In this lecture we will complete the proof of Matrix Bernstein's inequality. We will also introduce the scalar Hoeffding and Bernstein's inequalities.

1 Notation

A quick summary of the notation:

For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we use $\|A\|_{op}$ to denote its operator/spectral norm (i.e., the largest singular value of $A$).

For two $d \times d$ symmetric matrices $A, B$, the positive-semidefinite ordering $A \succeq B$ means that $A - B$ is positive-semidefinite matrix, i.e., $\lambda_i(A - B) \geq 0, \forall i$. It implies that $\lambda_i(A) \geq \lambda_i(B)$, for $i = 1, ..., d$.

2 Recap

We have introduced the statement of Matrix Bernstein’s Inequality, and covered several theorems which would be leveraged for the proof of Matrix Bernstein’s Inequality.

**Theorem 1** (Matrix Bernstein’s Inequality). Suppose $X_1, ..., X_n \in \mathbb{R}^{d_1 \times d_2}$ are independent random matrices satisfying the following conditions:

- $E[X_i] = 0$, for all $i \in \{1, 2, ..., n\}$,
- $\|X_i\|_{op} \leq b$ almost surely for all $i \in \{1, 2, ..., n\}$,
- $\max \left[ \left\| \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \right\|_{op}, \left\| \frac{1}{n} \sum_{i=1}^{n} X_i^\top X_i \right\|_{op} \right] \leq \sigma^2$,

then for every $t > 0$, we have

$$\mathbb{P}\left[ \left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\|_{op} \geq t \right] \leq (d_1 + d_2) \exp \left( \frac{-t^2/2}{\sigma^2 + bt/3} \right). \quad (1)$$

**Lemma 1** (Matrix Laplace Transform). Suppose $Y$ is a random symmetric matrix, then for any $t \in \mathbb{R}$,

$$\mathbb{P}(\lambda_{\max}(Y) \geq t) \leq \inf_{\theta > 0} e^{-\theta t} \cdot E[\text{tr}(e^{\theta Y})],$$

where $\lambda_{\max}(Y)$ is the maximum eigenvalue of $Y$.

**Theorem 2** (Lieb's Theorem). For a fixed symmetric matrix $H$, the matrix function $f$ defined through

$$f(A) \triangleq \text{tr} \exp(H + \log A)$$

is concave on the space of positive symmetric matrices with the same size as $H$.

The Lieb’s Theorem together with Jensen’s inequality implies that

$$E[\text{tr} \exp(H + X)] \leq \text{tr} \exp(h + \log E[e^X]). \quad (2)$$

3 Matrix Theory Background, cont’d

With Lieb’s theorem, we can proof the following lemma, which is the key in the proof of Matrix Bernstein’s Inequality.

**Lemma 2 (Sub-additivity of Matrix MGF).** Suppose $X_1, ..., X_n$ are independent symmetric matrices, then

$$
\mathbb{E} \left[ \text{tr} \exp(\theta \sum_{i=1}^{n} X_i) \right] \leq \text{tr} \exp \left( \sum_{i=1}^{n} \log \mathbb{E}[e^{\theta X_i}] \right), \quad \text{for all } \theta. \quad (3)
$$

**Proof** We note that the summation of $\theta \sum_{i=1}^{n} X_i$ can be decomposed and the Equation (2) implied by the Lieb’s Theorem can be applied iteratively as follows:

- Step (i): follows from invoking the inequality $[2]$, and step (ii) follows from independence.

By combining Lemma 1 and Lemma 2, the following Theorem 3 (Master Bound) can be derived.

**Theorem 3 (Master Bound).** Suppose $X_1, ..., X_n$ are independent symmetric matrices, then for any $t \in \mathbb{R},$

$$
\mathbb{P} \left[ \lambda_{\max}(Y) \geq t \right] \leq \inf_{\theta > 0} e^{-\theta t} \cdot \text{tr} \exp \left( \sum_{i=1}^{n} \log \mathbb{E}[e^{\theta X_i}] \right).
$$

The master bound (and its variants) can be used to prove the matrix versions of different inequalities, such as Hoeffding, Bernstein, Chernoff, Azuma, Bounded Difference, Bennett, Freeman.

4 Proof of the Matrix Bernstein’s Inequality

We shall prove the following symmetric version of the Matrix Bernstein’s inequality.

**Theorem 4 (Matrix Bernstein’s Inequality: Symmetric Case).** Suppose $X_1, ..., X_n \in \mathbb{R}^{d \times d}$ are independent symmetric random matrices with the following conditions:

- $\mathbb{E}[X_i] = 0$, for all $i \in \{1, 2, ..., n\},$
- $\lambda_{\max}(X_i) \leq b$ almost surely for all $i \in \{1, 2, ..., n\},$
- $\|\sum_{i=1}^{n} X_i\|_{op} \leq \sigma^2,$
then for every $t > 0$, we have

$$
P \left[ \lambda_{\max} \left( \sum_{i=1}^{n} X_i \right) \geq t \right] \leq d \exp \left( \frac{-t^2/2}{\sigma^2 + bt/3} \right). \quad (4)$$

**Remark** To prove the rectangular version of Matrix Bernstein’s Inequality in Theorem [1] we can apply Theorem 4 to the symmetric matrix $Y = \left[ \begin{array}{c} X \end{array} \right]$ where $X \in \mathbb{R}^{d_1 \times d_2}$ is a general rectangular matrix and $Y \in \mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2)}$ called the symmetric dilation of $X$. We can then use $\lambda_{\max}(Y) = ||X||_{op}$ to finish the proof.

**Proof** Define the scalar function $f$ as

$$f(x) \triangleq \frac{e^{\theta x} - 1 - \theta x}{x^2}.$$  

which is an increasing function.

For any symmetric matrix $X$ with $\lambda_{\max}(X) \leq b$, we have

$$e^{\theta X} = I + \theta X + X f(X) X$$

$$\leq I + \theta X + f(b) X^2.$$  

We use a scalar inequality: for any $\theta : 0 < \theta < \frac{3}{b}$, it holds that

$$f(b) = \frac{e^{\theta b} - 1 - \theta b}{b^2}$$

$$= \frac{1}{b^2} \sum_{k=2}^{\infty} \frac{(\theta b)^k}{k!}$$

$$\leq \frac{\theta^2}{2} \sum_{k=2}^{\infty} \frac{(\theta b)^{k-2}}{3^{k-2}}$$

$$= \frac{\theta^2/2}{1 - \theta b/3}.$$  

Combining the above inequalities, we have for any $\theta : 0 < \theta < \frac{3}{b}$, and any $X : \lambda_{\max}(X) \leq b$, it holds that

$$e^{\theta X} \leq I + \theta X + \frac{\theta^2/2}{1 - \theta b/3} X^2.$$  

It follows that

$$E[e^{\theta X}] \leq I + 0 + \frac{\theta^2/2}{1 - \theta b/3} E[X^2]$$

$$\leq \exp \left( \frac{\theta^2/2}{1 - \theta b/3} E[X^2] \right),$$  

Since matrix logarithm is operator monotone, we obtain that

$$\log E[e^{\theta X}] \leq \frac{\theta^2/2}{1 - \theta b/3} E[X^2]. \quad (5)$$
Letting \( g(\theta) := \frac{\theta^2}{2 - 3\theta} \), we have

\[
P\left[ \lambda_{\max}(\sum_{i=1}^{n} X_i) \geq t \right] \leq \inf_{\theta > 0} \frac{\text{tr} \exp\left( \sum_{i=1}^{n} \log E[e^{\theta X_i}] \right)}{e^{\theta t}}
\]

\[
\leq \inf_{0 < \theta < \frac{t}{\sigma^2}} \frac{\text{tr} \exp\left( \sum_{i=1}^{n} g(\theta) \ E[X_i^2] \right)}{e^{\theta t}}
\]

\[
\leq \inf_{0 < \theta < \frac{t}{\sigma^2}} \frac{d \cdot \exp (g(\theta)\sigma^2)}{e^{\theta t}},
\]

Here the first inequality is due to the Theorem 3 (Master Bound), and the second inequality is due to the Equation (5).

Taking \( \theta = t/(\sigma^2 + bt/3) \) and simplifying the expression, we have the desired bound

\[
P\left[ \lambda_{\max}(\sum_{i=1}^{n} X_i) \geq t \right] \leq d \cdot \exp \left( \frac{-t^2}{2 \sigma^2} + \frac{bt}{3} \right).
\]

\[ \square \]

Remark

- The proof is quite similar to that of the scalar Bernstein’s Inequality. Most of the hard work is done by the Lieb’s Theorem.

- The dimension factor on the right hand side of Equation (4) leads to the \( \sqrt{\log d} \) factor in the user-friendly form of the matrix Bernstein’s inequality (see last lecture).

- To prove a tighter bound and relax this dimension dependence, one must better capture non-commutativity. There is active research on achieving such improvement but it is outside the scope of this course.

5 Sub-Gaussian/Exponential Random Variables and Scalar Hoeffding/Bernstein Inequalities

In this section, we will briefly introduce the scalar versions of Hoeffding’s Inequality and Bernstein’s Inequality.

Definition 1. A variable \( X \) is called sub-Gaussian with parameter \( \sigma^2 \), denoted as sub-Gaussian(\( \sigma^2 \)), if

\[
E e^{\lambda (X - E[X])} \leq e^{\lambda^2 \sigma^2/2}, \quad \text{for all } \lambda \in \mathbb{R}.
\]

(6)

Note that the right hand side above is the MGF of a zero-mean Gaussian random variable with variance \( \sigma^2 \).

Example 1 (Rademacher). If a random variable \( X \in \{-1, +1\} \) with the equal probability 1/2 for -1 and 1, then \( X \) is sub-Gaussian(1).

---

2 Reading: Martin J. Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Section 2.1
3 Reading: Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. Section 2
4 Reading: John Duchi’s Lecture Notes, Section 3.1 [https://web.stanford.edu/class/stats311/lecture-notes.pdf]
**Proof** For all $\lambda$, we have

$$E e^{\lambda X} = \frac{1}{2} e^{\lambda} + \frac{1}{2} e^{-\lambda}$$

$$= \frac{1}{2} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + \frac{1}{2} \sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!}$$

$$= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}$$

$$\leq \sum_{k=0}^{\infty} \frac{(\lambda^2)^k}{2^k k!}$$

$$= e^{\lambda^2/2}.$$  

---

**Example 2** (Bounded RV). If a random variable $X \in [a, b]$ with probability 1, then $X$ is sub-Gaussian($(b-a)^2$).

**Proof** We prove this by a symmetrization argument. Let $\varepsilon \in \{-1, +1\}$ be a random variable with equal probability $1/2$ for $-1$ and $+1$, and $X'$ be an independent copy of $X$. Then we have $E[X'] = E[X]$ and

$$E e^{\lambda(X - E[X])} = E e^{\lambda(X - E[X'])} $$

$$\leq E e^{\lambda(X - X')} $$

$$= E e^{\lambda (X - X')} $$

$$= E[E[e^{\lambda (X - X')}|X, X']] $$

$$\leq E[e^{\lambda^2 (X - X')^2/2}] $$

$$\leq e^{\lambda^2 (b-a)^2/2}.$$  

The first inequality follows from the Jensen’s Inequality, and the second inequality follows from the previous example.

The lemma below provides an equivalent characterization of a sub-Gaussian random variable in terms of its tail probability.

**Lemma 3.** A variable $X$ is sub-Gaussian($\sigma^2$) if and only if for some universal constant $c > 0$,

$$P(|X| \geq t) \leq 2e^{-t^2/c\sigma^2}, \quad \text{for all } t \geq 0.$$  

(7)

We now state the Hoeffding’s inequality for sum of independent sub-Gaussian random variables. It generalizes the more commonly known Hoeffding’s inequality for sum of bounded random variables.

**Theorem 5** (Hoeffding’s Inequality). If $X_i$’s are independent sub-Gaussian($\sigma_i^2$) random variables, then for any $t \geq 0$,

$$P \left[ \left| \sum_i (X_i - E X_i) \right| \geq t \right] \leq 2 \exp \left( - \frac{t^2}{2 \sum_i \sigma_i^2} \right),$$

i.e., $\sum_i X_i$ is sub-Gaussian($\sum_i \sigma_i^2$).
One consequence of Hoeffding’s Inequality is

$$\left| \frac{1}{n} \sum_i (X_i - \mathbb{E} X_i) \right| \leq \frac{\sigma \sqrt{\log 1/\delta}}{\sqrt{n}},$$

with probability $1 - \delta$.

**Definition 2.** A variable $X$ is called sub-exponential with parameters $(\tau^2, b)$, denoted as sub-exponential$(\tau^2, b)$, if

$$\mathbb{E} e^{\lambda (X - \mathbb{E}[X])} \leq e^{\lambda^2 \tau^2 / 2}, \quad \text{for all } \lambda \in \mathbb{R} \text{ with } |\lambda| \leq \frac{1}{b}. \quad (8)$$

**Example 3 (Gaussian Squared).** If a variable $Z$ follows the normal distribution $N(0, 1)$ and $X \triangleq Z^2$, then $X$ is sub-exponential$(2, 4)$.

**Example 4 (Bounded RV).** If a variable $X \in [-b, b]$, with mean $\mathbb{E}[X] = 0$ and variance $\text{var}(X) = \sigma^2$, then $X$ is sub-Gaussian$((2b)^2)$, and is also sub-exponential$(6\sigma^2 / 5, 2b)$.

We now state the Bernstein’s inequality for sum of independent sub-exponential random variables. It generalizes the more commonly known Bernstein’s inequality for sum of bounded random variables.

**Theorem 6 (Bernstein’s Inequality).** If $X_i$’s are independent sub-exponential$(\tau_i^2, b_i)$ random variables with $\mathbb{E}[X_i] = 0$, then for any $t \geq 0$,

$$\mathbb{P} \left[ \left| \sum_i X_i \right| \geq t \right] \leq 2 \exp \left( -\frac{1}{2} \min \left\{ \frac{t^2}{\sum_i \sigma_i^2}, \frac{t}{\max_i b_i} \right\} \right).$$