightarrow 0} \frac{(A - (z+t)I)^{-1} t I (A - zI)^{-1}}{t} = (A - zI)^{-2} \end{aligned}$$

Local uniform analytic convergence of the resolvent away from the spectrum also gives convergence of its derivative, including at the random outlier $\lambda \rightarrow \beta + 1/\beta$. Thus

$$u^*(\lambda I - X)^{-2}u = m'_{sc}(\lambda) + o(1) = -\frac{1}{2} + \frac{\lambda}{2\sqrt{\lambda^2 - 4}} + o(1) = \frac{1}{\beta^2 - 1} + o(1)$$

implying that

$$|\langle u, \tilde{u} \rangle|^2 = \frac{1/\beta^2}{1/(\beta^2 - 1)} + o(1) = 1 - \frac{1}{\beta^2} + o(1)$$

in probability. For $0 \leq \beta < 1$, choose $\beta < \beta' < 1$ and write $L_n(b) = \lambda_1(X + buu^T)$. This is a convex function of b . Since the top eigenvalue is simple almost surely at each fixed b in the GOE setting, the Hellmann–Feynman formula gives $L'_n(\beta) = |\langle u, \tilde{u} \rangle|^2$. Convexity yields

$$0 \leq |\langle u, \tilde{u} \rangle|^2 \leq \frac{L_n(\beta') - L_n(\beta)}{\beta' - \beta} \rightarrow 0,$$

because both eigenvalues tend to 2. This completes the location and overlap claims. $\square$

The BBP phase transition tells us that with spectral algorithm, we can (asymptotically) know whether there exists a spike if the signal-to-noise ratio β is larger than 1. Then, a natural question is to ask whether we can do better (in a lower signal-to-noise ratio regime) with some other algorithms. We will show under certain priors on the signal, consistent detection below the spectral threshold is impossible. This does not rule out tests with a nonvanishing advantage over random guessing.

Also, now we formulate the problem as a detection problem, meaning that we would like to detect whether there is a signal (alternative hypothesis) or not (null hypothesis).

Definition 4.2.1. Let $\mathbb{P}_n, \mathbb{Q}_n$ be probability measures on measurable spaces $(\Omega_n, \mathcal{F}_n)$ for each $n \in \mathbb{N}$. Then $\mathbb{Q}_n$ is contiguous with respect to $\mathbb{P}_n$ (denoted as $\mathbb{Q}_n \triangleleft \mathbb{P}_n$) if

$$\lim_{n \rightarrow \infty} \mathbb{P}_n(A_n) = 0 \Rightarrow \lim_{n \rightarrow \infty} \mathbb{Q}_n(A_n) = 0$$

for every sequence of measurable events $A_n \in \mathcal{F}_n$.

The concept of contiguity is important for hypothesis testing, as shown by the following theorem.

Theorem 4.2.2. *If $\mathbb{Q}_n \triangleleft \mathbb{P}_n$, then there exists no test such that*

$$\mathbb{P}(\text{Type I error}) + \mathbb{P}(\text{Type II error}) \rightarrow 0$$

as $n \rightarrow \infty$.

Proof. Let A_n be the rejection region of the test. Assuming the contrary, we have

$$\mathbb{P}_n(A_n) + \mathbb{Q}_n(A_n^c) \rightarrow 0$$

implying $\mathbb{P}_n(A_n) \rightarrow 0$. Then by contiguity, we have $\mathbb{Q}_n(A_n) \rightarrow 0$ and thus $\mathbb{Q}_n(A_n^c) \rightarrow 1$, a contradiction. $\square$

In general, it is difficult to show contiguity directly. Instead, we can use the following second moment method.

Lemma 4.2.3 (Second moment method). *Let $\mathbb{P}_n, \mathbb{Q}_n$ be probability measures with $\mathbb{Q}_n \ll \mathbb{P}_n$. If*

$$\mathbb{E}_{\mathbb{P}_n} \left(\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \right)^2 < C \text{ for all } n \in \mathbb{N}$$

for some finite constant C independent of n then $\mathbb{Q}_n \triangleleft \mathbb{P}_n$.

Proof. Let A_n be any sequence such that $\mathbb{P}_n(A_n) \rightarrow 0$. Then

$$\begin{aligned} \mathbb{Q}_n(A_n) &= \int \mathbf{1}_{A_n} d\mathbb{Q}_n = \int \mathbf{1}_{A_n} \frac{d\mathbb{Q}_n}{d\mathbb{P}_n} d\mathbb{P}_n \\ &\leq \mathbb{P}_n(A_n)^{1/2} (\mathbb{E}_{\mathbb{P}_n} (\frac{d\mathbb{Q}_n}{d\mathbb{P}_n})^2)^{1/2} \\ &\leq C^{1/2} \mathbb{P}_n(A_n)^{1/2} \rightarrow 0 \end{aligned}$$

□

The argument is unchanged by conditioning on an event whose probability tends to one:

Lemma 4.2.4. *Let W_n be such that $\mathbb{Q}_n(W_n) = 1 - o(1)$ and define*

$$\tilde{\mathbb{Q}}_n(\cdot) = \frac{\mathbb{Q}_n(\cdot \cap W_n)}{\mathbb{Q}_n(W_n)}$$

Then $\tilde{\mathbb{Q}}_n \triangleleft \mathbb{Q}_n$ and $\mathbb{Q}_n \triangleleft \tilde{\mathbb{Q}}_n$. If

$$\mathbb{E}_{\mathbb{P}_n} (\frac{d\tilde{\mathbb{Q}}_n}{d\mathbb{P}_n})^2 < C \text{ for all } n \in \mathbb{N}$$

for some finite constant C independent of n then $\tilde{\mathbb{Q}}_n \triangleleft \mathbb{P}_n$, which implies $\mathbb{Q}_n \triangleleft \mathbb{P}_n$.

Given such definitions, we now consider

$$\mathbb{P}_n : Y = \frac{1}{\sqrt{n}} W, W \sim \text{GOE}(n)$$

$$\mathbb{Q}_n : Y = \lambda x x^T + \frac{1}{\sqrt{n}} W, x \sim \mathcal{X}_n, W \sim \text{GOE}(n) \text{ are independent}$$

where $\|x\|_2 \rightarrow 1$ in probability. Here W is the unnormalized GOE, with off-diagonal variance one and diagonal variance two, and $\lambda \geq 0$ is fixed. In other words, we consider the signal detection as a hypothesis testing. Here are some examples for the prior.

Example 4.2.5. $\mathcal{X}_n = N(0, \frac{1}{n} I_n)$

Example 4.2.6. $\mathcal{X}_n = \{(x_1, \dots, x_n) : x_i \stackrel{\text{iid}}{\sim} \text{Uniform}(\pm n^{-1/2})\}$

Example 4.2.7. $\mathcal{X}_n = \text{Uniform}(\mathbb{S}^{n-1})$

Using second moment method, we show that with such priors we cannot do better than the spectral algorithm. The following result shows that for $\sigma = 1$ (which hold for the three priors above), if $\lambda < 1$ then no test can distinguish between the two hypotheses with vanishing error.

Theorem 4.2.8. *Suppose $\|x\|_2 \rightarrow 1$ in probability and $\mathcal{X}_n$ is centered and σ^2/n -sub-Gaussian in the moment-generating-function sense:*

$$\mathbb{E} e^{t\langle v, x \rangle} \leq \exp \left(\frac{\sigma^2 t^2 \|v\|^2}{2n} \right), \quad t \in \mathbb{R}, \quad v \in \mathbb{R}^n.$$

If $0 \leq \lambda < 1/\sigma$, then $\mathbb{Q}_n \triangleleft \mathbb{P}_n$.

Proof. We first calculate the likelihood-ratio second moment, allowing its value to be infinite, and then apply the identity to a truncated prior. Consider $\mathbb{E}_{\mathbb{P}_n}(\frac{d\mathbb{Q}_n}{d\mathbb{P}_n})^2$. From

$$\begin{aligned} d\mathbb{P}_n &= \frac{1}{Z} \exp(-\sum_{i,j} \frac{n}{4} y_{ij}^2) dY = \frac{1}{Z} \exp(-\frac{n}{4} \text{tr}(YY^T)) dY \\ d\mathbb{Q}_n &= \mathbb{E}_{x \sim \mathcal{X}_n} \frac{1}{Z'} \exp(-\frac{n}{4} \text{tr}((Y - \lambda xx^T)(Y - \lambda xx^T)^T)) dY \end{aligned}$$

(dY is Lebesgue measure on the independent upper-triangular entries, and $Z' = Z$ because translation does not change the normalizing constant) we have

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} = \mathbb{E}_{x \sim \mathcal{X}_n} \exp(\frac{n\lambda}{2} \text{tr}(Yxx^T) - \frac{n\lambda^2}{4} \text{tr}(xx^Txx^T))$$

and thus

$$\mathbb{E}_{\mathbb{P}_n}(\frac{d\mathbb{Q}_n}{d\mathbb{P}_n})^2 = \mathbb{E}_{Y \sim \mathbb{P}_n} \mathbb{E}_{x, x' \sim \mathcal{X}_n} \exp(\frac{n\lambda}{2} \text{tr}(Y(xx^T + x'x'^T)) - \frac{n\lambda^2}{4} \text{tr}(xx^Txx^T + x'x'^Tx'x'^T))$$

where

$$\begin{aligned} &\mathbb{E}_{Y \sim \mathbb{P}_n} \exp(\frac{n\lambda}{2} \text{tr}(Y(xx^T + x'x'^T))) \\ &= \int \exp(\frac{n\lambda}{2} \text{tr}(Y(xx^T + x'x'^T))) \frac{1}{Z} \exp(-\frac{n}{4} \text{tr}(YY^T)) dY \\ &= \int \frac{1}{Z} \exp(-\frac{n}{4} \text{tr}((Y - \lambda(xx^T + x'x'^T))(Y - \lambda(xx^T + x'x'^T))^T)) dY \exp(\frac{n\lambda^2}{4} \text{tr}((xx^T + x'x'^T)^2)) \\ &= \exp(\frac{n\lambda^2}{4} \text{tr}((xx^T + x'x'^T)^2)) \end{aligned}$$

and thus

$$\mathbb{E}_{\mathbb{P}_n}(\frac{d\mathbb{Q}_n}{d\mathbb{P}_n})^2 = \mathbb{E}_{x, x' \sim \mathcal{X}_n} \exp(\frac{n\lambda^2}{4} \text{tr}(2xx^Tx'x'^T)) = \mathbb{E}_{x, x' \sim \mathcal{X}_n} \exp(\frac{n\lambda^2}{2} \langle x, x' \rangle^2)$$

According to the second moment method, we should bound this value.

Claim 4.2.9. *If $\langle x, x' \rangle$ is $\frac{\sigma^2}{n}$ -sub-Gaussian then for $\lambda < 1/\sigma$,*

$$\mathbb{E} e^{\frac{n\lambda^2}{2} \langle x, x' \rangle^2}$$

is bounded.

Proof. The case $\lambda = 0$ is immediate. For $\lambda > 0$, by the sub-Gaussian condition,

$$\mathbb{P}(|\langle x, x' \rangle| \geq t) \leq 2e^{-\frac{nt^2}{2\sigma^2}}$$

and thus

$$\begin{aligned}
\mathbb{E} e^{\frac{n\lambda^2}{2} \langle x, x' \rangle^2} &= \int_0^\infty \mathbb{P}(e^{\frac{n\lambda^2}{2} \langle x, x' \rangle^2} \geq u) du \\
&= 1 + \int_1^\infty \mathbb{P}(|\langle x, x' \rangle| \geq \sqrt{\frac{2 \log u}{n\lambda^2}}) du \\
&\leq 1 + 2 \int_1^\infty e^{-\frac{\log u}{\sigma^2 \lambda^2}} du \\
&= 1 + 2 \int_1^\infty u^{-\frac{1}{\sigma^2 \lambda^2}} du < \infty
\end{aligned}$$

if $\frac{1}{\sigma\lambda} > 1 \Leftrightarrow \lambda < \frac{1}{\sigma}$. □

The independent Gaussian prior does not make the unconditioned overlap of two Gaussian vectors globally sub-Gaussian, so truncation is essential for that example. Given $\lambda < 1/\sigma$, fix $\delta > 0$ such that $\lambda < \frac{1}{\sigma(1+\delta)}$ and let $W_n = \{x : \|x\| \leq 1 + \delta\}$. Also, let $\tilde{\mathcal{X}}_n$ be the conditional law of $\mathcal{X}_n$ on W_n . Then for

$$\tilde{\mathbb{Q}}_n : Y = \lambda x x^T + \frac{1}{\sqrt{n}} W$$

where the only difference is $x \sim \tilde{\mathcal{X}}_n$, put $q_n = \mathcal{X}_n(W_n) \rightarrow 1$. The total variation distance between the two data laws is at most $1 - q_n$, by coupling the latent priors and using the same noise. Thus they are mutually contiguous; this conclusion does not require W_n to be observable from Y . Also,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_n} \left(\frac{d\tilde{\mathbb{Q}}_n}{d\mathbb{P}_n} \right)^2 &= \mathbb{E}_{x, x' \sim \tilde{\mathcal{X}}_n} e^{\frac{n\lambda^2}{2} \langle x, x' \rangle^2} \\
&\leq q_n^{-1} \mathbb{E}_{x' \sim \tilde{\mathcal{X}}_n} \mathbb{E}_{x \sim \mathcal{X}_n} e^{\frac{n\lambda^2}{2} \langle x, x' \rangle^2}
\end{aligned}$$

where conditioning on $x' \sim \tilde{\mathcal{X}}_n$, $\langle x, x' \rangle$ is $\frac{\sigma^2(1+\delta)^2}{n}$ -sub-Gaussian. From $\lambda < \frac{1}{\sigma(1+\delta)}$, with the corresponding bound $0 < C < \infty$ on $\mathbb{E}_{x \sim \mathcal{X}_n} e^{\frac{n\lambda^2}{2} \langle x, x' \rangle^2}$ we have

$$\mathbb{E}_{\mathbb{P}_n} \left(\frac{d\tilde{\mathbb{Q}}_n}{d\mathbb{P}_n} \right)^2 \leq q_n^{-1} \mathbb{E}_{x' \sim \tilde{\mathcal{X}}_n} C = C/q_n = O(1)$$

which proves that $\mathbb{E}_{\mathbb{P}_n} (\frac{d\tilde{\mathbb{Q}}_n}{d\mathbb{P}_n})^2$ is bounded. Thus by the second moment method, $\mathbb{Q}_n \triangleleft \mathbb{P}_n$. □

4.3 Lecture 9: Spiked Tensor Model

Consider 3-tensor $(Z_{ijk})_{1 \leq i \leq n_1, 1 \leq j \leq n_2, 1 \leq k \leq n_3} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. Similarly, we can define $[Z_{i_1 \dots i_k}] \in \mathbb{R}^{n_1 \times \dots \times n_k}$. It is interesting to study the “low rank tensor” plus noise model

$$\sum_{i=1}^k \beta_i v_i \otimes v_i \otimes v_i + Z, \quad v_i \in \mathbb{R}^n,$$

where $\{v_i\}_{i=1}^k$ is an orthonormal family and Z is a noise tensor. An illuminating example is the recovery of the multi-dimensional Gaussian mixture model. Suppose $X \in \mathbb{R}^d$ has distribution $\sum_{i=1}^k w_i N(\mu_i, \sigma^2 I_d)$, with $w_i > 0$ and $\sum_i w_i = 1$. Suppose $\mathbb{P}[h = i] = w_i, 1 \leq i \leq k$. Conditionally on h , $X \sim N(\mu_h, \sigma^2 I_d)$; equivalently, $X = \mu_h + Z$, where $Z \sim N(0, \sigma^2 I_d)$ is independent of h . Given the observation $X_1, \dots, X_n$, we have to estimate $\{w_i, \mu_i\}_{i=1}^k, \sigma^2$. Consider the following moments.

1. $\mathbb{E}[X] = \mathbb{E}[\mu_h] = \sum_{i=1}^k w_i \mu_i = \bar{\mu}$;
2. $\mathbb{E}[(X - \mathbb{E}[X])(X^\top - \mathbb{E}[X^\top])] = \sigma^2 I_d + \sum_{i=1}^k w_i (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top$;
3. $\mathbb{E}[XX^\top] - \sigma^2 I = \sum_{i=1}^k w_i \mu_i \mu_i^\top := M_2$;
4. The noise-corrected third moment is

$$\begin{aligned} M_3 &= \mathbb{E}[X^{\otimes 3}] - \sigma^2 \sum_{j=1}^d (\bar{\mu} \otimes e_j \otimes e_j + e_j \otimes \bar{\mu} \otimes e_j + e_j \otimes e_j \otimes \bar{\mu}) \\ &= \sum_{i=1}^k w_i \mu_i^{\otimes 3}. \end{aligned}$$

This follows by expanding $(\mu_h + Z)^{\otimes 3}$: terms with one or three Gaussian factors vanish, and terms with two factors give the displayed correction.

If $k \leq d$, the between-component covariance has rank at most $k-1$, so σ^2 is the smallest population covariance eigenvalue. Now suppose $k = d$ and the means $\mu_1, \dots, \mu_d$ are linearly independent, so M_2 is invertible. Define $v_i := \sqrt{w_i} M_2^{-1/2} \mu_i$. Therefore $v_1, \dots, v_d$ are orthonormal vectors. We write $V = [v_1, \dots, v_d]$ and

$$\begin{aligned} VV^\top &= \sum_{i=1}^k v_i v_i^\top \\ &= \sum_{i=1}^k \sqrt{w_i} M_2^{-1/2} \mu_i \mu_i^\top M_2^{-1/2} \sqrt{w_i} \\ &= I. \end{aligned}$$

Thus

$$\begin{aligned} M_3(M_2^{-1/2}, M_2^{-1/2}, M_2^{-1/2}) &= \sum_{i=1}^k w_i M_2^{-1/2} \mu_i \otimes M_2^{-1/2} \mu_i \otimes M_2^{-1/2} \mu_i \\ &= \sum_{i=1}^k w_i^{-1/2} v_i \otimes v_i \otimes v_i \end{aligned}$$

from which an orthogonal tensor decomposition recovers $v_i, w_i^{-1/2}$ up to permutation (choosing signs so the coefficients are positive). Then $\mu_i = w_i^{-1/2} M_2^{1/2} v_i$. We define the operator norm for the tensor X

$$\|X\|_{op} = \sup_{\|u^{(i)}\|=1} \langle X, u^{(1)} \otimes \cdots \otimes u^{(k)} \rangle$$

where $\langle X, Y \rangle = \sum X_{i_1 \dots i_k} Y_{i_1 \dots i_k}$ for X, Y to be two tensors. We say a tensor X to be symmetric if for any permutation $\pi \in \mathfrak{S}_k$ such that $X_{i_1 \dots i_k} = X_{i_{\pi(1)} \dots i_{\pi(k)}}$. Under symmetry, we can write $\|X\|_{op} = \sup_{\|u\|=1} |\langle X, u \otimes \cdots \otimes u \rangle|$. To find the individual w_i, v_i , an orthogonal CP decomposition is needed. Tucker's decomposition identifies multilinear subspaces and a core tensor, and can be useful as a preliminary dimension reduction; it does not by itself identify the individual components. For $X \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $X = T \times (U^{(1)}, U^{(2)}, U^{(3)})$ where

1. $T \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor;
2. $U^{(1)} \in \mathbb{R}^{n_1 \times r_1}, U^{(2)} \in \mathbb{R}^{n_2 \times r_2}, U^{(3)} \in \mathbb{R}^{n_3 \times r_3}$ have orthonormal columns.

Then we can write

$$X_{i_1 i_2 i_3} = \sum_{j_1, j_2, j_3} T_{j_1 j_2 j_3} U_{i_1 j_1}^{(1)} U_{i_2 j_2}^{(2)} U_{i_3 j_3}^{(3)}.$$

To obtain the $U^{(1)}$ matrix for example, we unfold the tensor X into a n_1 -by- $(n_2 \times n_3)$ matrix and compute the singular value decomposition so that the left singular matrix is $U^{(1)}$. The following identity holds when each $U^{(j)}$ spans the full column space of the corresponding unfolding. For truncated ranks it is an approximation unless the tensor has those multilinear ranks:

$$\begin{aligned} X &= X \times (U^{(1)}(U^{(1)})^\top, U^{(2)}(U^{(2)})^\top, U^{(3)}(U^{(3)})^\top) \\ &= X \times ((U^{(1)})^\top, (U^{(2)})^\top, (U^{(3)})^\top) \times (U^{(1)}, U^{(2)}, U^{(3)}) \\ &= T \times (U^{(1)}, U^{(2)}, U^{(3)}). \end{aligned}$$

Next we consider the rank 1 spiked tensor model. First, similar to the spiked matrix case, it can be divided into two cases:

- (Symmetric)

$$X = \beta v \otimes v \otimes \cdots \otimes v + Z,$$

where $\|v\| = 1$ and $\beta \geq 0$. To fix the normalization for the constants below, let

$$Z = \frac{1}{k!} \sqrt{\frac{k}{n}} \sum_{\pi \in \mathfrak{S}_k} G^\pi, \quad G_{i_1, \dots, i_k} \text{ i.i.d. } N(0, 1),$$

where $G_{i_1, \dots, i_k}^\pi = G_{i_{\pi(1)}, \dots, i_{\pi(k)}}$. Thus entries with distinct indices have variance $1/(n(k-1)!)$, while repeated indices have the variances imposed by symmetrization. In particular, $\mathbb{E}[Z(u, \dots, u)Z(v, \dots, v)] = (k/n)\langle u, v \rangle^k$ for unit u, v . This is the convention in Montanari and Richard, *A statistical model for tensor PCA* (2014).

- (Asymmetric)

$$X = \beta v_1 \otimes v_2 \otimes \cdots \otimes v_k + Z,$$

where the entries of Z are independent $N(0, 1/n)$, and $\|v_j\| = 1$.

For the symmetric case, the maximum likelihood estimator is written as

$$\hat{\mu}_{\text{ML}} \in \operatorname{argmax}_{\|u\|=1} \{ \beta \langle u, v \rangle^k + Z(u, u, \dots, u) \}.$$

Claim 4.3.1. *There exists $\mu_k = \sqrt{k \log k} (1 + o_k(1))$ for large k , such that, almost surely for the preceding symmetric Gaussian model,*

$$\begin{aligned} \lim_{n \rightarrow \infty} \|Z\|_{\text{op}} &= \mu_k \text{ (for } k \text{ is even),} \\ \limsup_{n \rightarrow \infty} \|Z\|_{\text{op}} &\leq \mu_k \text{ (for } k \text{ is odd).} \end{aligned}$$

Combining this proposition, we have the following theorem for the maximum likelihood estimator:

Theorem 4.3.2. *If $\beta > \mu_k$, then for every fixed $\delta > 0$, with probability tending to one, $|\langle \hat{\mu}_{\text{ML}}, v \rangle|^k \geq 1 - \frac{2\mu_k + \delta}{\beta}$ for large n .*

Proof. By definition,

$$\beta \langle v, \hat{\mu}_{\text{ML}} \rangle^k + Z(\hat{\mu}_{\text{ML}}, \dots, \hat{\mu}_{\text{ML}}) \geq \beta + Z(v, \dots, v),$$

which gives

$$\langle v, \hat{\mu}_{\text{ML}} \rangle^k \geq 1 - \frac{Z(\hat{\mu}_{\text{ML}}, \dots, \hat{\mu}_{\text{ML}}) - Z(v, \dots, v)}{\beta} \geq 1 - \frac{2\|Z\|_{\text{op}}}{\beta} \geq 1 - \frac{2\mu_k + \delta}{\beta}$$

with probability tending to one. The displayed bound is nontrivial when $\beta > 2\mu_k + \delta$. To obtain a nontrivial guarantee for every $\beta > \mu_k$, use additionally $Z(v, \dots, v) \sim N(0, k/n) = o_{\mathbb{P}}(1)$, with v deterministic or independent of Z . The same comparison then gives the stronger bound

$$|\langle v, \hat{\mu}_{\text{ML}} \rangle|^k \geq 1 - \frac{\mu_k + \delta}{\beta}$$

with probability tending to one. □

Note that when $k \geq 3$, the argmax function above is nonconvex. Under the worst case, this question is NP hard. For randomly initialized tensor power iteration, $n^{(k-2)/2}$ is a signal-strength scale appearing in more refined analyses and heuristics; this is not a guarantee for an arbitrary gradient-descent scheme. The result depends on the noise convention, initialization, iteration rule, and possible logarithmic factors.

To improve this rate, we introduce the method of tensor unfolding when $k = 3$. Unfold the tensor $v \otimes v \otimes v \in \mathbb{R}^{n \times n \times n}$ to a matrix in $\mathbb{R}^{n \times n^2}$ by merging the last two dimensions. Similarly, the tensor $Z \in \mathbb{R}^{n \times n \times n}$ is unfolded to a matrix $\tilde{Z} \in \mathbb{R}^{n \times n^2}$. For the asymmetric independent-entry noise model, one can use the BBP transition for long matrices. Symmetric tensor noise produces dependencies after unfolding, so the following independent-entry theorem cannot be applied to it directly.

Proposition 4.3.3 (BBP transition for long matrix). *Consider the matrix*

$$\beta v u^\top + W \in \mathbb{R}^{n \times m}, \quad n \leq m \leq n^l.$$

Assume v, u are deterministic unit vectors, W has independent identically distributed centered entries of variance $1/n$ with $\mathbb{E}|\sqrt{n}W_{ij}|^q \leq C_q$ for every fixed q , and l is fixed. Let $\phi = \sqrt{m/n}$ and $\beta = \lambda\sqrt{\phi}$, with $\lambda > 0$ fixed. Let s_1 be the largest singular value and v_1 the corresponding left singular vector. The following asymptotic profiles hold in probability (Ben Arous, D. Huang, and J. Huang, Long Random Matrices and Tensor Unfolding):

$$1). \quad \lambda > 1, \quad s_1 = \sqrt{\phi^2 + (\lambda^2 + \lambda^{-2})\phi + 1} + o_{\mathbb{P}}(1), \quad |\langle v_1, v \rangle| \approx \sqrt{\frac{\lambda^4 - 1}{\lambda^4 + \lambda^2/\phi}}.$$

$$2). \quad 0 < \lambda \leq 1: \quad s_1 = \phi + 1 + o_{\mathbb{P}}(1) \quad \text{and} \quad |\langle v_1, v \rangle| \rightarrow 0.$$

Rescale $W = \frac{\tilde{W}}{\sqrt{n}}$, then the eigenvalues of $\frac{1}{m}\tilde{W}\tilde{W}^\top$ converges to the Marchenko-Pastur distribution with $\tau = \frac{n}{m}$. In the spiked k-tensor problem, take $m = n^{k-1}$, then $\phi = \sqrt{\frac{m}{n}} = n^{\frac{k-2}{2}}$, which gives

$$\beta = \lambda n^{\frac{k-2}{4}} (\lambda > 1), \quad |\langle v_1, v \rangle| \approx \sqrt{\frac{\lambda^4 - 1}{\lambda^4}}.$$

Combining tensor unfolding with a suitably initialized refinement gives algorithms at the $n^{(k-2)/4}$ signal scale in the independent-entry setting above, rather than the larger scale associated with naive random initialization. These are algorithm-specific sufficient scales, not an unconditional lower bound on all polynomial-time algorithms.

The statistical lower bound must be specified separately from the noise-norm constant μ_k . For example, for a uniform spherical spike and the symmetric normalization above, define

$$\beta_k^{2\text{nd}} = \inf_{0 < q < 1} \sqrt{-\frac{k \log(1 - q^2)}{2q^k}}.$$

For $\beta < \beta_k^{2\text{nd}}$, a bounded likelihood-ratio second moment rules out consistent detection. Indeed, the second moment is $\mathbb{E} \exp((n\beta^2/k)\langle v, v' \rangle^k)$, and the overlap of independent spherical vectors has density proportional to $(1 - q^2)^{(n-3)/2}$. The defining inequality makes the exponential rate negative away from $q = 0$; near zero the integral is bounded (and for $k \geq 3$ it tends to one). This is a lower bound, not an identification of the exact statistical threshold. The arguments in these notes do not settle the optimal polynomial-time threshold between the constant statistical scale and the unfolding scale.

Chapter 5

Statistical Learning

5.1 Lecture 10: How is Random Matrix Theory Used in Machine Learning? Guest Lecture by Zhun Deng

5.1.1 Overview

Some decades ago, deep learning was a very empirical field, leading mathematicians to believe it was founded on tricks and heuristics. In 2017, however, a number of theoreticians began to study how overparameterized neural network models generalized well to test data, seemingly contradicting conventional notions from statistical learning theory. Around that time, Horng-Tzer Yau started studying the dynamics of neural networks. Others began studying why local minima achieved by usual gradient-based optimization algorithms generally work well.

Today, we're entering a new era of AI which is meaningful for theoreticians and mathematicians (e.g. recent developments in automated theorem proving). This talk studies some interesting theoretical problems in machine learning.

Transfer Learning

Transfer learning involves training a model on a source scenario, then deploying on a target scenario. This sounds simple, but it wasn't until recently that pretrained models were able to work very well, such as in large language models.

In transfer learning, instead of training a model from scratch, we take a model trained on a different domain for a source task, and adapt it for a target task. There are two variants: one is same domain, different task (e.g. classifying a common group of people based on gender, then race) and different domain, same task (e.g. train on ImageNet, and transfer to CIFAR). Another example of the latter regime is pre-training a neural network for classification on a cat/dog dataset, then modifying the parameters slightly when working on a new dataset.

Transfer learning is needed for data and task augmentation. For example, for gene data, datasets are often imbalanced: just training on the given data might lead to high accuracy only for certain categories. Another frequent application is for large language models. Another reason is to reduce computational costs: transfer learning involves "fixing" part of the model, so the costs of training are reduced.

In *fixed-feature learning*, one trains from scratch on source data, freeze some layers of the network, then fine-tune remaining layers on the target data (often just the last layer or the last few layers, although some have experimented with fine-tuning the initial layer too). In *full-model fine-tuning*, we fine-tune all layers, and use the pre-trained parameters for initialization.

For some papers which study theoretical aspects of transfer learning, see:

- Provable Meta-Learning of Linear Representations (Tripuraneni et al.),
- Few-Shot Learning via Learning the Representation, Provably (Du et al.),
- Adversarial Training Helps Transfer Learning via Better Representations (Deng et al.).

We shall now proceed to discuss such aspects now.

5.1.2 Transfer Learning: Theoretical Development

Setting

Assume there are T source tasks in total, and that for each task $t \in [T]$, we have a corresponding training data set

$$S_t = \left\{ (x_1^{(t)}, y_1^{(t)}), \dots, (x_{n_t}^{(t)}, y_{n_t}^{(t)}) \right\},$$

where $x_i^{(t)} \in \mathcal{X} \subseteq \mathbb{R}^p$ and $y_i^{(t)} \in \{-1, 1\}$ are i.i.d., drawn from some distribution $\mathcal{P}_{x,y}^{(t)}$. The aim of transfer learning is to learn a representation from the T tasks, freeze this representation, then run the resulting model on the $T+1^{\text{th}}$ task, training some extra task-specific layers.

The data-generating process is given by

$$x_i^{(t)} = \eta_i^{(t)} + y_i^{(t)} \mu_t,$$

for i.i.d. noise $\eta_i^{(t)}$ and $\eta_i^{(t)} \perp y_i^{(t)}$, where

$$\mu_t = B a_t \in \mathbb{R}^p, \quad a_t \in \mathbb{R}^r, \quad B \in \mathbb{R}^{p \times r}$$

is an orthonormal matrix, i.e., $B^T B = I_r$. We further assume that $p \gg r$, so we map a low-dimensional data signal into a high-dimension (it is generally regarded as a fact that high-dimensional image data lie on low-dimensional manifolds), and that $\eta_i^{(t)}$ is zero-mean and ρ_t^2 -sub-Gaussian¹, i.e.,

$$\mathbb{E} \left[\exp \left(\lambda v^T \eta_i^{(t)} \right) \right] \leq \exp \left(\frac{\lambda^2 \rho_t^2}{2} \right)$$

for all $v \in \mathbb{S}^{p-1}$ and $\lambda \in \mathbb{R}$. Throughout this talk, we assume $\rho_t = \Theta(1)$ for all t .

We now define the relevant loss functions for our setup. For each task $t \in [T+1]$,

$$\ell \left(x, y, W_2^{(t)}, W_1 \right) = -y f^{(t)}(x),$$

where $f^{(t)}(\cdot)$ is a two-layer linear neural network, i.e. $f^{(t)}(x) = (W_2^{(t)})^T W_1^T x$ for $W_1 \in O_{p \times r}$, $W_2^{(t)} \in \mathbb{R}^r$, and $\|W_2^{(t)}\| \leq 1$. Here $O_{p \times r} = \{U \in \mathbb{R}^{p \times r} : U^T U = I_r\}$ is the Stiefel manifold. We interpret W_1 as a “good” common representation, and $W_2^{(t)}$ as a task-specific linear classifier. The empirical loss for task t is given by

$$\hat{L} \left(S_t, W_2^{(t)}, W_1 \right) = -\frac{1}{n_t} \sum_{i=1}^{n_t} y_i^{(t)} \left\langle W_1 W_2^{(t)}, x_i^{(t)} \right\rangle$$

while the expected loss for task t is

$$L \left(\mathcal{P}_{x,y}^{(t)}, W_2^{(t)}, W_1 \right) = -\mathbb{E}_{(x,y) \sim \mathcal{P}_{x,y}^{(t)}} \left[y \langle W_1 W_2^{(t)}, x \rangle \right].$$

Finally, the evaluation criterion is the excess population risk on the target task:

$$\begin{aligned} \mathcal{R}(\hat{W}_1, \hat{W}_2^{(T+1)}) &= L(\mathcal{P}_{x,y}^{(T+1)}, \hat{W}_2^{(T+1)}, \hat{W}_1) \\ &\quad - \min_{\|W_2\| \leq 1, W_1 \in O_{p \times r}} L(\mathcal{P}_{x,y}^{(T+1)}, W_2, W_1). \end{aligned}$$

¹Sub-Gaussianity is a fairly flexible condition: it can be applied to image sets that are all bounded, which implies sub-Gaussianity.

How to quantify “good”

We begin by defining a notion of *subspace distance*. Consider two subspaces S_1, S_2 in the same p -dimensional space, represented by full-column-rank matrices $S_1, S_2 \in \mathbb{R}^{p \times r}$. Write both in terms of their QR decompositions to yield

$$S_1 = Q_1 R_1, \quad S_2 = Q_2 R_2.$$

Writing $Q_1^T Q_2 = U \Sigma V^T$ in terms of its singular value decomposition, where $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$, let us define

$$\theta_1 = \cos^{-1}(\sigma_1), \theta_2 = \cos^{-1}(\sigma_2), \dots, \theta_r = \cos^{-1}(\sigma_r).$$

Then the subspace distance between S_1, S_2 is defined as

$$\|\sin \Theta(Q_1, Q_2)\|_F, \text{ where } \Theta(Q_1, Q_2) = \text{diag}(\theta_1, \dots, \theta_r).$$

For one more remark, note that during our setup, we have a model non-identifiability issue, since for any $r \times r$ orthonormal matrix Ω ,

$$W_2^T W_1^T = W_2^T \Omega^{-1} \Omega W_1^T$$

However, we can still hope to learn a good $\hat{W}_1$, for which

$$\|\sin \Theta(\hat{W}_1, B)\|_F \rightarrow 0.$$

In other words, the subspace distance between the learned $\hat{W}_1$ and the matrix B used for the data generating process vanishes.

Algorithm

Given input $\{S_t\}_{t=1}^{T+1}$ (the datasets in our setup), we proceed as follows.

1. Optimize the loss function on each individual source task $t \in [T]$ and obtain estimates

$$\hat{\beta}_t = \underset{\|\beta_t\| \leq 1}{\text{argmin}} \frac{1}{n_t} \sum_{i=1}^{n_t} -y_i^{(t)} \langle \beta_t, x_i^{(t)} \rangle.$$

2. Let $\hat{W}_1 \Sigma V^T$ denote the rank- r truncation of the SVD of $[\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_T]$, so that Σ is an $r \times r$ diagonal matrix consisting of singular values, and V is a $T \times r$ matrix with orthonormal columns.
3. Update

$$\hat{W}_2^{(T+1)} = \underset{\|\hat{W}_2^{(T+1)}\| \leq 1}{\text{argmin}} \frac{1}{n_{T+1}} \sum_{i=1}^{n_{T+1}} -y_i^{(T+1)} \langle \hat{W}_1 \hat{W}_2^{(T+1)}, x_i^{(T+1)} \rangle.$$

Return $\hat{W}_1, \hat{W}_2^{(T+1)}$.

Analysis

We now analyze the performance of the optimization algorithm. Let us define, for $t \in [T + 1]$,

$$\hat{\mu}_t = \sum_{i=1}^{n_t} \frac{x_i^{(t)} y_i^{(t)}}{n_t} \quad \mu_t = B a_t.$$

Notice that²

$$\hat{J} := \left(\frac{\hat{\mu}_1}{\|\hat{\mu}_1\|}, \dots, \frac{\hat{\mu}_T}{\|\hat{\mu}_T\|} \right) = (\hat{\mu}_1, \dots, \hat{\mu}_T) \text{diag} \left(\frac{1}{\|\hat{\mu}_1\|}, \dots, \frac{1}{\|\hat{\mu}_T\|} \right),$$

where $\hat{\beta}_t = \hat{\mu}_t / \|\hat{\mu}_t\|$ when $\hat{\mu}_t \neq 0$. If an empirical mean vanishes, any vector in the unit ball minimizes that task loss; the normalized formula below is used on the event all empirical means are nonzero. Column normalization preserves the full column span but generally changes the leading singular subspace, so we must analyze the normalized matrix

$$\hat{\Phi} = (\hat{\mu}_1, \dots, \hat{\mu}_T) \text{diag} \left(\frac{1}{\|\hat{\mu}_1\|}, \dots, \frac{1}{\|\hat{\mu}_T\|} \right).$$

Assume $\mu_t \neq 0$ for all source tasks and that the vectors a_t span $\mathbb{R}^r$ (task diversity). Then the population normalized matrix has rank r and column space equal to that of B . Denote

$$\Phi = (\mu_1, \dots, \mu_T) \text{diag} \left(\frac{1}{\|\mu_1\|}, \dots, \frac{1}{\|\mu_T\|} \right).$$

We briefly mention³ the techniques used to achieve bounds on the evaluation criterion: the heart of the proof follows from a variant of the Davis-Kahan Theorem. Assume $\min\{T, p\} > r$. For simplicity, we denote $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \dots \geq \hat{\sigma}_r$ as the top largest r singular values of $\hat{\Phi}$, and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ as the top r singular values of Φ . Let $V = (v_1, \dots, v_r)$ contain the left singular vectors of Φ associated with $\sigma_1, \dots, \sigma_r > 0$, and let $\hat{V}$ be the corresponding matrix for $\hat{\Phi}$. Then,

$$\left\| \sin \Theta \left(\hat{V}, V \right) \right\|_F \lesssim \frac{\left(2\sigma_1 + \|\hat{\Phi} - \Phi\|_{op} \right) \min \left\{ \sqrt{r} \|\hat{\Phi} - \Phi\|_{op}, \|\hat{\Phi} - \Phi\|_F \right\}}{\sigma_r^2}.$$

Moreover, there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{r \times r}$ such that

$$\|\hat{V}\hat{O} - V\|_F \leq \sqrt{2} \left\| \sin \Theta \left(\hat{V}, V \right) \right\|_F$$

and

$$\|\hat{V}\hat{O} - V\|_F \lesssim \frac{\left(2\sigma_1 + \|\hat{\Phi} - \Phi\|_{op} \right) \min \left\{ \sqrt{r} \|\hat{\Phi} - \Phi\|_{op}, \|\hat{\Phi} - \Phi\|_F \right\}}{\sigma_r^2}.$$

²The columns of the matrix $\hat{J}$ can be thought of as the gradients of the t objectives in Step 1 with respect to the corresponding parameters β_t , normalized to have unit norm.

³See the Canvas notes for the full treatment; we stay consistent to the lecture.

5.1.3 Optimization of Over-Parametrized Neural Networks

One of the biggest mysteries in theoretical deep learning today is that gradient-based optimization algorithms generalize well to unseen data, despite the loss landscape being highly non-convex and the model being overparametrized. This requires a better understanding of three directions: the power of first-order algorithms, the structure of neural networks, and the properties of inputs such as images.

5.2 Lecture 11: Random Feature Model

Consider a general learning problem where the data distribution is given by:

$$x_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \quad , y_i = \theta_{\text{teach}}(\xi^T x_i), \quad \|\xi\| = 1$$

where θ_{teach} represents the underlying teacher model. Our prediction model is defined as:

$$y = f_w(x)$$

In most cases, the model has a sequential structure and can be expressed as:

$$f_w(x) = \theta(w^T r) \quad , r = \mathcal{T}(x)$$

where $\mathcal{T}(\cdot)$ is the feature map and θ is the transform function applied to it. Feature maps are essential to every learning model and studying them can help us understand what a model can learn and how it processes complex data. [Rahimi & Recht '08] first proposed mapping input data to a randomized low-dimensional feature space and then applying fast linear methods.

In the Random Feature Model (RFM) setting, the learner receives n samples of the form $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ and aims to learn the weights $w \in \mathbb{R}^p$ that make the following model fit the data:

$$f_w(x) = \frac{a^T w}{\sqrt{p}} \quad \text{for } a = \sigma(Fx),$$

where $F \in \mathbb{R}^{p \times d}$ is a random transform matrix and $\sigma(\cdot)$ acts coordinatewise. To state the Gaussian equivalence used below, assume the rows of F are independent $N(0, I_d/d)$, independent of all inputs, and $p/d, n/d$ tend to positive finite constants. A sufficient set of regularity hypotheses is: σ is odd with bounded first three derivatives; $r(w) = \sum_{j=1}^p h(w_j)$ with $h'' \geq c > 0$ and bounded $h^{(3)}$; $\ell(y, s) \geq 0$ is convex and three times differentiable in s , its third derivative is bounded by a polynomial in the teacher argument uniformly in s , and its value and first two derivatives at $s = 0$ have the same polynomial growth. Assume also that the teacher has polynomial growth. These are the hypotheses relevant here from Hu and Lu, *Universality Laws for High-Dimensional Learning with Random Features* (arXiv:2009.07669); equivalence is not asserted for arbitrary feature maps, losses, or regularizers. The discussion of more general activations below describes the surrogate construction; the proof outline uses the stated odd-activation hypotheses. We refer to a as the random feature vector. The learner trains by solving the optimization problem

$$\min_w \sum_{i=1}^n \ell(y_i, \frac{a_i^T w}{\sqrt{p}}) + r(w),$$

where $\ell(\cdot, \cdot)$ is a loss function and $r(\cdot)$ is a regularizer. The optimal solution found by ERM is denoted as $\hat{w}$. The train and test errors are defined as

$$\varepsilon_{\text{train}} = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \frac{a_i^T \hat{w}}{\sqrt{p}}) \quad \text{and} \quad \varepsilon_{\text{test}} = \mathbb{E} \ell(y_{\text{new}}, \frac{a_{\text{new}}^T \hat{w}}{\sqrt{p}})$$

The explicitly nonlinear transform hinders a straightforward analysis of the learning performance of the RFM under random features $\{a_i\}_{i=1}^n$. In these notes, we will take a step towards showing that the above RFM setup is asymptotically equivalent to the following linear Gaussian model setup

$$b_i = \mu_0 \mathbf{1}_p + \mu_1 F x_i + \mu_2 z_i$$

where, after fixing $\mu_G(dx)$ as the standard Gaussian measure, we define the coefficients as

$$\begin{aligned}\mu_0 &= \int \sigma(x) \mu_G(dx) \\ \mu_1 &= \int \sigma(x) x \mu_G(dx) \\ \mu_2 &= \sqrt{\int \sigma(x)^2 \mu_G(dx) - \mu_0^2 - \mu_1^2}\end{aligned}$$

and $z_i \sim N(0, I_p)$ are independent across i and independent of F and the inputs. Here $\mathbf{1}_p$ is the all-ones vector; under oddness, $\mu_0 = 0$. This is exactly consistent with the well-known Gaussian equivalence conjecture, which states that as $d, p, n \rightarrow \infty$ while $p/d \rightarrow c_1$ and $n/d \rightarrow c_2$, “swapping out” the difficult-to-analyze random features by their Gaussian feature counterparts does not change the train and test error by much, i.e.

$$\varepsilon_{\text{train}}(A) \approx \varepsilon_{\text{train}}(B) \quad \text{and} \quad \varepsilon_{\text{test}}(A) \approx \varepsilon_{\text{test}}(B),$$

where A, B are respectively the stacks of random features a_i and Gaussian features b_i .

The proof outline below compares the regularized optimal objective. Deducing unregularized training error and test error requires additional perturbation and joint-comparison arguments; those conclusions do not follow from a single objective comparison alone. Our proof will use the classic Lindeberg technique, which can be thought of as interpolating between the train and test errors with respect to A and B by swapping the matrices row by row, then showing the change in the train and test errors with respect to each swap is small. We can think of the interpolation at step k to be

$$S_k = \begin{bmatrix} b_1^T \\ \vdots \\ b_k^T \\ a_{k+1}^T \\ \vdots \\ a_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

More rigorously, we will show that for

$$\Phi_A = \min_w \sum_{i=1}^n \ell(y_i, \frac{a_i^T w}{\sqrt{p}}) + r(w) \quad \text{and} \quad \Phi_B = \min_w \sum_{i=1}^n \ell(y_i, \frac{b_i^T w}{\sqrt{p}}) + r(w)$$

and a smooth enough test function φ , we have that

$$\left| \mathbb{E} \varphi\left(\frac{\Phi_A}{n}\right) - \mathbb{E} \varphi\left(\frac{\Phi_B}{n}\right) \right| \rightarrow 0$$

Keeping the above interpolation scheme in mind, we can define

$$\Phi_k = \min_w \left\{ \sum_{i=1}^k \ell(y_i, b_i^T w / \sqrt{p}) + \sum_{i=k+1}^n \ell(y_i, a_i^T w / \sqrt{p}) + r(w) \right\}.$$

Then we have that

$$\varphi(\Phi_B/n) - \varphi(\Phi_A/n) = \sum_{k=1}^n [\varphi(\Phi_k/n) - \varphi(\Phi_{k-1}/n)].$$

It will suffice to show that

$$\sup_{1 \leq k \leq n} |\mathbb{E}[\varphi(\Phi_k/n) - \varphi(\Phi_{k-1}/n)]| = o(1/n).$$

Furthermore, if we define

$$F_{\setminus k}(w) = \sum_{i=1}^{k-1} \ell(y_i, \frac{b_i^T w}{\sqrt{p}}) + \sum_{i=k+1}^n \ell(y_i, \frac{a_i^T w}{\sqrt{p}}) + r(w),$$

then we can write, for this fixed swap k ,

$$\begin{aligned} \Phi_k &= \min_w \{F_{\setminus k}(w) + \ell(y_k, b_k^T w / \sqrt{p})\} \approx \Phi_{\setminus k} + \Delta(b_k) =: \Psi_k, \\ \Phi_{k-1} &= \min_w \{F_{\setminus k}(w) + \ell(y_k, a_k^T w / \sqrt{p})\} \approx \Phi_{\setminus k} + \Delta(a_k) =: \Psi_{k-1}. \end{aligned}$$

The approximations Ψ_k, Ψ_{k-1} in this paragraph are local to the k -th swap; they are not one globally defined interpolation sequence. Thus

$$\Phi_k - \Phi_{k-1} \approx \Psi_k - \Psi_{k-1} = \Delta(b_k) - \Delta(a_k).$$

To be clear, we are viewing the k -th swap as a perturbation and expanding the leave-one-out objective to second order. Set

$$\Phi_{\setminus k} = \min_w F_{\setminus k}(w), \quad w_{\setminus k}^* = \operatorname{argmin}_w F_{\setminus k}(w), \quad H_{\setminus k} = \nabla^2 F_{\setminus k}(w_{\setminus k}^*).$$

Strong convexity ensures uniqueness and $H_{\setminus k} \succeq cI_p$. The linear Taylor term is zero at the minimizer. Accordingly, for $r_k \in \{a_k, b_k\}$ define

$$\Delta(r_k) = \min_w \left\{ \frac{1}{2} (w - w_{\setminus k}^*)^T H_{\setminus k} (w - w_{\setminus k}^*) + \ell(y_k, r_k^T w / \sqrt{p}) \right\}.$$

The additive constant $\Phi_{\setminus k}$ is not part of Δ .

The required quantitative leave-one-out estimate is, uniformly in k ,

$$\mathbb{E}|\Phi_k - \Psi_k|^2 + \mathbb{E}|\Phi_{k-1} - \Psi_{k-1}|^2 \leq C \frac{\text{polylog}(n)}{n}. \quad (5.2.1)$$

Here and below $\text{polylog}(n)$ means a fixed power of $\log n$, with constants uniform in the swap index. This is a quantitative input, not a consequence of a formal Taylor expansion alone; its proof controls the optimizing weights and the Taylor remainder under the hypotheses above (Hu–Lu).

Actually, we can simplify $\Delta(r_k)$ using its first-order condition:

$$H_{\setminus k}(w - w_{\setminus k}^*) + \frac{r_k}{\sqrt{p}} \ell'(y_k, r_k^T w / \sqrt{p}) = 0,$$

where the prime differentiates the second argument of ℓ . Put

$$s_k = \frac{r_k^T w_{\setminus k}^*}{\sqrt{p}}, \quad \gamma_k(r_k) = \frac{r_k^T H_{\setminus k}^{-1} r_k}{p}.$$

For fixed $x = r_k^T w / \sqrt{p}$, minimizing the quadratic form subject to this linear constraint gives $(x - s_k)^2 / (2\gamma_k(r_k))$. Therefore

$$\Delta(r_k) = \min_{x \in \mathbb{R}} \left\{ \frac{(x - s_k)^2}{2\gamma_k(r_k)} + \ell(y_k, x) \right\} =: \mathcal{M}_{\ell(y_k, \cdot)}(s_k; \gamma_k(r_k)).$$

This is the Moreau envelope. When $r_k = 0$ and $\gamma_k = 0$, interpret it as $\ell(y_k, 0)$. Its first-order condition is $x - s_k + \gamma_k \ell'(y_k, x) = 0$; multiplying the entire objective by γ_k would preserve the minimizer but would change the value and hence would not equal $\Delta(r_k)$. The corresponding moment bound needed below is

$$\mathbb{E}\Delta(a_k)^2 + \mathbb{E}\Delta(b_k)^2 \leq C \text{polylog}(n). \quad (5.2.2)$$

These estimates can first be proved conditionally on a high-probability regularity event for F and then averaged, controlling the exceptional event by moment bounds.

Now let $\mathcal{H}_k$ be the sigma-field generated by F and all samples and auxiliary Gaussian vectors except those at index k . The leave-one-out quantities are $\mathcal{H}_k$ -measurable. For a test function φ with bounded first two derivatives, Taylor expansion at $\Phi_{\setminus k}/n$ gives

$$\begin{aligned} & |\mathbb{E}[\varphi(\Phi_k/n) - \varphi(\Phi_{\setminus k-1}/n)]| \\ & \leq \frac{\|\varphi'\|_\infty}{n} \mathbb{E}|\mathbb{E}[\Phi_k - \Phi_{\setminus k-1} \mid \mathcal{H}_k]| \\ & \quad + \frac{\|\varphi''\|_\infty}{2n^2} \mathbb{E}[(\Phi_k - \Phi_{\setminus k})^2 + (\Phi_{\setminus k-1} - \Phi_{\setminus k})^2] \\ & \leq \frac{\|\varphi'\|_\infty}{n} \mathbb{E}|\mathbb{E}[\Delta(b_k) - \Delta(a_k) \mid \mathcal{H}_k]| + \mathcal{O}\left(\frac{\text{polylog}(n)}{n^{3/2}}\right). \end{aligned}$$

The last line uses Cauchy–Schwarz, (5.2.1), and (5.2.2). Thus the remaining required comparison is

$$\sup_{k \leq n} \mathbb{E}|\mathbb{E}[\Delta(b_k) - \Delta(a_k) \mid \mathcal{H}_k]| \leq C \frac{\text{polylog}(n)}{\sqrt{n}}.$$

Together these estimates give a total Lindeberg error $\mathcal{O}(\text{polylog}(n)/\sqrt{n}) = o(1)$ after summing over the n swaps.

The linear projection $b_k^T w_{\setminus k}^* / \sqrt{p}$ is Gaussian conditionally on $\mathcal{H}_k$, but this fact alone does not prove the comparison: the label y_k and the random quadratic form $\gamma_k(r_k)$ also enter the Moreau envelope. One needs a joint Gaussian approximation including the teacher coordinate and concentration of the quadratic form, as well as delocalization of the leave-one-out weights. These are

the additional model-dependent inputs in the Hu–Lu proof. The scalar calculation below illustrates Stein’s method; it is not by itself a proof of that dependent-feature comparison. Stein’s method exploits the Gaussian identity

$$\mathbb{E}[f'(G) - Gf(G)] = 0, \quad G \sim N(0, 1).$$

Stein’s method works by using a characterization equation for the normal distribution, which allows one to compare any distribution to the normal distribution by solving a differential equation. Stein’s method can also handle different types of probability metrics, such as total variation, Kolmogorov, and Wasserstein.

So to show that for a test function $\varphi(\cdot)$ and a sequence of distributions $\{Z_n\}$ that

$$\mathbb{E}[\varphi(Z_n) - \varphi(G)] \rightarrow 0,$$

we can solve for f that satisfies the differential equation

$$\varphi(x) - \mathbb{E}\varphi(G) = f'(x) - xf(x),$$

then show that

$$\mathbb{E}[f'(Z_n) - Z_nf(Z_n)] \rightarrow 0.$$

Consider the following toy example to build intuition for this method. Suppose $\{X_i\}_{i=1}^n$ is i.i.d. sample from distribution $\mathcal{D}$ s.t. $\mathbb{E}X_1 = 0, \mathbb{E}X_1^2 = 1, \mathbb{E}|X_1|^3 < \infty$. We want to show a CLT result for $Z_n = \frac{X_1 + \dots + X_n}{\sqrt{n}}$. Observe that

$$\mathbb{E}Z_nf(Z_n) = \sum_{i=1}^n \mathbb{E} \frac{X_i f(Z_n)}{\sqrt{n}}.$$

For each term in the sum,

$$\mathbb{E} \left[\frac{X_i}{\sqrt{n}} f \left(\frac{\sum_{j \neq i} X_j}{\sqrt{n}} + \frac{X_i}{\sqrt{n}} \right) \right] = \mathbb{E} \frac{X_i f(Z_{\setminus i})}{\sqrt{n}} + \mathbb{E} \frac{X_i^2 \int_0^1 f'(Z_{\setminus i} + \frac{X_i t}{\sqrt{n}}) dt}{n} = \mathbb{E} \frac{X_i^2 \int_0^1 f'(Z_{\setminus i} + \frac{X_i t}{\sqrt{n}}) dt}{n}.$$

So

$$\mathbb{E}[f'(Z_n) - Z_nf(Z_n)] = \frac{\sum_{i=1}^n \mathbb{E}[(1 - X_i^2)f'(Z_n)]}{n} + \mathbb{E} \left[\frac{\sum_{i=1}^n X_i^2 [f'(Z_n) - \int_0^1 f'(Z_{\setminus i} + \frac{X_i t}{\sqrt{n}}) dt]}{n} \right].$$

For a Lipschitz test function φ , the bounded solution to the Gaussian Stein equation has f' Lipschitz with a constant L controlled by the Lipschitz constant of φ . Write $Z_{\setminus i} = n^{-1/2} \sum_{j \neq i} X_j$. This implies that for $t \in [0, 1]$,

$$\left| f'(Z_{\setminus i} + \frac{X_i t}{\sqrt{n}}) - f'(Z_{\setminus i}) \right| \leq \frac{L|X_i|}{\sqrt{n}}.$$

Then

$$\left| \mathbb{E} \left[(1 - X_i^2) f'(Z_n) \right] \right| = \left| \mathbb{E} \left[(1 - X_i^2) (f'(Z_n) - f'(Z_{\setminus i})) \right] \right| \leq L \frac{\mathbb{E}|X_i| + \mathbb{E}|X_i^3|}{\sqrt{n}},$$

which implies that

$$\left| \frac{\sum_{i=1}^n \mathbb{E}[(1 - X_i^2)f'(Z_n)]}{n} \right| \leq L \frac{\mathbb{E}|X_i| + \mathbb{E}|X_i^3|}{\sqrt{n}}.$$

Symmetrically, we can show that

$$\left| \mathbb{E} \left[\frac{\sum_{i=1}^n X_i^2 [f'(Z_n) - \int_0^1 f'(Z_{\setminus i} + \frac{X_i t}{\sqrt{n}}) dt]}{n} \right] \right| \leq L \frac{\mathbb{E}|X_i|^3}{\sqrt{n}},$$

which lets us conclude

$$|\mathbb{E}[f'(Z_n) - Z_n f(Z_n)]| = O(n^{-1/2}).$$

For the random-feature comparison, the coordinates within one row $a_i = \sigma(Fx_i)$ are correlated through the common input x_i when F is fixed. The different sample rows are independent conditionally on F . Extending the scalar Stein calculation requires the joint and quadratic-form estimates specified above.

5.3 Lecture 12.1: Approximate message passing

Consider the linear model

$$M = \lambda v^* (v^*)^\top + W$$

with

- $W_{i,j} = W_{j,i} \sim N(0, 1/n)$ for $i < j$, and $W_{i,i} \sim N(0, 2/n)$; the upper-triangular entries are independent, and W is independent of v^* .
- $\lambda = SNR$ (signal-to-noise ratio) with $\|v^*\|_2 = 1$.
- Goal: Estimate v^* from M
- Phase Transition at $\lambda > 1$: the top eigenvalue separates from bulk, eigenvector correlates non-trivially with v^* .

The linear model has multiple applications e.g., spin-glass problems, community detection, image alignment:

- $\mathbb{Z}_2$ synchronization: $\sqrt{n}v_i^* \sim \text{Unif}\{+1, -1\}$.
- sparse Wigner model: $\|v^*\|_0 = k$.
- non-negative Wigner model: $v_i^* \geq 0$.
- cone-constrained spiked models: $v^* \in \mathbf{K}$ (e.g., monotone, convex)

We consider the approximate message passing algorithm for spiked models:

$$x_{t+1} = M\eta_t(x_t) - \langle \eta'_t(x_t) \rangle \cdot \eta_{t-1}(x_{t-1}), \quad \text{for } t \geq 1$$

where $\langle x \rangle = n^{-1} \sum_i x_i$, while $\langle u, v \rangle = u^T v$ denotes the Euclidean inner product. Unless stated otherwise, η_t acts coordinatewise and $\eta_0(x_0) = 0$. We denote η_t to be the denoising function selected a priori:

- $\mathbb{Z}_2$ synchronization: $\eta_t(x) = \rho_t \tanh(\pi_t x)$.
- sparse estimation: $\eta_t(x) = \rho_t \text{Sign}(x)(|x| - \tau_t)_+$.
- general closed convex cone: $\eta_t(x) = \rho_t \text{Proj}_{\mathbf{K}}(x)$. This is generally nonseparable; its Onsager coefficient is a normalized Jacobian trace, not the coordinatewise derivative above. A separate nonseparable AMP theorem is needed in this case.

We first consider the exact asymptotic state evolution. Since $\|v^*\| = 1$, the scalar coordinates of order one are $\sqrt{n}v_i^*$, not v_i^* .

Theorem 5.3.1. *Suppose the empirical law of $\{\sqrt{n}v_i^*\}_{i=1}^n$ converges in Wasserstein distance of order two to μ_V , with $\mathbb{E}V^2 = 1$. Let $f_t : \mathbb{R} \rightarrow \mathbb{R}$ be fixed continuously differentiable Lipschitz functions with Lipschitz derivatives, and use*

$$\eta_t(z) = n^{-1/2} f_t(\sqrt{n} z).$$

Initialize $x_1 = \alpha_1 v^* + \beta_1 g / \sqrt{n}$, where $g \sim N(0, I_n)$ is independent of W, v^* , and α_1, β_1 are fixed. For each fixed t , in probability,

$$\frac{1}{n} \sum_{i=1}^n (\sqrt{n} x_{t,i} - \sqrt{n} v_i^*)^2 = \|x_t - v^*\|_2^2 \longrightarrow \mathbb{E}[(\alpha_t V + \beta_t G - V)^2],$$

where $V \sim \mu_V$ and $G \sim N(0, 1)$ are independent.

Within the theorem, the deterministic state evolution is characterized as

$$\begin{aligned} \alpha_{t+1} &= \lambda \mathbb{E}[V f_t(\alpha_t V + \beta_t G)], \\ \beta_{t+1}^2 &= \mathbb{E}[f_t(\alpha_t V + \beta_t G)^2]. \end{aligned}$$

The empirical joint convergence of $(\sqrt{n} v_i^*, \sqrt{n} x_{t,i})$ underlying this statement is the Gaussian AMP state-evolution theorem, used here as an input. The displayed square error follows by applying it to a pseudo-Lipschitz function of order two. Lipschitz nonsmooth denoisers require the corresponding version of that theorem. Classical fixed-iteration state evolution does not itself give a finite-sample guarantee for a growing number of iterations. Therefore, the question is that: Is it possible to develop non-asymptotic analysis of AMP beyond $o(\log(n)/\log(\log(n)))$ iterations?

Theorem 5.3.2. *Initialize AMP with x_1 independent of W and $\eta_0(x_0) = 0$, and suppose the scalar denoisers are differentiable except at finitely many points. For $1 \leq t < n$, AMP admits the decomposition*

$$x_{t+1} = \alpha_{t+1} v^* + \sum_{k=1}^t \beta_t^k \phi_k + \xi_t,$$

where the ϕ_k are independent $N(0, I_n/n)$ vectors,

$$\alpha_{t+1} = \lambda (v^*)^T \eta_t(x_t), \quad \|\beta_t\|_2 = \|\eta_t(x_t)\|_2.$$

This is the decomposition theorem of Li and Wei, *A Non-Asymptotic Framework for Approximate Message Passing in Spiked Models* (arXiv:2208.03313, Theorem 1). The coefficients may depend on the Gaussian vectors, so their weighted sum is not automatically Gaussian. In its construction, z_k are the orthonormal directions obtained by orthogonalizing the denoised iterates (completing the basis if needed), and $\beta_t^k = \langle \eta_t(x_t), z_k \rangle$.

For $t \geq 2$, put

$$\begin{aligned} v_t &= \alpha_t v^* + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k = x_t - \xi_{t-1}, \\ \delta_t &= \eta_t(x_t) - \eta_t(v_t), \quad \delta'_t = \eta'_t(x_t) - \eta'_t(v_t). \end{aligned}$$

The same theorem gives a unit vector $\mu_t \in \mathbb{R}^{t-1}$ such that, with probability at least $1 - O(n^{-11})$,

$$\begin{aligned} \|\xi_t\|_2 &= \left\langle \sum_{k < t} \mu_t^k \phi_k, \delta_t \right\rangle - \langle \delta'_t \rangle \sum_{k < t} \mu_t^k \beta_{t-1}^k + \Delta_t + O\left(\sqrt{\frac{t \log n}{n}} \|\beta_t\|_2\right), \\ \Delta_t &= \sum_{k < t} \mu_t^k [\langle \phi_k, \eta_t(v_t) \rangle - \langle \eta'_t(v_t) \rangle \beta_{t-1}^k]. \end{aligned}$$

At $t = 1$, the corresponding initialization residual satisfies $\|\xi_1\|_2 = O(\sqrt{\log n/n}\|\beta_1\|_2)$; no unit vector in $\mathbb{R}^0$ is required.

We have the following error bound, which makes explicit the quantities that must be controlled. Assume for this bound that η_t is twice continuously differentiable and $\|\eta'_t\|_\infty \leq \rho$. Define ϕ_n to be standard Gaussian measure on $\mathbb{R}^n$, set $b = \|\beta_{t-1}\|_2$, and suppose

$$\sup_{\|\mu\|=1} \left| \sum_{k < t} \mu^k [\langle \phi_k, \eta_t(v_t) \rangle - \langle \eta'_t(v_t) \rangle \beta_{t-1}^k] \right| \leq A_t,$$

$$\left| (v^*)^T \left[\eta_t(v_t) - \int \eta_t(\alpha_t v^* + bx/\sqrt{n}) \phi_n(dx) \right] \right| \leq B_t.$$

For $q_s = v_t + s\xi_{t-1}$, let

$$\kappa_t = \sup_{\substack{\|\mu\|=1 \\ 0 \leq s \leq 1}} \left\| \left(\sum_{k < t} \mu^k \phi_k \right) \circ \eta'_t(q_s) - \frac{\sum_{k < t} \mu^k \beta_{t-1}^k}{n} \eta''_t(q_s) \right\|_2,$$

where $\circ$ is coordinatewise multiplication.

Theorem 5.3.3. *On the event of the preceding bounds and the decomposition estimate,*

$$\alpha_{t+1} = \lambda(v^*)^T \int \eta_t(\alpha_t v^* + bx/\sqrt{n}) \phi_n(dx) + \lambda \Delta_{\alpha,t},$$

$$|\Delta_{\alpha,t}| \leq B_t + \rho \|\xi_{t-1}\|_2,$$

$$\|\xi_t\|_2 \leq \kappa_t \|\xi_{t-1}\|_2 + A_t + O\left(\sqrt{\frac{t \log n}{n}} \|\beta_t\|_2\right).$$

The normalization $\|\beta_t\|_2 = 1$ holds only when the denoiser has been chosen so that $\|\eta_t(x_t)\|_2 = 1$.

Proof. The signal estimate follows from the Lipschitz bound and $\|v^*\| = 1$. For the residual, write $\delta_t = \int_0^1 \eta'_t(q_s) \circ \xi_{t-1} ds$ and $\delta'_t = \int_0^1 \eta''_t(q_s) \circ \xi_{t-1} ds$. The first two terms of the residual formula are the inner product of ξ_{t-1} with the integral of the vector defining κ_t at $\mu = \mu_t$. Cauchy-Schwarz and $|\Delta_t| \leq A_t$ give the result. $\square$

It suffices to control $\kappa_t \leq 1 - c$, A_t , and B_t over the desired range of iterations. None of these bounds follows from the decomposition alone. The supremum above includes the dependence of the coefficients on the Gaussian vectors; replacing it by a fixed-coefficient Gaussian expectation needs a separate uniform concentration argument. For nonsmooth denoisers, additional errors near their nondifferentiable points must be included.

The prior art on hybrid procedure is the following:

- Setting: $M = \lambda v^*(v^*)^\top + W$ where $\sqrt{n}v_i^* \sim \text{Unif}\{+1, -1\}$.
- Goal: recover v^* given M .

For the following quantitative result, specify the spectral initialization and the denoiser. Let $g \sim N(0, I_n)$ be independent of M , and take

$$s_0 = \left\lceil C \frac{\lambda^2 \log n}{(\lambda - 1)^2} \right\rceil, \quad x_1 = \frac{\lambda M^{s_0} g}{\|M^{s_0} g\|_2}, \quad \eta_0(x_0) = x_1/\lambda,$$

with a sufficiently large absolute constant C . For each subsequent step use

$$\eta_t(x) = \rho_t \tanh(\pi_t x), \quad \pi_t = \sqrt{n(\|x_t\|_2^2 - 1)}, \quad \rho_t = \|\tanh(\pi_t x_t)\|_2^{-1}.$$

The scalar parameters are held fixed when computing η'_t . These formulas are used on the high-probability event where their square roots and denominators are well defined; the algorithm may be defined arbitrarily on its complement. The sign of the initialization is immaterial for recovery up to sign.

Theorem 5.3.4. *Suppose $1 + \log n/n^{1/16} < \lambda \leq 1.2$. Define*

$$\tau_1 = \lambda^2 - 1, \quad \tau_{t+1} = \lambda^2 \int \tanh(\tau_t + \sqrt{\tau_t} x) \phi(dx),$$

where ϕ is standard Gaussian measure on $\mathbb{R}$. For any deterministic $T_n = o(n(\lambda - 1)^{10}/\log^7 n)$, with probability at least $1 - O(n^{-11})$ the spectrally initialized AMP satisfies, simultaneously for $1 \leq t \leq T_n$,

$$\begin{aligned} x_{t+1} &= \alpha_{t+1} v^* + \sum_{k=-2s_0}^t \beta_t^k \phi_k + \xi_t, \\ \alpha_{t+1} &= \lambda(v^*)^T \eta_t(x_t), \quad \|\beta_t\|_2 = 1, \\ \left| \frac{\alpha_{t+1}^2}{\tau_{t+1}} - 1 \right| &\leq C \left(\sqrt{\frac{t \log n}{(\lambda - 1)^8 n}} + \sqrt{\frac{\log^7 n}{(\lambda - 1)^{14} n}} \right), \\ \|\xi_t\|_2 &\leq C \left(\sqrt{\frac{t \log n}{(\lambda - 1)^3 n}} + \sqrt{\frac{\log^7 n}{(\lambda - 1)^9 n}} \right). \end{aligned}$$

Here the ϕ_k are independent $N(0, I_n/n)$ vectors; the extra indices account for the dependence introduced by the power initialization.

This is the specialization of Li-Wei, Theorem 4; it is not a consequence of applying the independent-initialization theorem to a spectral eigenvector. Comparing to state evolution, let $\tau_* > 0$ be its positive fixed point. In this range of λ , the scalar recursion converges geometrically, and the estimates above imply

$$\begin{aligned} |\alpha_t^2 - \tau_t| &\leq C \left(\sqrt{\frac{t \log n}{(\lambda - 1)^8 n}} + \sqrt{\frac{\log^7 n}{(\lambda - 1)^{14} n}} \right), \\ |\alpha_t^2 - \tau_*| &\leq C(\lambda^2 - 1)(1 - c(\lambda - 1))^{t-1} + C \left(\sqrt{\frac{t \log n}{(\lambda - 1)^8 n}} + \sqrt{\frac{\log^7 n}{(\lambda - 1)^{14} n}} \right). \end{aligned}$$

The deterministic geometric term is a bound, not an equality with a signed random error.

5.4 Lecture 12.2: Ridge Regression

We consider the linear model

$$y = X\beta + \epsilon \quad (5.4.1)$$

with the following assumptions on data generation:

- (a) The population covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is deterministic and symmetric positive definite.
- (b) The rows x_i^T of $X \in \mathbb{R}^{n \times d}$ are independent, with $x_i \sim N(0, \Sigma)$.
- (c) The noise coordinates ϵ_i are independent $N(0, \sigma^2)$ variables.
- (d) The signal β has mean zero and covariance $(\alpha^2/d)I_d$, where $\alpha, \sigma > 0$; X, β, ϵ are mutually independent.

For the asymptotic statements, assume $d/n \rightarrow \gamma \in (0, \infty)$, that $\|\Sigma\|$ is uniformly bounded, and that its empirical spectral measures converge weakly to a probability measure. These conditions ensure the sample spectral limit used below.

We estimate the signal β in (5.4.1) with the Ridge Regression estimator, for $\lambda > 0$:

$$\begin{aligned} \hat{\beta}_\lambda &= \operatorname{argmin}_{b \in \mathbb{R}^d} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\} \\ &= (X^T X + n\lambda I_d)^{-1} X^T y. \end{aligned}$$

One natural question is: What is the prediction risk $r_\lambda(X)$ of $\hat{\beta}_\lambda$? Let $x_0 \sim N(0, \Sigma)$ and $\epsilon_0 \sim N(0, \sigma^2)$ be independent of the training variables, and set $y_0 = x_0^T \beta + \epsilon_0$. Then

$$\begin{aligned} r_\lambda(X) &= \mathbb{E}[(y_0 - x_0^T \hat{\beta}_\lambda)^2 \mid X] \\ &= \sigma^2 + \mathbb{E}[(x_0^T (\hat{\beta}_\lambda - \beta))^2 \mid X] \\ &= \sigma^2 + \mathbb{E}[(\hat{\beta}_\lambda - \beta)^T \Sigma (\hat{\beta}_\lambda - \beta) \mid X]. \end{aligned} \quad (5.4.2)$$

Furthermore, writing $Q = (X^T X + n\lambda I_d)^{-1}$, we have

$$\begin{aligned} \hat{\beta}_\lambda - \beta &= Q X^T (X\beta + \epsilon) - \beta \\ &= -n\lambda Q\beta + Q X^T \epsilon. \end{aligned}$$

Plugging this into (5.4.2), the cross term vanishes by independence and centering, and

$$\begin{aligned} r_\lambda(X) &= \sigma^2 + (n\lambda)^2 \mathbb{E}[\beta^T Q \Sigma Q \beta \mid X] + \mathbb{E}[\epsilon^T X Q \Sigma Q X^T \epsilon \mid X] \\ &= \sigma^2 + \frac{\lambda^2 n^2 \alpha^2}{d} \operatorname{Tr}(\Sigma Q^2) + \sigma^2 \operatorname{Tr}(\Sigma Q X^T X Q). \end{aligned}$$

Let $\hat{\Sigma} = X^T X/n$, $G = (\hat{\Sigma} + \lambda I_d)^{-1}$, and $\gamma_d = d/n$. Since $G \hat{\Sigma} G = G - \lambda G^2$,

$$r_\lambda(X) = \sigma^2 + \frac{\sigma^2 \gamma_d}{d} \operatorname{Tr}(\Sigma G) + \frac{\lambda(\lambda \alpha^2 - \sigma^2 \gamma_d)}{d} \operatorname{Tr}(\Sigma G^2).$$

Thus, to analyze the asymptotic risk, we should look at the two normalized traces

$$\frac{1}{d} \operatorname{Tr}(\Sigma(\hat{\Sigma} + \lambda I_d)^{-1}), \quad \frac{1}{d} \operatorname{Tr}(\Sigma(\hat{\Sigma} + \lambda I_d)^{-2}).$$

We define the Stieltjes transform of a probability measure H on $[0, \infty)$ by

$$m_H(z) = \int_0^\infty \frac{dH(l)}{l - z}, \quad z \in \mathbb{C} \setminus [0, \infty).$$

For the empirical spectral measure of $\hat{\Sigma}$,

$$m_{\hat{\Sigma}}(z) = \frac{1}{d} \operatorname{Tr}(\hat{\Sigma} - z I_d)^{-1}.$$

The companion transform is

$$v_n(z) = \frac{1}{n} \operatorname{Tr}(X X^T / n - z I_n)^{-1} = \gamma_d m_{\hat{\Sigma}}(z) - \frac{1 - \gamma_d}{z}.$$

If $m_{\hat{\Sigma}}(z) \rightarrow m(z)$, write $v(z) = \gamma m(z) - (1 - \gamma)/z$.

The main step in deriving the limiting risk is the following weighted-resolvent limit.

Theorem 5.4.1. *Under the assumptions above, for each $\lambda > 0$, almost surely,*

$$\frac{1}{d} \operatorname{Tr}(\Sigma(\hat{\Sigma} + \lambda I_d)^{-1}) \longrightarrow \frac{1}{\gamma} \left(\frac{1}{\lambda v(-\lambda)} - 1 \right) =: \kappa(\lambda).$$

The weighted trace is not determined by the empirical sample spectrum alone for arbitrary matrices. Its identification here uses the sample covariance structure and the Gaussian design. Before proving the theorem, we compute its consequence for the risk.

The normalized weighted resolvents are locally uniformly bounded and analytic away from $[0, \infty)$. The theorem and analytic convergence therefore also give

$$\frac{1}{d} \operatorname{Tr}(\Sigma(\hat{\Sigma} + \lambda I_d)^{-2}) \longrightarrow -\kappa'(\lambda).$$

Consequently, the limiting risk is $\sigma^2 + \sigma^2 \gamma \kappa(\lambda) - \lambda(\lambda \alpha^2 - \sigma^2 \gamma) \kappa'(\lambda)$. After simplification,

$$r_\lambda(X) \longrightarrow \frac{1}{\lambda v(-\lambda)} \left\{ \sigma^2 + \left(\frac{\lambda \alpha^2}{\gamma} - \sigma^2 \right) \left(1 - \frac{\lambda v'(-\lambda)}{v(-\lambda)} \right) \right\}. \quad (5.4.3)$$

Here v' is differentiation with respect to its argument, not with respect to λ .

We make the following observations about (5.4.3):

- (a) The optimal limiting ridge parameter is $\lambda^* = \sigma^2 \gamma / \alpha^2$, and

$$r_{\lambda^*}(X) \longrightarrow \frac{\sigma^2}{\lambda^* v(-\lambda^*)}.$$

In fact, the finite-sample conditional risk is minimized at $\lambda_d^* = \sigma^2 \gamma_d / \alpha^2$. To see this, diagonalize $\hat{\Sigma}$ with eigenvalues $s_j \geq 0$ and unit eigenvectors u_j , and put $c_j = u_j^T \Sigma u_j > 0$. Differentiating the exact risk gives

$$r'_\lambda(X) = 2 \left(\frac{\alpha^2 \lambda}{d} - \frac{\sigma^2}{n} \right) \sum_j \frac{c_j s_j}{(s_j + \lambda)^3}.$$

(b) When $\Sigma = I_d$, the Marchenko–Pastur theorem gives

$$r_\lambda(X) \longrightarrow \sigma^2 + \sigma^2 \gamma m_I(-\lambda; \gamma) + \lambda(\lambda \alpha^2 - \sigma^2 \gamma) m'_I(-\lambda; \gamma).$$

The formulas with irreducible noise risk 1 are the special case $\sigma^2 = 1$.

Now we prove Theorem 5.4.1. Let $G = (\lambda I_d + \hat{\Sigma})^{-1}$. We are interested in $\text{Tr}(G\Sigma)/d$ and $\text{Tr}(G)/d$. To this end, we invoke Stein's formula.

Theorem 5.4.2. *Let $X \sim N(0, \Sigma)$ and let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be continuously differentiable with f and its gradient of at most polynomial growth. For $i_0 \in [d]$,*

$$\mathbb{E}[X_{i_0} f(X)] = \sum_{k=1}^d \Sigma_{i_0 k} \mathbb{E}[\partial_k f(X)].$$

As a consequence, we have the following matrix-valued version.

Theorem 5.4.3. *For a continuously differentiable matrix-valued function F with the same integrability conditions,*

$$\mathbb{E}[X^T F(X) X] = \mathbb{E} \text{Tr}(\Sigma F(X)) + \mathbb{E} \sum_{k=1}^d (\Sigma(\partial_k F(X)) X)_k.$$

The identity follows by applying Stein's formula to $(F(X)X)_i$ and summing over i . It allows us to compute $\mathbb{E} \text{Tr}(G\hat{\Sigma})$. Regard the i -th row of the data matrix as the column vector $X(i) = x_i$. Conditionally on all other rows,

$$\begin{aligned} \mathbb{E} \text{Tr}(G\hat{\Sigma}) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X(i)^T G X(i)] \\ &= \mathbb{E} \text{Tr}(G\Sigma) + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^d \mathbb{E} \left[e_k^T \Sigma \left(\frac{\partial G}{\partial X(i)_k} \right) X(i) \right]. \end{aligned}$$

Differentiating an inverse matrix gives the crucial minus sign:

$$\frac{\partial G}{\partial X(i)_k} = -\frac{1}{n} G(e_k X(i)^T + X(i)e_k^T)G.$$

Substitution and summation therefore yield the exact identity

$$\mathbb{E} \text{Tr}(G\hat{\Sigma}) = \mathbb{E} \text{Tr}(G\Sigma) - \frac{1}{n} \mathbb{E}[\text{Tr}(\Sigma G) \text{Tr}(G\hat{\Sigma})] - \frac{1}{n} \mathbb{E} \text{Tr}(\Sigma G \hat{\Sigma} G).$$

Set $a_n = n^{-1} \text{Tr}(G\hat{\Sigma})$ and $b_n = n^{-1} \text{Tr}(G\Sigma)$. The final trace is $O(d)$ for fixed λ , since $\|G\hat{\Sigma}G\| \leq 1/(4\lambda)$ and $\|\Sigma\|$ is bounded. After division by n its contribution is $O(1/n)$.

For completeness, the required concentration follows directly from Gaussian concentration. Write $X = Z\Sigma^{1/2}$ with independent standard Gaussian entries in Z . Differentiation with respect to X gives

$$\nabla_X a_n = \frac{2\lambda}{n^2} X G^2, \quad \nabla_X b_n = -\frac{2}{n^2} X G \Sigma G.$$

The squared Frobenius norms of the corresponding gradients with respect to Z are $O(n^{-2})$: use $X^T X = n\hat{\Sigma}$, the bound on Σ , and the bounded scalar functions $s/(s + \lambda)^j$ for $j \geq 2$. Thus both functions have Lipschitz constant $O(n^{-1})$. Their fluctuations vanish almost surely by Gaussian concentration and Borel–Cantelli, and $\mathbb{E}[a_n b_n] - \mathbb{E}[a_n]\mathbb{E}[b_n] = O(n^{-2})$.

It follows that, almost surely,

$$a_n = b_n - a_n b_n + o(1), \quad b_n = \frac{a_n}{1 - a_n} + o(1).$$

There is no singular denominator: b_n is bounded and the first equation implies $(1 + b_n)(1 - a_n) = 1 + o(1)$. Finally,

$$a_n = \gamma_d(1 - \lambda m_{\hat{\Sigma}}(-\lambda)), \quad 1 - a_n = \lambda v_n(-\lambda).$$

Therefore

$$\begin{aligned} \frac{1}{d} \text{Tr}((\lambda I_d + \hat{\Sigma})^{-1} \Sigma) &= \frac{b_n}{\gamma_d} = \frac{1}{\gamma_d} \left(\frac{1}{\lambda v_n(-\lambda)} - 1 \right) + o(1) \\ &\rightarrow \frac{1}{\gamma} \left(\frac{1}{\lambda v(-\lambda)} - 1 \right), \end{aligned}$$

which proves the theorem.

5.5 Lecture 13: Mean Field Limit of Two-Layer Neural Networks

For a deep neural network $f(x; \theta)$, the loss function can be highly non-convex. However, when training neural networks with gradient descent, it is often observed that the loss goes to zero, i.e., the global optimum, (or close to zero) in practice. This intriguing phenomenon motivates us to study the training dynamic of neural networks.

There are different trends of studying neural networks theoretically. For example, the random feature model has been widely analyzed due to its simplicity. A random feature model is a two-layer neural network defined as

$$f(x; \theta) = a^\top \sigma(Wx)$$

where W and a are the weights associated with the first and second layer, respectively. In particular, W is randomly initialized but becomes frozen once initialized. In other words, only a is trainable in a random feature model.

The second popular theory trend is the neural tangent kernel (NTK). For full-batch gradient flow of the empirical loss $(2n)^{-1} \sum_j (f(x_j; \theta) - y_j)^2$, the exact chain rule gives

$$\partial_t f(x_i; \theta_t) = -\frac{1}{n} \sum_{j=1}^n k_t(x_i, x_j) (f(x_j; \theta_t) - y_j)$$

where kernel $k_t(x_i, x_j) = \langle \nabla_\theta f(x_i; \theta_t), \nabla_\theta f(x_j; \theta_t) \rangle$. Under the appropriate NTK parameterization, initialization, and regularity assumptions, the kernel variation tends to zero as width $N \rightarrow \infty$ on a fixed training-time interval. The rate is parameterization-dependent, and is not universally $1/N$. Thus, in this regime k_t is nearly constant. As a result, we can replace k_t by k_0 and use

$$\partial_t f(x_i; \theta_t) = -\frac{1}{n} \sum_{j=1}^n k_0(x_i, x_j) (f(x_j; \theta_t) - y_j).$$

The evolution obtained is kernel gradient flow. With an invertible training Gram matrix its long-time limit interpolates the training data, with the corresponding kernel predictor and initialization offset. SGD has additional sampling noise and is not exactly this deterministic ODE.

Below, we will focus on another perspective named as the mean-field limit of two-layer neural networks. Recall that a two-layer neural network can be written as

$$f(x; \theta) = \frac{1}{N} \sum_{i=1}^N a_i \tilde{\sigma}(\langle w_i, x \rangle)$$

where $\tilde{\sigma}$ is the nonlinear activation function and $\theta = \{\theta_i\}_{i=1}^N$ with $\theta_i = (a_i, w_i)$ are the parameters to be learned. We absorb (a_i, w_i) and $\tilde{\sigma}$ as $\sigma(x; \theta_i) \triangleq a_i \tilde{\sigma}(\langle w_i, x \rangle)$. We assume data (x, y) follows an unknown distribution and ℓ is the squared loss function, i.e., $\ell(f(x; \theta), y) = (f(x; \theta) - y)^2$. Ideally, to obtain good generalization performance, we hope to minimize the population risk $\mathbb{E}[\ell(f(x; \theta), y)] = \mathbb{E}[(f(x; \theta) - y)^2]$. We let $R_N(\theta)$ denote the population risk where subscript N indicates the number

of neurons. Note that

$$\begin{aligned}
R_N(\theta) &= \mathbb{E} \left[\left(\frac{1}{N} \sum_{i=1}^N \sigma(x; \theta_i) - y \right)^2 \right] \\
&= \mathbb{E}[y^2] - \frac{2}{N} \sum_{i=1}^N \mathbb{E}[\sigma(x; \theta_i)y] + \frac{1}{N^2} \sum_{i,j=1}^N \mathbb{E}[\sigma(x; \theta_i)\sigma(x; \theta_j)] \\
&= \mathbb{E}[y^2] + \frac{2}{N} \sum_{i=1}^N V(\theta_i) + \frac{1}{N^2} \sum_{i,j=1}^N U(\theta_i, \theta_j)
\end{aligned}$$

where $V(\theta_i) := -\mathbb{E}[\sigma(x; \theta_i)y]$ and $U(\theta_i, \theta_j) := \mathbb{E}[\sigma(x; \theta_i)\sigma(x; \theta_j)]$.

A very special property of two-layer neural network is the permutation invariance, namely, permuting $\{\theta_i\}_{i=1}^N$ does not change the output. Thus, its predictive performance only relates to set $\{\theta_i\}_{i=1}^N$. If we encode the parameters by the empirical probability measure $\hat{\rho} = N^{-1} \sum_i \delta_{\theta_i}$, we have $f(x; \theta) = \int \sigma(x; \vartheta) \hat{\rho}(d\vartheta)$. The empirical measure need not have a density. For a general probability measure ρ , define $f_\rho(x) = \int \sigma(x; \vartheta) \rho(d\vartheta)$. Similarly, we can write the population risk as

$$R(\rho) = \mathbb{E}[y^2] + 2 \int V(\theta) \rho(d\theta) + \iint U(\theta, \theta') \rho(d\theta) \rho(d\theta').$$

We note that parameter θ is learned by stochastic gradient descent (SGD) in practice

$$\theta(t+1) = \theta(t) - \gamma \nabla_\theta (f(x_t; \theta(t)) - y_t)^2$$

where (x_t, y_t) is the data sampled at the t -th iteration and $\gamma > 0$ is the learning rate. Considering the component θ_i , we have

$$\begin{aligned}
\theta_i(t+1) &= \theta_i(t) - 2\gamma \nabla_{\theta_i} f(x_t; \theta(t)) (f(x_t; \theta(t)) - y_t) \\
&= \theta_i(t) - \frac{2\gamma}{N} \nabla_{\theta_i} \sigma(x_t; \theta_i(t)) (f(x_t; \theta(t)) - y_t) \\
&= \theta_i(t) - 2\eta \nabla_{\theta_i} \sigma(x_t; \theta_i(t)) (f(x_t; \theta(t)) - y_t)
\end{aligned}$$

where we denote $\eta \triangleq \gamma/N$. Similarly, for the neural network $f(x; \theta) = N^{-1} \sum_{i=1}^N \sigma(x; \theta_i)$ with N neurons, we denote the empirical measure of parameters $\{\theta_i\}_{i=1}^N$ as $\hat{\rho}_t = \frac{1}{N} \sum_{i=1}^N \delta_{\theta_i(t)}$. Thus we have $f(x; \theta_t) = \int \sigma(x; \theta) \hat{\rho}_t(d\theta)$.

The first result requires regularity and initialization assumptions. Here is one sufficient setting on a fixed interval $[0, T]$. The initial parameters are i.i.d. with a compactly supported law ρ_0 , independently of the i.i.d. training samples. Assume that the sample update field

$$b(\theta, \rho; x, y) = -2\nabla_\theta \sigma(x; \theta) (f_\rho(x) - y)$$

is uniformly bounded and uniformly Lipschitz in (θ, ρ) , using the Euclidean distance and the Wasserstein distance W_1 , respectively. It is enough that these bounds hold on a common invariant bounded region containing all trajectories under consideration. Assume sufficient differentiability to justify the gradients of V, U and their interchange with expectation.

Theorem 5.5.1. *If $N \rightarrow \infty$, $\eta = \eta_N \rightarrow 0$, and $\eta_N \log^2 N \rightarrow 0$, then, in probability,*

$$\sup_{0 \leq t \leq T} W_1(\hat{\rho}_{\lfloor t/\eta \rfloor}, \rho_t) \rightarrow 0,$$

where ρ_t is the unique weak solution with initial law ρ_0 of

$$\begin{aligned} \partial_t \rho_t &= 2 \nabla_\theta \cdot (\rho_t \nabla_\theta \psi(\theta; \rho_t)), \\ \psi(\theta; \rho_t) &= V(\theta) + \int U(\theta, \theta') \rho_t(d\theta'). \end{aligned} \tag{5.5.1}$$

The condition on η_N is a convenient sufficient condition for the uniform-particle estimate below, not an optimal restriction.

Remark 5.5.2. 1. Theorem 5.5.1 gives a full characterization of the limit of the training dynamics of a two-layer neural network.

2. (5.5.1) essentially corresponds to the gradient descent of the population risk $R(\rho)$ in the measure space, with respect to the Wasserstein metric.

We provide some intuition for (5.5.1). Let the continuous particles start from the same initial parameters as SGD and solve

$$\frac{d}{ds} \bar{\theta}_i(s) = -2 \nabla \psi(\bar{\theta}_i(s); \bar{\rho}_s), \quad \bar{\rho}_s = \frac{1}{N} \sum_{j=1}^N \delta_{\bar{\theta}_j(s)}.$$

The relation $\bar{\theta}_i(s) \approx \theta_i(\lfloor s/\eta \rfloor)$ is an approximation, not a definition. The conditional mean of one SGD increment is exactly $-2\eta \nabla \psi(\theta_i; \hat{\rho})$, which explains the time scale. This is called mean-field dynamics because each particle interacts with the others through their empirical measure.

For a smooth compactly supported test function g , differentiation gives

$$\frac{d}{ds} \int g(\theta) \bar{\rho}_s(d\theta) = -2 \int \nabla g(\theta) \cdot \nabla \psi(\theta; \bar{\rho}_s) \bar{\rho}_s(d\theta).$$

This is the weak form of (5.5.1). The Lipschitz assumptions give existence, uniqueness, and stability of the characteristic flow. Coupling the particles with nonlinear characteristics having the same initial values, and applying Gronwall's inequality, gives

$$\sup_{s \leq T} W_1(\bar{\rho}_s, \rho_s) \leq e^{CT} W_1(\bar{\rho}_0, \rho_0).$$

The right-hand side converges to zero by the law of large numbers for the initial empirical measure. Combining this with the next proposition proves Theorem 5.5.1.

Based on the limit dynamics, we can derive the approximation for a neural network with N neurons.

Proposition 5.5.3. *Under the boundedness and Lipschitz assumptions above, for $0 < \delta < 1$ and $0 < \eta \leq 1$, there is a constant C_T independent of N, η, δ such that*

$$\max_{1 \leq i \leq N} \sup_{0 \leq s \leq T} \|\bar{\theta}_i(s) - \theta_i(\lfloor s/\eta \rfloor)\| \leq C_T \left(\sqrt{\eta \log(2N/\delta)} + \eta \right)$$

with probability at least $1 - \delta$. The constant may depend on the fixed parameter dimension and the regularity bounds.

Proof. Subtract the conditional mean from each SGD increment. For each neuron, the resulting martingale differences have norm bounded by $C\eta$. A maximal martingale concentration bound over the $O(T/\eta)$ steps, followed by a union bound over the N neurons, bounds all their partial sums by $C_T\sqrt{\eta\log(2N/\delta)}$. The Lipschitz drift and the Euler discretization error then give the stated estimate by the discrete Gronwall inequality. The neurons need not have independent martingale increments: the union bound only uses a bound for each neuron separately. Taking $\delta = e^{-c\log^2 N}$ gives a bound of order $C_T(\sqrt{\eta}\log N + \eta)$, as used above. $\square$

More general versions of Theorem 5.5.1 and Proposition 5.5.3 are developed in A mean field view of the landscape of two-layers neural networks (S Mei, A Montanari, P Nguyen) and Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit (S Mei, T Misiakiewicz, A Montanari).

The takeaway messages of the mean-field perspective are

1. Due to the permutation invariance, the measure of parameters is more essential to study in the training of two-layer neural networks than the parameters themselves.
2. The limit of training dynamics (with infinite width and infinitesimal learning rate) of two-layer neural networks is the gradient descent of the population risk in the measure space, with respect to the Wasserstein metric. This identifies the evolution, but does not by itself prove convergence to a global optimum. Indeed, when the differentiation is justified,

$$\frac{d}{dt}R(\rho_t) = -4 \int \|\nabla\psi(\theta; \rho_t)\|^2 \rho_t(d\theta) \leq 0.$$

Monotonicity of the risk gives dissipation, not global optimality of every stationary measure. Although R is convex under linear interpolation of probability measures, it need not be displacement convex in the Wasserstein geometry. Global convergence requires additional assumptions.

3. The approximation result for two-layer neural networks with finite width and learning rate can be established. This partially explains why the training of two-layer neural networks works surprisingly well in practice.