36-710: Advanced Statistical Theory **Fall 2018**

Lecture 11: October 8

Lecturer: Alessandro Rinaldo *Scribes: Yue Li*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

11.1 Matrix Bernstein Theorem

Last time we came to half way of proof of matrix Bernstein theorem; we first complete the proof.

Theorem 11.1 *(Matrix Bernstein Theorem) Let $X_1, \dots, X_n$ be mean-zero, independent $d \times d$ symmetric matrices, s.t. $\|X_i\|_{op} \le C$, a.e., $\forall i$. Then $\forall t \ge 0$, we have*

$$\mathbb{P}\left(\| \sum_{i=1}^n X_i \| \ge t \right) \le 2d \exp\left\{ -\frac{t^2}{2(\sigma^2 + Ct/3)} \right\},$$

where $\sigma^2 = \| \sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \|_{op} = \| \sum_{i=1}^n Var(X_i) \|_{op}$. As a result,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^n X_i \right) \ge t \right) \le \inf_{\lambda > 0} \left\{ e^{-\lambda t} t \right\}.$$

Proof: In last lecture we completed the first two steps. We know

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^n X_i) \ge t \right) \le e^{-\lambda t} tr \exp\left\{ \sum_{i=1}^n \log\left(\mathbb{E}\left[\exp(\lambda X_i)\right] \right) \right\}.$$

Step 3: We handle term $\mathbb{E}\left[\exp(\lambda X_i)\right]$.

We will assume that $\mathbb{E}\left[\exp(\lambda X_i)\right] \le \exp\left\{g(\lambda) A_i\right\}$ for some A_i and function g. We first prove the following lemma:

Lemma 11.2 *Let $g : (0, \infty) \mapsto [0, \infty)$ and $A_1, \dots, A_n$ be $d \times d$ symmetric PSD matrices such that*

$$\mathbb{E}\left[\exp(\lambda X_i)\right] \preceq \exp(g(\lambda) A_i), \forall \lambda > 0.$$

Then it holds that

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^n X_i) \ge t \right) \le d \inf_{\lambda > 0} \exp\left\{ -\lambda t + g(\lambda)\sigma^2 \right\},$$

where $\sigma^2 = \lambda_{\max}(\sum_{i=1}^n A_i)$.

Proof: By the proof of Step 1 and Step 2, we know

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^n X_i) \geq t\right) \leq e^{-\lambda t} tr \exp\left\{\sum_{i=1}^n \log\left(\mathbb{E}\left[\exp(\lambda X_i)\right]\right)\right\}.$$

Recall two properties of matrix-valued functions:

Proposition 11.3 *1) operator monotonicity of matrix logarithm function:* $0 \prec A \preceq B \Rightarrow \log(A) \preceq \log(B)$*; 2) monotonicity of* $tr \exp(\cdot)$ *function:* $A \preceq B \Rightarrow tr \exp(A) \preceq tr \exp(B)$.

By 1) and $\mathbb{E}\left[\exp(\lambda X_i)\right] \preceq \exp(g(\lambda)A_i)$, we have $\log \mathbb{E}\left[\exp(\lambda X_i)\right] \preceq \log \exp(g(\lambda)A_i)$. So $\sum_{i=1}^n \log \mathbb{E}\left[\exp(\lambda X_i)\right] \preceq \sum_{i=1}^n g(\lambda)A_i$. By 2), we have

$$tr \exp\left\{\sum_{i=1}^n \log \mathbb{E}\left[\exp(\lambda X_i)\right]\right\} \leq tr \exp\left\{\sum_{i=1}^n g(\lambda)A_i\right\},$$

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^n X_i) \geq t\right) \leq e^{-\lambda t} tr \exp\left\{\sum_{i=1}^n \log\left(\mathbb{E}\left[\exp(\lambda X_i)\right]\right)\right\} \leq e^{-\lambda t} tr \exp\left\{g(\lambda)\sum_{i=1}^n A_i\right\}$$

$$(By \sum_{i=1}^n A_i \preceq \sigma^2 I) \leq e^{-\lambda t} tr \exp\left\{g(\lambda)\sigma^2 I\right\} = d \exp\left\{-\lambda t + g(\lambda)\sigma^2\right\}.$$

Since λ can take any values in $\mathbb{R}^+$, we complete the proof of Lemma 11.2. ■

Now go back to original proof. Again we assume $\mathbb{E}\left[\exp(\lambda X_i)\right] \leq \exp\left\{g(\lambda)A_i\right\}$.

Step 4: We prove Bernstein inequality in this step. We need the auxiliary result:

Lemma 11.4 *Let* $X \in \mathbb{R}^{d\times d}$ *be symmetric mean-zero, such that* $\lambda_{\max}(X) \leq 1$*, a.e. Then*

$$\mathbb{E}\left[\exp(\lambda X)\right] \preceq \exp\left\{(e^\lambda - \lambda - 1)\mathbb{E}\left[X^2\right]\right\}.$$

Proof: The function

$$f : X \in \mathbb{R} \mapsto \begin{cases} \frac{e^{\lambda x} - \lambda x - 1}{x^2}, & x \neq 0; \\ \frac{\lambda^2}{2}, & x = 0 \end{cases}$$

is increasing in x. So that $f(x) \leq f(1)$ is $x \geq 1$. By transfer rule, $f(X) \preceq f(1)I_d$.

Next,

$$\begin{aligned} \exp(\lambda X) =& I_d + \lambda X + \exp(\lambda X) - \lambda X - I_d = I_d + \lambda X + Xf(X)X \\ \preceq& I_d + \lambda X + Xf(1)I_d X = I_d + \lambda X + f(1)X^2. \end{aligned}$$

Taking expectation,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq I_d + \mathbb{E}\left[\lambda X\right] + f(1)\mathbb{E}\left[X^2\right] = I_d + f(1)\mathbb{E}\left[X^2\right].$$

Thus we complete the proof. ■

We can see $\lambda_{\max}(X_i/C) \leq 1$. Applying Lemma 11.4., for $g(\lambda) = e^\lambda - \lambda - 1$,

$$\mathbb{E}\left[\exp(\lambda X_i/C)\right] \leq \exp\left\{g(\lambda)\mathbb{E}\left[X_i^2\right]/C^2\right\}.$$

By Lemma 11.2,

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^n X_i) \geq Ct\right) \leq d \exp\left\{-\lambda t + \frac{g(\lambda)}{C^2}\sigma^2\right\}.$$

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^{n} X_i) \geq t\right) \leq d \exp\left\{-\frac{\lambda}{C}t + \frac{g(\lambda)}{C^2}\sigma^2\right\}.$$

Minimizing over $\lambda > 0$, minimum happens at $\lambda = \log(1 + Ct/\sigma^2)$, $t \geq 0$. Plug this in, we get

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^{n} X_i) \geq t\right) \leq d \exp\left\{-\frac{\sigma^2}{C^2} h(\frac{Ct}{\sigma^2})\right\},$$

where $h(\mu) = (1+\mu)\log(1+\mu) - \mu$ for $\mu > 0$. We know $h(\mu) \geq \frac{\mu^2}{2(1+\mu/3)}$, and thus

$$\mathbb{P}\left(\lambda_{\max}(\sum_{i=1}^{n} X_i) \geq t\right) \leq d \exp\left\{-\frac{t^2}{2(\sigma^2 + Ct/3)}\right\} \begin{cases} \leq d \exp\left\{-\frac{3t^2}{8\sigma^2}\right\}, & t \leq \sigma^2/C; \\ \leq d \exp\left\{-\frac{3t}{8C}\right\}, & t > \sigma^2/C. \end{cases}$$

■

11.2 Matrix Hoeffding Bound

Theorem 11.5 *(Matrix Hoeffding Bound) Let $X_1, \ldots, X_n$ be independent $d \times d$ symmetric mean-zero matrices, s.t. $X_i^2 \preceq A_i^2$ for all i and some positive-definite matrices $A_1, A_2, \ldots, A_n$. Then*

$$\mathbb{P}\left(\|\sum_{i=1}^{n} X_i\|_{op} \geq t\right) \leq 2d \exp\left\{-\frac{t^2}{8\sigma^2}\right\}, \forall t \geq 0,$$

where $\sigma^2 = \|\sum_{i=1}^{n} A_i^2\|_{op}$.

Define $X \in \mathbb{R}^{d\times d}$ mean-zero symmetric random matrix to be

1. sub-Gaussian with parameters $\Sigma \in \mathbb{R}^{d\times d}$ positive definite if

$$\mathbb{E}\left[\exp(\lambda X)\right] \preceq \exp\left\{\frac{\lambda^2}{2}\Sigma\right\}, \forall \lambda \in \mathbb{R}.$$

2. sub-exponential with parameters (N, α), $N \in \mathbb{R}^{d\times d}$ positive definite, $\alpha > 0$ if

$$\mathbb{E}\left[\exp(\lambda X)\right] \preceq \exp\left\{\frac{\lambda^2}{2}N\right\}, \forall |\lambda| < 1/\alpha.$$

(There are other ways to generalize sub-Gaussian and sub-exponential random variables. For example, a mean-zero $d \times d$ matrix is sub-Gaussian(σ^2) if $\langle A, V\rangle \in SG(\sigma^2)$ for all V with $\|V\|_F = 1$. This is a weaker condition. Here we define $\langle A, B\rangle = tr(A^T B)$, and then $\|A\|_F^2 = \langle A, A\rangle$.)

Remarks:

1. If $\text{rank}(\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]) = r \leq d$. Then we can replace d by r;
2. Extension to non-symmetric matrices: $B \in \mathbb{R}^{d_1\times d_2}$ or $B \in \mathbb{R}^{d\times d}$ non-symmetric. In this case, let $A = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$, then $A^2 = \begin{pmatrix} BB^T & 0 \\ 0 & B^T B \end{pmatrix}$, and $\|A\|_{op} = \|B\|_{op}$. Applying matrix Bernstein inequality, with $\sigma^2 = \max\left\{\|\sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^T\right]\|_{op}, \|\sum_{i=1}^{n} \mathbb{E}\left[X_i^T X_i\right]\|_{op}\right\}$, and replace d by $(d_1 + d_2)$;
3. When d is not too large, this bound can be sharp. When d is too large, the bound is loose. There are results in which d is replaced by $d_{INT}(\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right])$, where $1 \leq d_{INT}(A) \leq \frac{tr(A)}{\lambda_{\max}(A)} \leq d$ for positive definite A.

11.3 Application to Covariance Matrices

Proposition 11.6 *Let $X_1, \ldots, X_n$ be independent mean-zero random vectors in $\mathbb{R}^d$ s.t. $\|X_i\| \leq \sqrt{C}$, $\forall i$, a.e.; then*

$$\mathbb{P}\left(\|\hat{\Sigma}_n - \Sigma\|_{op} \geq t\right) \leq 2d \exp\left\{-\frac{nt^2}{2C(\|\Sigma\|_{op} + t/3)}\right\}.$$

Proof: Let $Q_i = X_i X_i^T - \Sigma$, then $\hat{\Sigma}_n - \Sigma = \frac{1}{n}\sum_{i=1}^n Q_i$. We need to:

1. Check $\|Q_i\|_{op}$ are uniformly bounded;
2. Bound $\sigma^2 = \| \sum_{i=1}^n V\left[Q_i\right] \|_{op}$.

To deal with 1):

$$\|Q_i\|_{op} = \|X_i X_i^T - \Sigma\|_{op} \leq \|X_i X_i^T\|_{op} + \|\Sigma\|_{op} = \|X_i\|^2 + \|\Sigma\|_{op} \leq 2C,$$

because

$$\begin{aligned}\|\Sigma\|_{op} = \lambda_{\max}(\Sigma) &= \max_{z \in \mathbb{S}^{d-1}} z^T \Sigma z \\ &= \max_{z \in \mathbb{S}^{d-1}} z^T \mathbb{E}\left[XX^T\right] z = \max_{z \in \mathbb{S}^{d-1}} \mathbb{E}\left[(X^T z)\right]^2 \\ &\leq \max_{z \in \mathbb{S}^{d-1}} \mathbb{E}\left[\|X\|^2 \|z\|^2\right] \leq C.\end{aligned}$$

As for 2),

$$\begin{aligned}V\left[Q_i\right] =& \mathbb{E}\left[(X_i X_i^T)^2\right] - \Sigma^2 \\ \leq& \mathbb{E}\left[(X_i X_i^T)^2\right] = \mathbb{E}\left[\|X_i\|^2 X_i X_i^T\right] \\ \leq& C\mathbb{E}\left[X_i X_i^T\right] = C\Sigma.\end{aligned}$$

Then $\|V\left[Q_i\right]\|_{op} \leq C\|\Sigma\|_{op}$. ■

For a simple application, since $\|X_i\| \leq K\sqrt{\mathbb{E}\left[\|X_i\|^2\right]} = K tr(\Sigma)$, we can always take $C = K\sqrt{d\|\Sigma\|_{op}}$. Then it can be shown that with high probability,

$$\frac{\|\hat{\Sigma}_n - \Sigma\|_{op}}{\|\Sigma\|_{op}} \leq const \max\left\{\sqrt{\frac{d}{n}\log d}, \frac{d}{n}\log d\right\}.$$