Lecture 13 :- Proof of matrix Bernstein.

Goal (1) was :- Get bounds on $\psi_M(\alpha) = E[\exp(\alpha M)]_{\alpha \geq 0}$,

Conditions :- $\|M\| \leq s$.

$$\exp(\alpha M) = I + \alpha M + \frac{\alpha^2 M^2}{2} + \frac{\alpha^3 M^3}{3!} + \cdots$$

→ All the summands commute, and are Hermition.

$$= I + \alpha M + \frac{\alpha^2 M^2}{2} + B.$$

$$B = \sum_{i \geq 2} \frac{\alpha^i M^i}{i!}$$

$$\|B\| \leq \sum_{i \geq 2} \frac{\alpha^i \|M\|^i}{i!} \leq \frac{\alpha^2 s^2}{2} \sum_{j \geq 0} \frac{\alpha^j s^j}{3^j}$$

$$\leq \frac{\alpha^2 s^2}{2} \left(\frac{1}{1 - \frac{\alpha s}{3}} \right) \quad \text{when } \frac{\alpha s}{3} < 1$$

Similarly.

$$\exp(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots$$

$$\leq 1 + x + \frac{x^2}{2}\left(1 + \frac{x}{3} + \frac{x^2}{3 \cdot 4} + \frac{x^3}{3 \cdot 4 \cdot 5} + \cdots \right)$$

$$\leq 1 + x + \frac{x^2}{2\left(1 - \frac{|x|}{3}\right)} \quad \text{when } |x| < 3.$$

If $|x| < s < 3$ then.

$$\exp(x) \leq 1 + x + \frac{x^2}{2\left(1 - \frac{s}{3}\right)}$$

So for every Hermitian H s.t. $\|H\| \leq K < 3$,

we have $-3 < -K \leq \lambda(H) \leq K < 3$

$$\exp(H) \preceq I + H + \frac{H^2}{2\left(1-\frac{K}{3}\right)}.$$

$\forall$ e.v. $\lambda(H)$ of H.

(Why?).

(∗) (In general if H, Hermitian has eigenvalues in $[a,b]$ and $f, g : \mathbb{R} \to \mathbb{R}$ are s.t. $f(x) \leq g(x)$ $\forall x \in [a,b]$, then $f(H) \preceq g(H)$.)

(It is important that all matrices involved have the same set of orthonormal vectors).

In our setting

$$\exp(\alpha M) \preceq I + \alpha M + \frac{\alpha^2 M^2}{2(1-|\alpha|s/3)}$$

provided that $|\alpha| < \frac{3}{s}$.

(Recall that $\|M\| \leq s$).

$\therefore$ By taking expectation (Why is this allowed?)

$$E[\exp(\alpha M)] \preceq I + \underbrace{\alpha E[M]}_{0} + \frac{\alpha^2 E[M^2]}{2(1-|\alpha|s/3)}.$$

$$= I + g(\alpha)\, E[M^2]$$

where $g(\alpha) := \frac{\alpha^2}{\left(1-\frac{|\alpha|s}{3}\right)}$, $|\alpha| < \frac{3}{s}$.

So we have.

$$\Psi_M(\alpha) \preceq I + g(\alpha)\, E[M^2] \quad \text{when } |\alpha| < \frac{3}{s}.$$

Now since log is operator monotone

$$\log \Psi_M(\alpha) \preccurlyeq \log\left(I + g(\alpha)\,E[M^2]\right).$$

Now, because $\log(1+x) \leq x$ when $x > -1$, we have,

$$\log\left(I + g(\alpha)E[M^2]\right) \preccurlyeq g(\alpha)E[M^2]$$

(From $(*)$).

when $g(\alpha)E[M^2] + I \succ 0$; which is true for all Hermitian M, and α where $g(\alpha)$ is defined).

So, the conclusion is: When $\|M\| \leq S$ then

$$\log E[\exp(\alpha M)] = \log \Psi_M(\alpha) \preccurlyeq g(\alpha)\,E[M^2] \quad \text{where}$$

$$g(\alpha) := \frac{\alpha^2}{1 - |\alpha| S/3},$$

for all $|\alpha| < \frac{3}{S}$.

Goal 2 :- Understand $\mathrm{Tr}\,\Psi_{\sum M_i}(\alpha)$.

$$\mathrm{Tr}\,\Psi_{\sum M_i}(\alpha) = \mathrm{Tr}\left(E\left[\exp\left(\alpha \sum_{i=1}^{n} M_i\right)\right]\right)$$

$$= E\left[\mathrm{Tr}\left[\exp\left(\alpha \sum_{i=1}^{n} M_i\right)\right]\right] \qquad \left(\mathrm{Tr}(E[\cdot]) = E[\mathrm{Tr}[\cdot]]\right)$$

Commutive case: $\exp\left(\alpha \sum_{i=1}^{n} M_i\right) = \prod_{i=1}^{n} \exp(\alpha M_i)$

$$= \exp\left(\sum_{i=1}^{n} \underbrace{\log(\exp(\alpha M_i))}_{\text{and push expectation here}}\right)$$

But $e^{A+B} \neq e^A e^B$ in general when A, B are matrices.

But we can use Lieb's theorem. Recall:

If H is fixed, Hermitian, and Z is a random Hermitian matrix then

$$E\left[\operatorname{tr}\left(\exp(H+Z)\right)\right] \le \operatorname{Tr}\left[\exp\left(H + \log E[\exp(Z)]\right)\right]$$

Let $H = \underbrace{\alpha \sum_{i=1}^{n-1} M_i}_{\text{Fixed realization.}}$. $Z = \alpha M_n$.

$$E\left[\operatorname{tr}\left(\exp\left(\alpha \sum_{i=1}^{n} M_i\right)\right)\right]$$

$$= E_{M_1 \dots M_{n-1}}\left[E_{M_n}\left[\operatorname{tr}\left(\exp\left(\alpha \sum_{i=1}^{n-1} M_i + \alpha M_n\right)\right) \Big| M_1 \dots M_{n-1}\right]\right]$$

$$\overset{\text{Lieb}}{\le} E_{M_1 \dots M_{n-1}}\left[\operatorname{tr}\left(\exp\left(\alpha \sum_{i=1}^{n-1} M_i + \log E\left[\exp(\alpha M_n) \big| M_1 \dots M_{n-1}\right]\right)\right)\right]$$

But since M_n is independent of $M_1 \dots M_{n-1}$, this becomes.

$$= E_{M_1 \dots M_{n-1}}\left[\operatorname{tr}\left(\exp\left(\alpha \sum_{i=1}^{n-1} M_i + \log E[\exp(\alpha M_n)]\right)\right)\right]$$

We can repeat this, conditioninoning on $M_1 \dots M_{n-i}$ at the ith step and defining

$$H = \sum_{j=1}^{n-i} \alpha M_i + \sum_{j=n-i+2}^{n} \log E[\exp(\alpha M_j)] .$$

Finally we get.

$$\mathrm{Tr}\left(\Psi_{\sum M_i}(\alpha)\right) \preccurlyeq \mathrm{Tr}\left(\exp\left(\sum_i \log E[\exp(\alpha M_i)]\right)\right)$$

$$= \mathrm{Tr}\left[\exp\left(\sum_{i=1}^{n} \log \Psi_{M_i}(\alpha)\right)\right]$$

Recall :-

$$\log \Psi_{M_i}(\alpha) \preccurlyeq g(\alpha)\, E[M_i^2]$$

when $\alpha < \frac{3}{S}$, $\|M_i\| \leq S$.

Therefore

$$\sum_{i=1}^{n} \log \Psi_{M_i}(\alpha) \preccurlyeq g(\alpha) \sum_{i=1}^{n} E[M_i^2]. \quad \text{— (1)}$$

Recall:- If f is increasing on $(-\infty, \infty)$ then $\forall A \preccurlyeq B$ Hermitian, $\mathrm{Tr}(f(A)) \leq \mathrm{Tr}(f(B))$. [We proved this for twice differentiable f.]

So (1) gives

$$\mathrm{Tr}\left(\exp\left(\sum_{i=1}^{n} \log \Psi_{M_i}(\alpha)\right)\right)$$

$$\leq \mathrm{Tr}\left(\exp\left(g(\alpha) \sum_{i=1}^{n} E[M_i^2]\right)\right)$$

$$\leq d\, \lambda_{max}\left(\exp\left(g(\alpha) \sum_{i=1}^{n} E[M_i^2]\right)\right).$$

$$= d \exp\left(g(\alpha)\, \lambda_{max}\left(\sum_{i=1}^{n} E[M_i^2]\right)\right)$$

$(\because g(\alpha) \geq 0)$.

$$= d \exp\left(g(\alpha) \left\| \sum_{i=1}^{n} E[M_i^2] \right\|\right)$$

(Since $M_i^2 \succcurlyeq 0$, and $\lambda_{max}(A) = \|A\|$ when $A \succcurlyeq 0$.)

[$|\mathrm{Tr}(A)| \leq d \cdot |\lambda_{max}(A)|$ for all sq. A.
$\mathrm{Tr}(A) \leq d\, \lambda_{max}(A)$ for Hermitian A]

$$\leq d \exp(n\, g(\alpha)\, \sigma^2)$$

$$\left(\because \text{ By assumption } \left\| \frac{1}{n}\sum_{i=1}^{n} E[M_i^2] \right\| \leq \sigma^2. \right.$$

So, we get.

$$\operatorname{Tr} \psi_{\sum_{i=1}^{n} M_i}(\alpha) \leq d \exp(n\, g(\alpha)\, \sigma^2).$$

Therefore by the previous calculation,

$$\Pr\left[\lambda_{max}\left(\frac{1}{n}\sum_{i=1}^{n} M_i\right) \geq t \right] = \Pr\left[\lambda_{max}\left(\sum_{i=1}^{n} M_i\right) \geq nt \right]$$

$$\leq \exp(-n\alpha t) \operatorname{Tr} \psi_{\sum_{i=1}^{n} M_i}(\alpha) \qquad \forall\, \alpha > 0$$

$$\leq d \exp(-n(\alpha t - g(\alpha)\sigma^2)) \qquad \forall\, \alpha > 0,\ \alpha < \frac{3}{S}.$$

$$h(\alpha) = \alpha t - \frac{\alpha^2/2}{\left(1 - \frac{\alpha S}{3}\right)}.$$

Put $\alpha = \dfrac{t}{\sigma^2 + St/3} \in \left(0, \dfrac{3}{S}\right)$, as $t > 0$, and get.

(Check this later, from Versynin's book).

$$\Pr\left(\lambda_{max}\left(\frac{1}{n}\sum_{i=1}^{n} M_i \right) \geq t \right)$$

$$\leq d \exp\left(- \frac{nt^2/2}{\sigma^2 + St/3} \right) \quad \text{— } (1)$$

For Hermitian M, $\|M\| = \max(|\lambda_{max}(M)|, |\lambda_{min}(M)|)$

$$= \max(|\lambda_{max}(M)|, |\lambda_{max}(-M)|).$$

So, by applying ① to the collection $(-M_i)_{i=1}^n$ as well we get:

$$\boxed{\begin{aligned} \Pr\left(\left\|\tfrac{1}{n}\sum M_i\right\| \geq t\right) &\leq 2d\exp\left(-\frac{n\,t^2/2}{\sigma^2+st/3}\right) \\ &\leq 2d\exp\left(-n\min\left(\frac{t^2}{4\sigma^2},\frac{3t}{4s}\right)\right) \end{aligned}}$$

__Application to covariance estimation:-__

$X^{(i)} \sim X$ random vectors, with $E[XX^T]=\Sigma$

__assumption:__ $\|X\| \leq \sqrt{K\,E[\|X\|^2]}$
$= \sqrt{K\,\mathrm{tr}\,\Sigma}$

Define $B^{(i)} = X^{(i)}X^{(i)T} - \Sigma$.

$E[B^{(i)}] = 0$, $B^{(i)}$ independent.

$$\begin{aligned} \|B^{(i)}\| &\leq \|\Sigma\| + \|X^{(i)}X^{(i)T}\| \\ &= \|\Sigma\| + \|X^{(i)}\|^2 \\ &\leq \|\Sigma\| + K^2\,\mathrm{Tr}\,\Sigma \\ &\leq 2K^2\,\mathrm{Tr}\,\Sigma \quad (\text{assuming } K \geq 1). \end{aligned}$$

So can take $s = 2K^2\,\mathrm{Tr}(\Sigma)$

$$\begin{aligned} 0 \preccurlyeq E[B^{(i)2}] &= E[\|X\|^2 X^{(i)}X^{(i)T}] - \Sigma^2. \\ &\preccurlyeq E[\|X\|^2 X^{(i)}X^{(i)T}] \end{aligned}$$

Now $0 \preccurlyeq \|X\|^2 \; X^{(i)} X^{(i)T} \preccurlyeq (K^2 \operatorname{tr} \Sigma) \; X^{(i)} X^{(i)T}$

$$\Rightarrow \quad \Rightarrow 0 \preccurlyeq E\left[\|X\|^2 X^{(i)} X^{(i)T}\right] \preccurlyeq K^2(\operatorname{Tr}(\Sigma)) \cdot \Sigma.$$

$$\Rightarrow \quad 0 \preccurlyeq E\left[B^{(i)^2}\right] \preccurlyeq K^2(\operatorname{Tr}(\Sigma)) \cdot \Sigma.$$

$$\Rightarrow \quad \left\| \frac{1}{n} \sum_{i=1}^{n} E\left[B^{(i)^2}\right] \right\| \leqslant K^2 \operatorname{Tr}(\Sigma) \cdot \|\Sigma\|.$$

So, we can take

$$\sigma^2 = K \operatorname{Tr}(\Sigma) \cdot \|\Sigma\|.$$

So, we get.

$$\Pr\left(\left\| \left(\frac{1}{n} \sum_{i=1}^{n} X^{(i)} X^{(i)T}\right) - \Sigma \right\| \geqslant \varepsilon \|\Sigma\| \right)$$

$$\leqslant 2d \cdot \exp\left(-n \|\Sigma\| \left(\frac{\varepsilon^2}{4K \operatorname{Tr}(\Sigma)}, \frac{\varepsilon}{8K \operatorname{tr}(\Sigma)}\right)\right)$$

$$= 2d \cdot \exp\left(-n \cdot \underbrace{\frac{\|\Sigma\|}{\operatorname{Tr}(\Sigma)}}_{\substack{\text{In general,} \\ \operatorname{Tr}(\Sigma) \leqslant d\|\Sigma\|}} \cdot \left(\frac{\varepsilon^2}{4K^2}, \frac{\varepsilon}{8K^2}\right)\right)$$

$$\leqslant 2d \cdot \exp\left(-\frac{n}{d}\left(\frac{\varepsilon^2}{4K^2}, \frac{\varepsilon}{8K^2}\right)\right).$$

So, $n \gtrsim \frac{d}{\varepsilon^2} \log(d)$ is sufficient to drive the error close to zero.