Chapter 4

Metric Entropy and Matrix Concentration

- Nets, covering numbers, and metric entropy
- Eigenvalues and Singular values of random matrices
- Matrix Concentration
- Operator norm vs. row/column norms
- Applications:
  - community detection,
  - (sparse) covariance matrix estimation
  - matrix completion

Let (T, ρ) be a metric space

Ex: $(\mathbb{R}^d, \|\cdot\|_2)$ with $\rho(x,y) = \|x-y\|_2$

- $(\{0,1\}^d, \rho_H)$ with Hamming metric

$$\rho_H(\theta, \tilde{\theta}) = \frac{1}{d} \sum_{j=1}^{d} \mathbb{1}[\theta_j \neq \tilde{\theta}_j]$$

- $L^2(\mu, [0,1])$ with

$$\|f-g\|_2 = \sqrt{\int_0^1 (f(x)-g(x))^2 \, d\mu(x)}$$

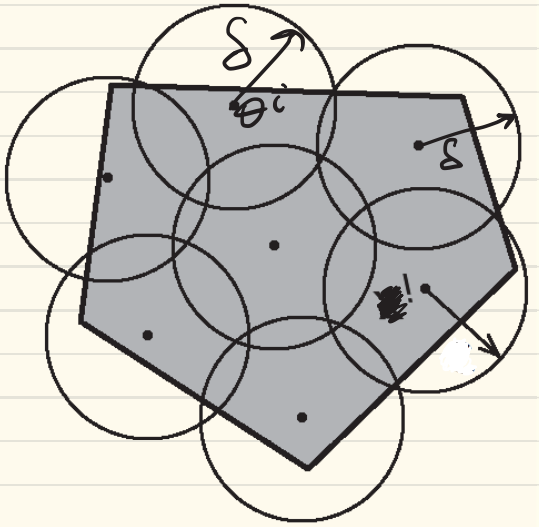
- $C[0,1]$ with

$$\|f-g\|_\infty = \sup_{x \in [0,1]} |f(x) - g(x)|$$

Defn: A δ-cover of a set T with respect to a metric ρ is a set $\{\theta^1, \ldots, \theta^N\} \subset T$ such that

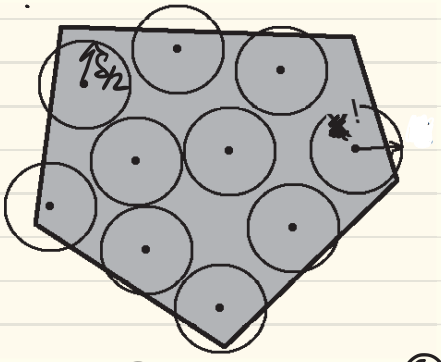
$\forall \theta \in T \; \exists i \in \{1, \ldots, N\}$ s.t. $\rho(\theta, \theta^i) \leq \delta$

The δ-covering number $N(\delta; T, \rho)$ is the cardinality of the smallest δ-cover. The quantity $\log(N(\delta; T, \rho))$ is called the metric entropy.



Defn: A δ-packing of T with respect to ρ is a set $\{\theta^1, \ldots, \theta^m\} \subseteq T$ such that $\rho(\theta^i, \theta^j) > \delta \quad \forall i,j \in \{1,2,\ldots,m\}$

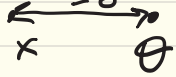
The δ-packing number $M(\delta; T, \rho)$ is the cardinality of the largest δ-packing.



Lemma:

$$M(2\delta; T, \rho) \overset{(a)}{\leq} N(\delta; T, \rho) \overset{(b)}{\leq} M(\delta; T, \rho)$$

Pf: (b) $M(\delta, T, \rho)$



$= M(\delta, T, \rho) \geq N(\delta, T, \rho)$

(a) $M(2\delta, T, \rho)$

No δ Ball can cover θ^i, θ^j simultaneously. ☆

Lemma: Let $\|\cdot\|$, $\|\cdot\|'$ be norms on $\mathbb{R}^d$ and set

$$\mathbb{B} = \{x : \|x\| \le 1\}$$
$$\mathbb{B}' = \{x : \|x\|' \le 1\}$$

Then

$$\left(\frac{1}{\delta}\right)^d \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')} \overset{(a)}{\le} N(\delta; \mathbb{B}, \|\cdot\|') \overset{(b)}{\le} \frac{\mathrm{vol}\left(\frac{2}{\delta}\mathbb{B} + \mathbb{B}'\right)}{\mathrm{vol}(\mathbb{B}')}$$

Remark: If $\mathbb{B}' \subseteq \mathbb{B}$, then simplifies

$$N(\delta; \mathbb{B}, \|\cdot\|') \le \left(1 + \frac{2}{\delta}\right)^d \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')}$$

pf of lemma: If $\{\theta^1, \ldots, \theta^n\}$ is a δ-covering of $\mathbb{B}$, then $\mathbb{B} \subseteq \bigcup_{i=1}^n \{\theta^i + \delta\mathbb{B}'\}$

$\Rightarrow \mathrm{vol}(\mathbb{B}) \le N\delta^d \mathrm{vol}(\mathbb{B}') \Rightarrow$ (a).

Let $\{\theta^1, \ldots, \theta^m\}$ be a maximal $\delta/2$-packing of $\mathbb{B}$. $\Rightarrow$ must also be δ-covering.

Taking into account:

$$\{\theta_j + \tfrac{\delta}{2}B'\} \subseteq B + \tfrac{\delta}{2}B'$$

are disjoint

get:

$$M \operatorname{vol}\left(\tfrac{\delta}{2}B'\right) \le \operatorname{vol}\left(B + \tfrac{\delta}{2}B'\right)$$

$$\left(\tfrac{\delta}{2}\right)^d M \operatorname{vol}(B') \quad \left(\tfrac{\delta}{2}\right)^d \operatorname{vol}\left(\tfrac{2}{\delta}B + B'\right)$$

∎

Cor: If $\|\cdot\|' = \|\cdot\|$, then

$$d \log\left(\tfrac{1}{\delta}\right) \le \log N(\delta; B, \|\cdot\|) \le d \log\left(1 + \tfrac{2}{\delta}\right)$$

Metric entropy can be very large...

Ex: Define

$$\mathcal{F}_L([0,1]^d) = \{f : [0,1]^d \to \mathbb{R} : f(0) = 0,\ |f(x) - f(x')| \le L\|x - x'\|_\infty \ \forall x, x' \in [0,1]^d\}$$

Then $\log N(\delta; \mathcal{F}_L([0,1]^d, \|\cdot\|_\infty) \asymp \left(\tfrac{L}{\delta}\right)^d$

See ex 5.10 in Wainwright.

One of the main uses of covering numbers is for random matrix theory.

Review: Any $A \in \mathbb{R}^{m \times n}$ can be written as

$$A = \sum_{i=1}^{r} s_i u_i v_i^T = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix} \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_r \end{bmatrix} \begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_r & - \end{bmatrix}$$

where $r = \text{rank}(A)$, $s_i \geq 0$, $\|u_i\|_2 = \|v_i\| = 1$

- Decomposition is called a <u>singular value decomposition</u>.
- s_i are the <u>singular values</u>. We define $s_{r+1}, \ldots, s_n = 0$. and arrange

$$s_1 \geq s_2 \geq s_3 \geq \cdots \geq s_n \geq 0$$

- $s_i(A) = \sqrt{\lambda_i(AA^T)} = \sqrt{\lambda_i(A^TA)} \quad \forall i = 1, \ldots, r.$

The set $\mathbb{R}^{m \times n}$ admits the inner-product

$$\langle A, B \rangle := \sum_{i,j} A_{ij} B_{ij} = \text{tr}(A^T B)$$

and the norms

$$\|A\|_F := \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\langle A, A \rangle} = \|s(A)\|_2$$

$$\|A\|_2 = \sup_{\|x\| \leq 1} \|Ax\|_2 = s_1(A) = \|s(A)\|_\infty$$

Remark: $s_1(A)$ and $s_n(A)$ are respectively the smallest $\bar{m}$ and largest $\underline{m}$ with

$$\underline{m}\|x\|_2 \leq \|Ax\|_2 \leq \bar{m}\|x\|_2 \quad \forall x \in \mathbb{R}^n$$

Thm (Eckart-Young-Mirsky)
Let $A = \sum_{i=1}^{n} s_i u_i v_i^T$ be an SVD of A
Then the matrix $A = \sum_{i=1}^{r} s_i u_i v_i^T$
solves the problem

$$\min_{A' : \operatorname{rank}(A') \leq r} \|A - A'\|$$

where $\|\cdot\|$ is either $\|\cdot\|_F$ or $\|\cdot\|_2$.

We next aim to show that for $A \in \mathbb{R}^{m \times n}$ with independent subGaussian entries, it holds:

$$\|A\|_2 \lesssim \sqrt{m} + \sqrt{n}$$

with high probability.

Lemma: Let $A \in \mathbb{R}^{m \times n}$ and $\varepsilon \in [0, \sqrt{2}-1)$. Then for any ε-nets $\mathcal{N}$ of S^{n-1} and $\mathcal{M}$ of S^{m-1}, it holds:

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \overset{@}{\leq} \|A\|_2 \overset{ⓑ}{\leq} \frac{1}{1-2\varepsilon-\varepsilon^2} \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle$$

pf: @ is clear. To see ⓑ, suppose

$\|A\|_2 = \langle Ax, y \rangle$ for some $x \in S^{n-1}$, $y \in S^{m-1}$.

$\exists \hat{x} \in \mathcal{N}$, $\hat{y} \in \mathcal{M}$ with $\|x - \hat{x}\| \leq \varepsilon$, $\|y - \hat{y}\| \leq \varepsilon$

$$\Rightarrow \langle Ax, y \rangle = \langle A\hat{x} - A(\hat{x}-x), \hat{y} - (\hat{y}-y) \rangle$$

$$= \langle A\hat{x}, \hat{y} \rangle - \langle A\hat{x}, \hat{y}-y \rangle - \langle A(\hat{x}-x), \hat{y} \rangle + \langle A(\hat{x}-x), \hat{y}-y \rangle$$

$$\leq \langle A\hat{x}, \hat{y} \rangle + 2\varepsilon\|A\|_2 + \|A\|_2 \varepsilon^2$$

$$\Rightarrow \|A\|_2 \leq \frac{1}{1-2\varepsilon-\varepsilon^2} \langle A\hat{x}, \hat{y} \rangle \quad \blacksquare$$

Thm: Let $A \in \mathbb{R}^{m \times n}$, where A_{ij} are independent mean-zero σ_{ij}-subGaussian. Then $\exists$ constant $C > 0$ s.t.

$$\mathbb{P}\left[\|A\|_2 \leq C \cdot \left(\max_{i,j} \sigma_{ij}\right) \cdot \left(\sqrt{m} + \sqrt{n} + t\right)\right] \geq 1 - 2e^{-t^2}$$

[**Remark:** when $A_{ij} \sim N(0,1)$, optimal constant is $C = 1$]

pf: **Step 1**: Approximation

We know $N(S^{n-1}, \varepsilon, \|\cdot\|_2) \leq \left(1 + \frac{2}{\varepsilon}\right)^n$

$N(S^{m-1}, \varepsilon, \|\cdot\|_2) \leq \left(1 + \frac{2}{\varepsilon}\right)^m$

$\Rightarrow$ Set $\varepsilon = \frac{1}{4}$. There are nets

$\mathrm{card}(\mathcal{N}) \leq 9^n$, $\mathrm{card}(\mathcal{M}) \leq 9^m$

$\overset{\text{Lemma}}{\Rightarrow} \|A\|_2 \leq 3 \cdot \max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle$

Step 2: Concentration

The RV $\langle Ax, y\rangle = \sum_{i,j} A_{ij} x_i y_j$ satisfies

$$e^{\lambda\langle Ax,y\rangle} = \prod_{i,j} e^{(\lambda x_i y_j) A_{ij}} \le \prod_{i,j} e^{\lambda^2 \sigma_{ij}^2 x_i^2 y_j^2}$$

$$= e^{\lambda^2 \left(\max_{ij} \sigma_{ij}\right)^2 \sum_{i,j} x_i^2 y_j^2} = e^{\lambda^2 \max_{i,j} \sigma_{ij}^2}$$

So Hoeffding $\Rightarrow$

$$\mathbb{P}\left[\langle Ax, y\rangle \ge u\right] \le e^{\frac{-u^2}{2\max \sigma_{ij}^2}} \quad \forall u > 0$$

$$\Rightarrow \mathbb{P}\left[\max_{\substack{x \in \mathcal{N} \\ y \in \mathcal{M}}} \langle Ax, y\rangle \ge u\right] \le 9^{n+m} e^{\frac{-u^2}{2\max_{ij} \sigma_{ij}^2}}$$

Choose $u = C\left(\max_{ij} \sigma_{ij}\right)\left(\sqrt{n} + \sqrt{m} + t\right)$

Then

$$RHS \le 9^{n+m} \cdot e^{-C(n+m+t^2)}$$

$$= e^{\ln(9)(n+m) - C(n+m+t^2)} \le e^{-t^2}$$

by choosing C large ∎

Cor: (HW) $\mathbb{E}\|A\|_2 \le C(\max_{ij} \sigma_{ij})(\sqrt{m} + \sqrt{n})$

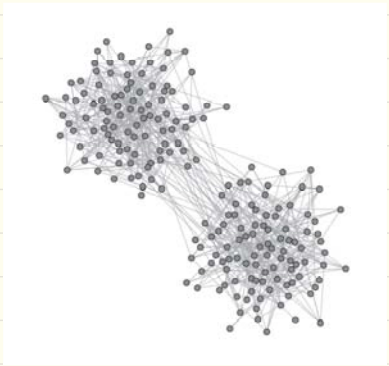
Obvious analogues hold for symmetric matrices.

Application: Community Detection

Defn: (Stochastic Block Model)

Divide n vertices into two sets Q_1, Q_2 of size $n/2$ each. Build the graph

$$\mathbb{P}[ij \text{ is an edge}] = \begin{cases} p & \text{if } i,j \in Q_1 \text{ or } i,j \in Q_2 \\ q & \text{o.w} \end{cases}$$



The goal is to find the communities from the realization of the graph.

Let A be the adjacency matrix.
Write it as

$$A = D + R \quad \text{where}$$

$$D = \mathbb{E}A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right]$$

Exercise: (HW)
Show that D has rank 2 and

$$\lambda_1 = \left(\frac{p+q}{2}\right)n, \quad u_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} / \sqrt{n}$$

$$\lambda_2 = \left(\frac{p-q}{2}\right)n, \quad u_2 = \begin{bmatrix} 1 \\ \vdots \\ -1 \end{bmatrix} / \sqrt{n}$$

If we knew u_2, we would be able to identify the communities.

Signal to noise ratio (SNR): $\frac{\|D\|_2}{\|R\|_2}$

We know $\|D\|_2 = \lambda_1 \sim n$

$$\mathbb{P}\left[\|R\| \leq C\sqrt{n}\right] \geq 1 - 4e^{-n}$$

So $SNR \sim \sqrt{n}$

$\Rightarrow$ problem should be solvable.

A natural idea is to use spectral clustering.

Alg: (Spectral Clustering)

Input: graph G

1) Compute adjacency matrix A

2) Compute eigenvector $V_2(A)$ corresponding to second largest eigenvalue of A.

3) Set $Q_1 = \{i : [V_2(A)]_i > 0\}$

$Q_2' = \{i : [V_2(A)]_i \leq 0\}$.

Perturbation Results:

Thm: (Weyl's inequality)
For any symmetric $X, Y \in \mathbb{R}^{n\times n}$, it holds

$$\max_{i=1,\dots,n} |\lambda_i(X) - \lambda_i(Y)| \le \|X - Y\|_2$$

Thm (Davis-Kahan)
Let $X, Y \in \mathbb{R}^{n\times n}$ be symmetric. Fix an index i and assume

$$\min_{j : j \neq i} |\lambda_i(X) - \lambda_j(X)| = \delta > 0$$

Then

$$\sin \angle (v_i(X), v_i(Y)) \le \frac{2\|X - Y\|_2}{\delta}$$

Remark: This implies

$$\min_{\theta \in \{\pm 1\}} \|v_i(X) - \theta v_i(\theta)\|_2 \le \frac{2^{3/2}\|X - Y\|_2}{\delta}$$

Let's compute the eigengap

$$\delta = \min\{\lambda_2, \lambda_1 - \lambda_2\} = \min\{\tfrac{p-q}{2}, q\} \cdot n$$

$$= \mu \cdot n$$

$$\Rightarrow \min_{\theta \in \{\pm 1\}} \|V_2(D) - \theta V_2(A)\|_2 \le \frac{\|D-A\|_2}{\mu n} = \frac{\|R\|_2}{\mu n}$$

$$\Rightarrow \mathbb{P}\left[\min_{\theta \in \{\pm 1\}} \|V_2(D) - \theta V_2(A)\|_2 \le \frac{c}{\mu\sqrt{n}}\right] \ge 1 - 4e^{-n}.$$

Since all coordinates of $\theta V_2(D)$ are $\pm \frac{1}{\sqrt{n}}$, the signs can only disagree in at most $\frac{c^2}{\mu^2}$ coordinates.

Thm: Set $\mu := \min\{q, \frac{p-q}{2}\}$. Then with probability $1 - 4e^{-n}$, the spectral clustering algorithm identifies the communities correctly up to $\left(\frac{c}{\mu}\right)^2$ misclassified vertices.

We next obtain two-sided bounds on the full spectrum

$$\sqrt{m} - C\sqrt{n} \le s_i(A) \le \sqrt{m} + C\sqrt{n}$$

and we relax independence of entries to independence of the rows.

Thm: Let $A \in \mathbb{R}^{m \times n}$ be a matrix with rows A_i that are independent and σ_i-sub-Gaussian isotropic. Then

$$\sqrt{m} - CK^2(\sqrt{n}+t) \le s_n(A) \le s_1(A) \le \sqrt{m} + CK^2(\sqrt{n}+t)$$

w.p. $1 - 2\exp(-t^2)$, where $K = \max_i \sigma_i$

pf: We first show

$$\left\| \frac{1}{m} A^T A - I_n \right\|_2 \le K^2 \max\{\delta, \delta^2\} \text{ where}$$

$$\delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right)$$

Step 1: As before, the exist a $\frac{1}{4}$-cover $\mathcal{N}$ of S^{n-1} with $\text{card}(\mathcal{N}) \le 9^n$.

For $B = \frac{1}{m} A^T A - I$ have $\|B\|_2 = \sup_{v \in S^{n-1}} \langle Bv, v \rangle$

If v achieves sup. Then $\exists w$ with $\|w - v\| \le \frac{1}{4}$ and

$$\|B\|_2 = \langle Bw - B(w-v),\ w - (w-v) \rangle$$

$$= \langle Bw, w \rangle - 2\langle Bw, w-v \rangle + \langle B(w-v), w-v \rangle$$

$$\le \sup_{w \in \mathcal{N}} \langle Bw, w \rangle + \frac{1}{4}\|B\|_2 + \frac{1}{16}\|B\|$$

$$\Rightarrow \|B\|_2 \le 2 \sup_{w \in \mathcal{N}} \langle Bw, w \rangle$$

$$= 2\left(\sup_{w \in \mathcal{N}} \frac{1}{m}\|Aw\|_2^2 - 1 \right)$$

Step 2: Concentration

Fix $w \in \mathcal{N}$ and write

$$\|Aw\|_2^2 = \sum_{i=1}^m \langle A_i, w \rangle^2 =: \sum_{i=1}^m X_i^2$$

$\Rightarrow X_i$ are independent, 6_i- subGaussian with $\mathbb{E} X_i^2 = 1$. Therefore, Bernstein

$$\mathbb{P}\left[\left| \frac{1}{m}\|Aw\|_2^2 - 1 \right| \ge t \right] \le 2\exp\left[-\frac{1}{2}\left(\frac{mt^2}{K^4} \wedge \frac{mt}{4K^2} \right) \right]$$

$$2\exp\left[-\frac{1}{2}\left(\frac{mt^2}{K^4} \wedge \frac{mt}{4K^2}\right)\right] = 2\exp\left[-c_1 m\left[\frac{t^2}{K^4} \wedge \frac{t}{K^2}\right]\right.$$

Set $t = K^2 \max\{\delta, \delta^2\}$

$$= 2\exp\{-c_1 m\delta^2\}$$

$\leq$

Step 3: Union bound

$$\mathbb{P}\left[\max_{w \in N} \left|\frac{1}{m}\|Aw\|_2^2 - 1\right| \geq K^2 \max\{\delta, \delta^2\}\right]$$

$$\leq 9^n \cdot 2\exp\{-c_1 C^2 (n + t^2)\} \leq 2e^{-t^2}$$

if C is big enough.

Finally result follows from

Lemma: Suppose $A \in \mathbb{R}^{m \times n}$ and $\delta > 0$ satisfy

$$\|A^T A - I\|_2 \leq \max\{\delta, \delta^2\}$$

Then $1 - \delta \leq s_n(A) \leq s_1(A) \leq 1 + \delta$ ∎

Cor: (HW) $\mathbb{E}\left\| \frac{1}{m} A^T A - I \right\|_2 \le CK^2 \left(\sqrt{\frac{n}{m}} + \frac{n}{m} \right)$

Covariance Estimation

Assume $X \in \mathbb{R}^n$ and $\mathbb{E}X = 0$

Define $\Sigma = \mathbb{E}XX^T$. Let's estimate Σ by the sample covariance

$$\Sigma_m := \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T$$

Clearly for $\Sigma - \Sigma_m$ to be small we need $m > n$. We'll see $m \sim n$ suffice

Thm: Suppose $X \in \mathbb{R}^n$ satisfies

$$\mathbb{E}_x e^{\lambda \langle X, w \rangle} \le e^{K^2 \lambda^2 \langle \Sigma w, w \rangle} \quad \forall w \in \mathbb{R}^n$$

for some $K > 0$.

Then

$$\mathbb{E} \frac{\|\Sigma_m - \Sigma\|_2}{\|\Sigma\|_2} \le CK^2 \left(\sqrt{\frac{n}{m}} + \frac{n}{m} \right)$$

pf. Fix random $X, X_1, \dots, X_m$. Define

$$Z = \Sigma^{-1/2} X \qquad Z_i := \Sigma^{-1/2} X_i$$

Recall Z, Z_i are independent and isotropic. Moreover for any $w \in S^{n-1}$, have

$$\mathbb{E} e^{\lambda \langle Z, w \rangle} = \mathbb{E} e^{\lambda \langle \Sigma^{-1/2} X, w \rangle}$$
$$= \mathbb{E} e^{\lambda \langle X, \Sigma^{-1/2} w \rangle} \le e^{\frac{1}{2}\lambda^2 K^2 \langle \Sigma (\Sigma^{-1/2} w), \Sigma^{-1/2} w \rangle}$$
$$= e^{\frac{1}{2}\lambda^2 K^2}$$

So Z and Z_i are K-sub-Gaussian

$$\Rightarrow \|\Sigma_m - \Sigma\|_2 = \left\| \Sigma^{1/2} \left(\frac{1}{m} \sum_{i=1}^m Z_i Z_i^T - I_n \right) \Sigma^{1/2} \right\|_2$$
$$\le \|\Sigma\|_2 \cdot \left\| \frac{1}{m} \sum_{i=1}^n Z_i Z_i^T - I \right\|_2$$

Define A whose rows are Z_i. Then

$$\frac{1}{m} A^T A - I = \frac{1}{m} \sum_{i=1}^m Z_i Z_i^T - I.$$

$\square$

So can get

$$\mathbb{E}\,\|\Sigma_m - \Sigma\|_2 \le \varepsilon\,\|\Sigma\|_2$$

as long as $m \sim \frac{n}{\varepsilon^2}$

Exc: (HW) Actually get the high probability estimate

$$\|\Sigma_m - \Sigma\|_2 \le CK^2\left(\sqrt{\frac{n+t}{m}} + \frac{n+t}{m}\right)\|\Sigma\|_2$$

with probability $1 - 2e^{-t}$

Remark: In particular, setting $t = \delta m$ get

$$\frac{\|\Sigma_m - \Sigma\|_2}{\|\Sigma\|_2} \le CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m} + \delta\right)$$

with probability $1 - 2e^{-\delta m}$

Clustering (Gaussian Mixture Model)

Generate m random points in $\mathbb{R}^n$ as follows.

Flip a fair coin

- if heads choose $x_i \sim N(\mu, I)$
- if tails choose $x_i \sim N(-\mu, I)$

Suppose we are given m points drawn from the Gaussian Mixture Model.

Goal: Identify which points belong to which cluster.

Spectral Clustering:

Input: $x_1, \ldots, x_m \in \mathbb{R}^n$

Output: a partition of points into 2 clusters.

1) Compute $\Sigma_m = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T$
2) Compute eigenvector of Σ_m cossesponding to $\lambda_1(\Sigma)$.
3) If $\langle v, x_i \rangle > 0$, put x_i in first community otherwise, put x_i in second.

Thm: Let $\delta > 0$ be s.t. $\|\mu\|_2 \geq C\sqrt{\log\left(\frac{1}{\delta}\right)}$.
Suppose $m \geq \left(\frac{n}{\|\mu\|_2}\right)^c$ where c is a constant. Then with probability $1 - 4e^{-n}$, the algorithm identifies communities correctly up to δm misclassified vertices.

pf sketch: Note $X = \varepsilon\mu + g$ where ε is a Rademacher RV and $g \sim N(0, I)$

$\Rightarrow \mathbb{E}\, XX^T = \mathbb{E}(\mu\mu^T + gg^T) = I + \mu\mu^T$

$\Rightarrow \mathbb{E}\, XX^T$ has eigenvalues $1 + \|\mu\|^2, 1, \ldots, 1$

Eigenvector corresponding to $1 + \|\mu\|^2$ is μ

So from the e-vector corresponding to the largest eigenvalue, you can "guess" at the communities by checking the sign of $\langle x_i, \mu \rangle$.

You do the rest in your HW ☺

Our next goal is to investigate tails of matrices generalizing bounds

$$\left\| \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T - \Sigma \right\|_2$$ from before.

Notation:

- S^d are $d \times d$ symmetric matrices
- $S_+^d = \{ Q \in S^d : \langle Qx, x \rangle \geq 0 \ \forall x \in \mathbb{R}^d \}$
- $\gamma_1(X) \geq \gamma_2(X) \geq \dots \geq \gamma_d(X)$ are the eigenvalues of $X \in S^d$
- $O^d = \{ U \in \mathbb{R}^{d \times d} : U^T U = I \}$.

Any function $f : \mathbb{R} \to \mathbb{R}$ defines a function $f : S^d \to S^d$ by

$$f(Q) = U \operatorname{diag}(f(\gamma_1(Q)), \dots, f(\gamma_d(Q))) U^T$$

where $Q = U \operatorname{diag}(\gamma(Q)) U^T$ is any eigenvalue decomposition of Q.

[The choice of U does not matter (check!)]

Note $\gamma(f(Q)) = \{f(\gamma_j(Q)) : j = 1, \ldots, d\}$

Matrix exponential: e^Q

Matrix Logarithm: $\log Q$ defined on $\mathbb{S}^d_{++}$

The Löwner ordering on

$X \succeq Y \iff Y - X \in \mathbb{S}^d_+$

Lemma: (HW) $Q_2 \succeq Q_1 > 0 \Rightarrow \log(Q_2) \geq \log(Q_1)$

[e^Q is not monotone]

Lemma: If $f: \mathbb{R} \to \mathbb{R}$ is continuous and non-decreasing, then

$$Q \succeq R \Rightarrow \operatorname{tr}(f(Q)) \leq \operatorname{tr}(f(R))$$

Lemma: $I + A \preceq e^A$

Thm (Lieb): For any $H \in \mathbb{S}^d$, the function $f: \mathbb{S}^d_{++} \to \mathbb{R}$ given by $f(A) = \operatorname{tr}(e^{H + \log(A)})$ is concave.

Defn: Moment generating function of a random matrix $Q \in \mathbb{S}^d$ is the function $\psi_Q : \mathbb{R} \to \mathbb{S}^d$ given by

$$\psi_Q(\lambda) = \mathbb{E} e^{\lambda Q} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[Q^k]$$

Defn: A zero-mean $Q \in \mathbb{S}^d$ is V-subGaussian (with $V \in \mathbb{S}^d_+$) if

$$\psi_Q(\lambda) \preceq e^{\frac{\lambda^2 V}{2}} \quad \forall \lambda \in \mathbb{R}$$

Ex: Suppose $Q = \varepsilon A$ where ε is a Rademacher RV and $A \in \mathbb{S}^d$ is fixed.

$$\mathbb{E} e^{\lambda Q} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}(\varepsilon A)^k = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} A^{2k}$$

$$\preceq \sum_{k=1}^{\infty} \frac{1}{k!} \left(\frac{\lambda^2 A^2}{2} \right)^k$$

$$= e^{\frac{\lambda^2 A^2}{2}}$$

So Q is A^2-subGaussian.

Exercise (HW)

If $Q = gA$, where $g \in \mathbb{R}$ is a symmetric distribution σ-subGaussian variable, then Q is $\sigma^2 A^2$ subGaussian.

Ex: Suppose $Q = \varepsilon A$ where ε is Rademacher and A is random with $\|A\|_2 \le b$. Then

$$\psi_Q(\lambda) = \mathbb{E} e^{\lambda Q} = \mathbb{E}_A \mathbb{E}_\varepsilon e^{\varepsilon A} \overset{\text{(why?)}}{\le} \mathbb{E}_A e^{\frac{\lambda^2 A^2}{2}} \le e^{\frac{\lambda^2 b^2 I}{2}}$$

So Q is $b^2 I$-subGaussian.

Defn: A zero-mean random matrix Q is subexponential with parameters (ν, α) if

$$\psi_Q(\lambda) \le e^{\frac{\lambda^2 \nu}{2}} \quad \forall |\lambda| \le \frac{1}{\alpha}.$$

Ex: $Q = \varepsilon g^2 A$ where ε is Rademacher and $g \sim N(0,1)$ independent.

[You'll compute V and α in HW]

Lemma: Suppose zero-mean $Q \in S^d$ satisfies $\|Q\|_2 \le b$. Then we have

$$\psi_Q(\lambda) \lesssim \exp\left(\frac{\lambda^2 \mathrm{Var}(Q)}{2(1-b|\lambda|)}\right) \quad \forall |\lambda| \le \frac{1}{b}$$

pf: $\mathbb{E} e^{\lambda Q} = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E} Q^k}{k!} =$

$$= I + \frac{\lambda^2 \mathrm{Var}(Q)}{2} + \sum_{j=3}^{\infty} \frac{\lambda^j \mathbb{E} Q^j}{j!}$$

$$\lesssim I + \frac{\lambda^2 \mathrm{Var}(Q)}{2} + \sum_{j=3}^{\infty} \frac{\lambda^2 \lambda^{j-2} \mathbb{E} Q^2 b^{j-2}}{j!}$$

$$\lesssim I + \frac{\lambda^2 \mathrm{Var}(Q)}{2} + \frac{\lambda^2 \mathrm{var}(Q)}{2} \sum_{j=3}^{\infty} \lambda^{j-2} b^{j-2}$$

$$= I + \frac{\lambda^2 \mathrm{var}(Q)}{2(1-b|\lambda|)} \lesssim \exp\left(\frac{\lambda^2 \mathrm{var}(Q)}{2(1-b|\lambda|)}\right) \blacksquare$$

Matrix Chernoff:

Lemma: Let $Q \in \mathbb{S}^d$ be zero mean with ψ_Q existing on $(-a, a)$. Then for any $t > 0$, it holds:

$$\mathbb{P}[\gamma_1(Q) \geq t] \leq \operatorname{tr}(\psi_Q(\lambda)) e^{-\lambda t} \quad \forall_{\lambda \in [0,a)}$$

Consequently

$$\mathbb{P}[\|Q\|_2 \geq t] \leq 2 \operatorname{tr}(\psi_Q(\lambda)) e^{-\lambda t} \quad \forall_{\lambda \in [0,a]}$$

pf: We begin as in scalar case:

$$\begin{aligned}
\mathbb{P}[\gamma_1(Q) \geq t] &= \mathbb{P}[e^{\gamma_1(\lambda Q)} \geq e^{\lambda t}] \\
&= \mathbb{P}[\gamma_1(e^{\lambda Q}) \geq e^{\lambda t}] \\
&\leq \mathbb{E}[\gamma_1(e^{\lambda Q})] e^{-\lambda t} \\
&\leq \mathbb{E}[\operatorname{tr} e^{\lambda Q}] e^{-\lambda t} \\
&\leq \operatorname{tr}[\psi_Q(\lambda)] e^{-\lambda t} \quad \square
\end{aligned}$$

Lemma: Let $Q_1, \dots, Q_n \in \mathbb{S}^n$ be independent with $\Psi_{Q_i}(\lambda)$ existing for all λ in an interval J. Define $S_n = \sum_{i=1}^n Q_i$. Then

$$\operatorname{tr}\left(\Psi_{S_n}(\lambda)\right) \le \operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right) \quad \forall \lambda \in J$$

[**Note:** together with the previous lemma, we deduce

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \ge t\right] \le 2\operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right) \cdot e^{-\lambda t n}$$

]

pf: Define $G(\lambda) := \operatorname{tr}\left(\Psi_{S_n}(\lambda)\right)$. Then

$$G(\lambda) = \operatorname{tr}\left(\mathbb{E}\, e^{\lambda S_{n-1} + \log \exp(\lambda Q_n)}\right)$$

$$= \mathbb{E}_{S_{n-1}} \mathbb{E}_{Q_n} \operatorname{tr}\left(e^{\lambda S_{n-1} + \log \exp(\lambda Q_n)}\right)$$

Lieb $\le \mathbb{E}_{S_{n-1}} \operatorname{tr}\left(e^{\lambda S_{n-1} + \log \Psi_{Q_n}(\lambda)}\right)$

$\le \dots \le \operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right)$ ∎

Thm (Hoeffding)
Let $Q_1, \dots, Q_n \in \mathbb{S}^d$ be independent, zero-mean, and V_i-subGaussian.
Then

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \geq t\right] \leq 2\,\mathrm{rank}\left(\sum_{i=1}^n V_i\right) e^{\frac{-nt^2}{2\sigma^2}}$$

$$\leq 2d\, e^{\frac{-nt}{2\sigma^2}} \qquad \forall t \geq 0,$$

where $\sigma^2 = \left\|\frac{1}{n}\sum_{i=1}^n V_i\right\|_2$.

pf: Define $V = \sum_{i=1}^n V_i$. Suppose first $\mathrm{rank}\, V = d$. Then

$$\sum_{i=1}^n \log \Psi_{Q_i}(\lambda) \preceq \frac{\lambda^2}{2}\sum_{i=1}^n V_i$$

because $\log(\cdot)$ is matrix monotone.
The function $t \mapsto e^t$ is increasing

$$\Rightarrow \mathrm{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right) \leq \mathrm{tr}\left(e^{\frac{\lambda^2}{2}\sum_{i=1}^n V_i}\right)$$

Chernoff

$$\Rightarrow \mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_i\right\|_2 \geq t\right] \leq 2\,\mathrm{tr}\left(e^{\frac{\lambda^2}{2}V}\right)e^{-\lambda n t}$$

Note $\mathrm{tr}\left(e^{A}\right) \leq d\,e^{\|A\|_2}$ for any $A \in \mathbb{R}^d$

$$\Rightarrow \leq 2d\,e^{\frac{\lambda^2}{2}n\sigma^2 - \lambda n t}$$

Optimize over $\lambda \Rightarrow \lambda = \frac{t}{\sigma^2}$. Done.

If $r := \mathrm{rank}(V) < d$, we can form an eigenvalue decomposition

$V = UDU^T$ where $U \in \mathbb{R}^{d \times r}$

Then replace Q by

$\widehat{Q} = U^T Q U \in \mathbb{R}^{r \times n}$. ∎

Thm: Let $Q_1, \dots, Q_n \in \mathbb{S}^d$ be zero-mean, independent, with $\|Q_i\|_2 \le b$. Then

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \ge t\right] \le 2\,\mathrm{rank}(V)\exp\left(-\frac{nt^2}{2\sigma^2 + bt}\right)$$

$$\le 2\,\mathrm{rank}(V)\exp\left(-\frac{nt^2}{4\sigma^2} \wedge \frac{nt}{2b}\right)$$

where $V = \sum_{i=1}^n \mathbb{E}Q_i^2$ and $\sigma^2 = \frac{1}{n}\|V\|_2$. (HW)

Cor: Under same assumptions,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \le \sqrt{\frac{\sigma^2 \log(d)}{n}} + \frac{b\log(d)}{n}$$

The Bernstein bound can be extended to rectangular matrices.

Idea: Given a sequence $A_1, \dots, A_n \in \mathbb{R}^{d_1 \times d_2}$ form $Q_i := \begin{bmatrix} 0_{d_1 \times d_2} & A_i \\ A_i^T & 0_{d_1 \times d_2} \end{bmatrix}$

Lemma: (HW) $\|Q_i\|_2 = \|A_i\|$

$$\sigma^2 \le \max\left\{ \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}[A_i A_i^T] \right\|_2, \left\| \frac{1}{n}\sum_{i=1}^n \mathbb{E}[A_i^T A_i] \right\|_2 \right\}$$

and therefore

$$\mathbb{P}\left[\left\| \frac{1}{n}\sum_{i=1}^n A_i \right\|_2 \ge t \right] \le 2(d_1+d_2) e^{-\frac{nt^2}{2\sigma^2 + bt)}}$$

Covariance Estimation for general distributions

Let $X \in \mathbb{R}^d$ be a random vector with $\mathbb{E}X = 0$. Define $\Sigma = \mathbb{E}XX^T$ and $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ where X_i are i.i.d realizations of X.

[If X_i were subGaussian, we proved $\frac{\mathbb{E}\|\Sigma_n - \Sigma\|_2}{\|\Sigma\|_2} \le K^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)$]

Thm: Assume for some $K > 0$, have

$$\|X\|_2 \le K\sqrt{\operatorname{tr}\Sigma} \qquad \left[= \sqrt{\mathbb{E}\|X\|_2^2}\right]$$

Then

$$\frac{\mathbb{E}\|\Sigma_n - \Sigma\|_2}{\|\Sigma\|_2} \le C\left(\sqrt{\frac{K^2 r \log(d)}{n}} + \frac{K^2 r \log(d)}{n}\right)$$

where $r := \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_2} \le d$

pf: Matrix Bernstein

$$\mathbb{E}\|\Sigma_n - \Sigma\|_2 = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n (x_i x_i^T - \Sigma)\right\|_2$$

$$\le C\left(\sqrt{\sigma^2 \log(d)/n} + M\log(d)/n\right)$$

where $\sigma^2 = \left\|\mathbb{E}(xx^T - \Sigma)^2\right\|_2$ and $M \ge \|xx^T - \Sigma\|_2$

Let's estimate σ^2 and M.

$$\mathbb{E}(xx^T - \Sigma)^2 = \mathbb{E}(xx^T)^2 - \Sigma^2 \lesssim \mathbb{E}(xx^T)^2$$

and

$$(xx^T)^2 = \|x\|_2^2 xx^T \lesssim K^2(\operatorname{tr}\Sigma)\, xx^T$$

Take expectations $\Rightarrow \mathbb{E}(xx^T)^2 \lesssim K^2 \operatorname{tr}(\Sigma)\Sigma$

$$\Rightarrow \sigma^2 \le K^2 \operatorname{tr}(\Sigma)\|\Sigma\|_2$$

Finally

$$\|xx^T - \Sigma_i\|_2 \le \|x\|_2^2 + \|\Sigma_i\|_2$$
$$= K^2 \operatorname{tr}\Sigma_i + \|\Sigma_i\|_2$$
$$\le 2K^2 \operatorname{tr}\Sigma_i =: M$$

So

$$\mathbb{E}\|\Sigma_n - \Sigma_i\|_2 \le C\left(\sqrt{\frac{K^2 \operatorname{tr}(\Sigma_i)\|\Sigma_i\|_2 \log(d)}{n}} + \frac{2K^2 \operatorname{tr}(\Sigma_i)\log(d)}{n}\right)$$

Simplifying completes the proof. □

Sparse covariance estimation

If we know a priori that Σ_i is sparse, we should be able to estimate it using fewer observations. Goal: replace $\|\Sigma_n - \Sigma_i\|_2 \le \sqrt{\frac{d}{n}}$ by $\|\hat{\Sigma}_n - \Sigma_i\|_2 \le \sqrt{\frac{\ell \log(d)}{n}}$ where "ℓ" is the sparsity level.

For $\lambda > 0$, define the hard-thresholding operator

$$T_\lambda(u) := \begin{cases} u, & \text{if } |u| > \lambda \\ 0, & \text{o.w.} \end{cases}$$

As a surrogate for sparsity we will use $\|A\|_2$ where

$$A_{ij} = \begin{cases} 1, & \text{if } i=j \text{ or } \Sigma_{ij} \neq 0 \\ 0, & \text{o.w.} \end{cases}$$

Lemma: If Σ has at most s nonzero entries per row, then $\|A\|_2 \leq s$.

pf: $A + (s-1)I$ is diagonally dominant and therefore positive semidefinite.

$\Rightarrow |\lambda_j(A)| \leq s-1$.

Let u be an eigenvector of A corresponding to $\gamma_1(A)$ and let j index its maximal entry. WLOG $u_j > 0$

Then $\gamma_1 u_j = (Au)_j = \langle A_j, u \rangle \leq s u_j$ ∎

Thm: Let $\{X_i\}_{i=1}^n$ be i.i.d zero mean with covariance Σ, and suppose X_{ij} is σ-subGaussian. Then as long as $n > \log(d)$ setting

$$\lambda_n = \sigma^2\left(8\sqrt{\frac{\log(d)}{n}} + \delta\right)$$

it holds

$$\mathbb{P}\left[\frac{\|T_{\lambda_n}(\Sigma_n) - \Sigma\|_2}{\sigma^2} \geq 2\|A\|_2\left(8\sqrt{\frac{\log(d)}{n}} + \delta\right)\right] \leq 8e^{-\frac{n}{16}\min\{\delta, \delta^2\}}$$

We proceed with the following lemmas.

Lemma: Suppose $\|\Sigma_n - \Sigma\|_{max} \leq \lambda_n$ (coordinate wise)
Then $|T_{\lambda_n}(\Sigma_n) - \Sigma| \leq 2\lambda_n A$ (coordinate wise)

pf: $A^{ij} = 0 \Rightarrow |\Sigma_n^{ij}| \leq \lambda_n \Rightarrow T_{\lambda_n}(\Sigma_n^{ij}) = 0$

$A^{ij} \neq 0 \Rightarrow |T_{\lambda_n}(\Sigma_n^{ij}) - \Sigma^{ij}| \leq |T_{\lambda_n}(\Sigma_n^{ij}) - \Sigma_n^{ij}| + |\Sigma_n^{ij} - \Sigma^{ij}|$

$\leq 2\lambda_n$ ∎

Lemma: For any symmetric matrices $0 \leq A \leq B$, it holds $\|A\|_2 \leq \|B\|_2$.

pf: Observe

$$\langle a_i, a_j \rangle \leq \langle b_i, b_j \rangle$$

$$\Rightarrow \quad A^2 \leq B^2$$

Since A^2 and B^2 are positive semidefinite, $\|A^2\|_2 = \gamma_1(A^2)$

$$\|B^2\|_2 = \gamma_1(B^2)$$

The Perron-Frobenius theorem guarantees that there exists an eigenvector $u \geq 0$ corresponding to $\gamma_1(A^2)$. In this symmetric setting, this follows directly from:

$$\gamma_1(A^2) = \sup_{\|u\| \leq 1} \sum (A^2)_{ij} u_i u_j \leq \sup_{\|u\| \leq 1} \sum (A^2)_{ij} |u_i| |u_j| \leq \gamma_1(A^2)$$

Therefore

$$\gamma_1(A) \leq \langle A^2 u, u \rangle \leq \sup_{\|v\|_2 \leq 1} \langle B^2 v, v \rangle = \|B^2\|_2$$

$$\Rightarrow \quad \|A\|_2^2 = \|A^2\|_2 \leq \|B^2\|_2 \leq \|B\|_2^2. \qquad \square$$

So $|\Sigma_n - \Sigma|_{max} \le \lambda_n$

Lemma: $\Rightarrow \|T_{\lambda_n}(\Sigma_n) - \Sigma\|_2 \le 2\lambda_n \|A\|_2$

$$\mathbb{P}\left[\frac{|\Sigma_n - \Sigma|_{max}}{\sigma^2} \ge t\right] \le 8d\, e^{2 - \frac{n}{16}t \wedge t^2}$$

pf: Set $\Delta_n := \Sigma_n - \Sigma$. Observe

$\Delta_n^{ii} = \frac{1}{n}\sum_{k=1}^n (X_k)_i^2 - \mathbb{E}X_i^2$. Apply Bernstein

If $i \neq j$, then

$$2\Delta_n^{ij} = \frac{1}{n}\sum_{k=1}^n 2(X_k)_i (X_k)_j - 2\Sigma_{ij}$$

$$= \frac{1}{n}\sum_{k=1}^n \Big(\big[(X_k)_i + (X_k)_j\big]^2 - \big[\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij}\big] - (X_k)_i^2 - (X_k)_j^2 - \Sigma_{ii} - \Sigma_{jj} \Big)$$

Apply Bernstein again + union bound.

∎

Notice the bound

$$\|T_{\lambda_n}(\Sigma_W) - \Sigma\|_2 \le C\|A\|_2\sqrt{\frac{\log(d)}{n}}$$

can be $\sim d\sqrt{\frac{\log d}{n}}$, if Σ is dense, with small entries. This is much worse than our bounds from before.

Operator vs row norm:

For any $A \in S^d$, it holds

$$\|A\|_2 \ge \max_i \|A_i\|_2$$

Thm: Let $A \in S^d$ whose entries on and above the diagonal are independent mean zero. Then

$$\mathbb{E}\|A\|_2 \le C\sqrt{\log(d)} \cdot \mathbb{E}\max_i \|A_i\|_2$$

The proof is based on Bernstein.

pf: Write

$$A = \sum_{i \le j} Z_{ij} \quad \text{where}$$

$$Z_{ij} = \begin{cases} A_{ij}(e_i e_j^T + e_j e_i^T), & \text{if } i \neq j \\ A_{ii}, & \text{if } i = j \end{cases}$$

$$\mathbb{E}\|A\|_2 = \mathbb{E}_{\varepsilon} \left\| \sum_{i \le j} Z_{ij} - \mathbb{E} \sum_{i \le j} Z'_{ij} \right\|_2$$

$$\le \mathbb{E}_{\varepsilon} \left\| \sum_{i \le j} \varepsilon_{ij} (Z_{ij} - Z'_{ij}) \right\|_2$$

$$\le 2 \mathbb{E}_{\varepsilon} \left\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \right\|_2$$

Recall $\varepsilon_{ij} Z_{ij}$ is Z_{ij}^2-subGaussian.

Therefore with $\sigma^2 = \left\| \sum_{ij} Z_{ij}^2 \right\|_2$ get

$$\mathbb{P}\left[\left\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \right\|_2 \ge t \right] \le 2d \cdot \exp\left(-\frac{t^2}{2\sigma^2} \right)$$

$$= \exp\left(\log(2d) - \frac{t^2}{2\sigma^2} \right)$$

set $t = \sqrt{2\sigma^2 \log(2d) + u^2}$

$$\Rightarrow \mathbb{P}\left[\left\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \right\| \ge \sqrt{2\sigma^2 \log(2d)} + u \right] \le e^{-\frac{u^2}{2\sigma^2}}$$

Integrating, get

$$\mathbb{E}_\varepsilon \left\| \sum_{i,j} \varepsilon_{ij} Z_{ij} \right\| \le C\sqrt{\log d} \cdot \underbrace{\sqrt{\left\| \sum_{i \le j} Z_{ij}^2 \right\|_2}}_{\sigma}$$

Next evaluate $\mathbb{E}_Z \sqrt{\left\| \sum_{ij} Z_{ij} \right\|_2}$

Quick computation shows

$$Z_{ij}^2 = \begin{cases} A_{ij}^2 (e_i e_i^T + e_j e_j^T), & i < j \\ A_{ii}^2 e_i e_i^T, & i = j. \end{cases}$$

↑ diagonal.

$$\Rightarrow \sum_{i \le j} Z_{ij}^2 = \begin{bmatrix} \|A_1\|_2^2 & & 0 \\ & \|A_2\|_2^2 & \\ & & \ddots \\ 0 & & \|A_d\|_2^2 \end{bmatrix}$$

$$\Rightarrow \sqrt{\left\| \sum_{i \le j} Z_{ij}^2 \right\|_2} = \max_{i=1,\dots,d} \|A_i\|_2$$

∎

Cor (HW) Let $A \in \mathbb{R}^{m \times n}$ whose entries are independent mean-zero. Then

$$\mathbb{E}\|A\|_2 \leq C\sqrt{\log(m+n)}\left(\mathbb{E}\max_i \|A_i\|_2 + \mathbb{E}\max_j \|A^j\|_2\right)$$

Application: Matrix Completion.

Consider $X \in \mathbb{R}^{d \times d}$ with $\text{rank}(X) = r \ll d$.

Suppose each entry X_{ij} is revealed independently with probability $p \in (0,1)$

So we get to see $Y \in \mathbb{R}^{d \times d}$ with

$$Y_{ij} = \delta_{ij} X_{ij} \text{ where } \delta_{ij} \sim \text{Ber}(p)$$

Thus we get to see

$$m := p \cdot d^2 \text{ entries on average}$$

Thm: Let $\hat{X}$ be a best rank-r approximation to $p^{-1}Y$. Then

$$\mathbb{E}\,\frac{1}{d}\|\hat{X}-X\|_F \le C\sqrt{\frac{rd\log(d)}{m}}\,\|X\|_{\max}$$

as long as $m \ge \log(d)$.

So in order to ensure

$$\mathbb{E}\sqrt{\frac{1}{d^2}\sum_{i,j}(\hat{X}_{ij}-X_{ij})^2} \le \varepsilon\,\|X\|_{\max},$$

it suffices to see $m = \frac{1}{\varepsilon^2} rd\log(d)$ entries on average.

pf: First bound error is operator norm

$$\|\hat{X}-X\|_2 \le \|\hat{X}-p^{-1}Y\|_2 + \|p^{-1}Y - X\|_2$$

$$\le 2\|p^{-1}Y - X\|_2 = \frac{2}{p}\underbrace{\|Y-pX\|_2}_{\text{easy to understand}}$$

$$(Y - pX)_{ij} = (\delta_{ij} - p) X_{ij}$$

δ_{ij}: independent mean-zero

$$\Rightarrow \mathbb{E}\|Y - pX\|_2 \le C\sqrt{\log(n)} \cdot \left(\mathbb{E}\max_i \|(Y-pX)_i\|_2 + \mathbb{E}\max_j \|(Y-pX)^j\|_2\right)$$

Observe

$$\|(Y-pX)_i\|_2^2 = \sum_j (Y-pX)_{ij}^2 = \sum_j (\delta_{ij}-p)^2 X_{ij}^2$$

$$\le \left(\sum_j (\delta_{ij}-p)^2\right) \|X\|_{max}^2$$

Can bound (HW) $\mathbb{E}\max_i \sum_j (\delta_{ij}-p)^2 \le Cpd$

using concentration and union bound.

$$\Rightarrow \mathbb{E}\|Y - pX\|_2 \le C\sqrt{pd\log(d)}\,\|X\|_{max}$$

$$\Rightarrow \mathbb{E}\|\hat{X} - X\|_2 \le C\sqrt{\frac{d\log(d)}{p}}\,\|X\|_{max}$$

Step 2: Pass to Frobenius.

$$\|\hat{X} - X\|_F \leq \sqrt{2r}\,\|\hat{X} - X\|_2$$

($\hat{X}$: rank $2r$)

$$\Rightarrow \mathbb{E}\,\frac{1}{d}\|\hat{X} - X\|_F \leq C\sqrt{\frac{rd\log(d)}{pd^2}}\,\|X\|_{\max}.$$

(pd^2: m)

∎