36-710: Advanced Statistical Theory **Spring 2019**

Lecture 14: February 28

Lecturer: Alessandro Rinaldo *Scribes: Kartik Gupta*

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

This lecture was dedicated to proving the Matrix Bernstein inequality in detail. The statement of Matrix Hoeffding inequality was also stated. Before discussing these statements we first state some preliminaries.

14.1 Some preliminaries on matrix calculus

Following is a list of some standard facts about symmetric $d \times d$ matrices which will be used in the proof. Wherever used, $A = U\Lambda U^T$ is the singular value decomposition of A.

1. Element-wise matrix function can be shifted to function on eigenvalues: $f(A) = Uf(\Lambda)U^T$.

2. PSD ordering: $A \preceq B$ iff $B - A \preceq 0$.

3. Transfer rule: $f, g : I \subseteq \mathbb{R} \to \mathbb{R}$ s.t. $f(x) \le g(x)\ \forall x \in I$ then $f(A) \preceq g(A)$ if the eigenvalues of A are contained in I.

4. Matrix exponentiation function: $exp(A) = I + \sum_{k=1}^{\infty} \frac{A^k}{k!} = Uexp(\Lambda)U^T$. Note that $exp(A)$ is always positive definite.

5. Matrix logarithm: The inverse function of exp on $\mathbb{S}^{++}$, the set of all positive definite matrices. Hence, $log(exp(A)) = A$.

6. Trace: $tr(A) = \sum_{i=1}^{d} A_{ii} = \sum_{i=1}^{d} \lambda_i$.

7. Trace-exponential inequality: $A \preceq B \implies tr(exp(A)) \preceq tr(exp(B))$.

8. Operator monotonicity for log: $0 \prec A \preceq B \implies \log(A) \preceq \log(B)$.

9. If A, B have their eigenvalues in $I \subseteq \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ is convex on I then f is operator convex if $f(\tau A + (1-\tau)B) \preceq \tau f(A) + (1-\tau)f(B)$ for all $\tau \in [0, 1]$.

10. Note that in general $exp(A + B) \neq exp(A)exp(B)$ but $tr(exp(A + B)) \le tr(exp(A)exp(B)))$. The latter is called the Golden-Thompson inequality which is unfortunately generally not true for addition of more than two matrices.

14.2 Matrix Bernstein Inequality

Theorem 14.1 *Let $X_1, \ldots, X_n$ be independent centered $d \times d$ symmetric random matrices such that $||X_i||_{op} \leq C$ for all i a.s. and some $C > 0$. Then*

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

Proof: The proof is divided into 4 steps.

Step 1. In this step we will bound the mgf of $S = \sum_{i=1}^{n} X_i$. Note that letting $\lambda_{max}(A) = \max_s \lambda_s(A)$ we get

$$||A||_{op} = \max\{\lambda_{max}(A), \lambda_{max}(-A)\}$$

where $\lambda_{max}(-A) = -\lambda_{min}(A)$. Thus we only need to bound $\lambda_{max}(A)$. A similar argument can be used for $\lambda_{max}(-A)$ to finish the proof.

Fix $\lambda \geq 0$ and $t \in \mathbb{R}$. Then

$$\begin{aligned} \mathbb{P}[\lambda_{max}(S) \geq t] &\leq e^{-\lambda t}\mathbb{E}[e^{\lambda \cdot \lambda_{max}(S)}] \\ &= e^{-\lambda t}\mathbb{E}[\lambda_{max}(exp\{\lambda S\})] \\ &\leq e^{-\lambda t}\mathbb{E}[tr(exp\{\lambda S\})] \end{aligned} \tag{14.1}$$

Step 2. Now,

$$\mathbb{E}[tr(exp(\lambda S))] = \mathbb{E}\left[tr\left(exp\left(\lambda \sum_{i=1}^{n-1} X_i + \lambda X_n\right)\right)\right]$$

condition on $X_1, \ldots, X_{n-1}$ and apply the Corollary 14.3 to get

$$\mathbb{E}[tr(exp(\lambda S))] \leq \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[tr\left(exp\left(\lambda \sum_{i=1}^{n-1} X_i + \log \mathbb{E}_{X_n} exp(\lambda X_n)\right)\right)\right]$$

successive application on X_{n-1} to X_1 gives us

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

Hence we have,

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

This is referred to as the **master tail bound theorem** for sums of random independent matrices.

Step 3. Now we will bound the terms $\mathbb{E}[exp(\lambda X_i)]$. Assume $C = 1$ in the theorem of the statement. By Lemma 14.5

$$\mathbb{E}[exp(\lambda X)] \preceq exp\{f(\lambda)\mathbb{E}[X^2]\}$$

where $f(\lambda) = e^{\lambda} - \lambda - 1$. Then

$$\mathbb{P}[\lambda_{max}(S) \geq t] \leq d \inf_{\lambda > 0} exp\{-\lambda t + f(\lambda)\sigma^2\}$$

where $\sigma^2 = ||\sum_{i=1}^{n} \mathbb{E}[X_i^2]||_{op}$

Step 4. Bernstein argument

The minimum is achieved at $\lambda = \log\left(1 + \frac{t}{\sigma^2}\right)$. So that

$$\begin{aligned} \mathbb{P}[\lambda_{max}(S) \geq t] &\leq dexp\left\{-\frac{\sigma^2}{C^2} h\left(\frac{Ct}{\sigma^2}\right)\right\} \\ &\leq dexp\left\{\frac{-t^2}{2\left(\sigma^2 + \frac{Ct}{3}\right)}\right\} \end{aligned} \tag{14.2}$$

here $h(\mu) = (1+\mu)\log(1+\mu) - \mu$ for all $\mu > 0$ and

$$h(\mu) \geq \frac{\mu^2}{2(1+\mu/3)}$$

■

14.3 Some useful theorems and lemmas

Theorem 14.2 *(Lieb's Theorem) Let B be a symmetric matrix. The function $tr(exp(B + \log(A)))$ defined for positive definite matrices A is operator concave.*

Corollary 14.3 *If B is a fixed $d \times d$ symmetric matrix and X is a symmetric $d \times d$ random matrix then*

$$\mathbb{E}\left[tr(exp(B + X))\right] \leq tr(exp(B + \log \mathbb{E}[exp(X)]))$$

Proof: Set $Y = exp(X) \in \mathbb{S}^{++}$. Use Jensen's inequality to get

$$\mathbb{E}[tr(exp(B + \log(Y))] \leq tr(exp(B + log\mathbb{E}[Y]))$$

■

Lemma 14.4 *Let $g : (0, \infty) \to [0, \infty)$ and $A_1, \ldots, A_n$ be $d \times d$ symmetric matrices s.t.*

$$\mathbb{E}[exp(\lambda X_i)] \preceq exp\{g(\lambda)A_i\}$$

for all i and $\lambda \geq 0$. Then,

$$\mathbb{P}\left[\lambda_{max}\left(\sum_{i=1}^{k} X_i\right) \geq t\right] \leq d \inf_{\lambda > 0} \left\{exp\left\{-\lambda t + g(\lambda)e\right\}\right\}$$

where $e = \lambda_{max}\left(\sum_{i=1}^{n} A_i\right)$.

Proof: We will use the following two properties:

1. Operator monotonicity of log: if $0 \prec A \preceq B$ then $\log(A) \preceq \log(B)$.
2. Monotonicity of $tr(exp(\cdot))$: if $A \preceq B$ then $tr(exp(A)) \preceq tr(exp(B))$

By the master tail bound theorem we have

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

We have

$$\begin{aligned} &\mathbb{E}[exp(\lambda X_i)] \preceq exp\{g(\lambda) A_i\} \\ \implies &\log(\mathbb{E}[exp(\lambda X_i)]) \preceq log(exp\{g(\lambda) A_i\}) = g(\lambda) A_i \end{aligned} \tag{14.3}$$

Hence,

$$\sum_{i=1}^{n} \log(\mathbb{E}[exp(\lambda X_i)]) \preceq g(\lambda) \sum_{i=1}^{n} A_i$$

By property 2,

$$\begin{aligned} \mathbb{P}\left[\lambda_{max}(S) \geq t\right] &\leq e^{-\lambda t} tr\left(exp\left(g\left(\lambda \sum_{i=1}^{n} A_i\right)\right)\right) \\ &\leq d e^{-\lambda t} exp\left\{g(\lambda) e\right\} \end{aligned} \tag{14.4}$$

■

Lemma 14.5 *Let X be a $d \times d$ symmetric mean 0 random matrix s.t. $||X||_{op} \leq 1$ a.s.. Then*

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

for all $\lambda > 0$.

Proof: The function

$$f(x) = \begin{cases} \dfrac{e^{\lambda x} - \lambda x - 1}{x^2} & x \neq 0 \\ \dfrac{\lambda^2}{2} & x = 0 \end{cases}$$

is increasing on $[0, \infty]$ such that $f(x) \leq f(1)$ for $x \in [0, 1]$. Then

$$exp(\lambda X) = I + \lambda X + Xf(X)X \preceq I + \lambda X + X(f(1).I)X = I + \lambda X + f(1)X^2$$

because $f(X) \preceq f(1)I$. Taking expectation on both the sides we get

$$\begin{aligned} \mathbb{E}[exp(\lambda X)] &\preceq I + \mathbb{E}[f(1)X^2] \\ &\preceq exp\{f(1)\mathbb{E}[X^2]\} = exp\{(e^{\lambda} - \lambda - 1)\mathbb{E}[X^2]\} \end{aligned} \tag{14.5}$$

where the last inequality follows from the fact that $1 + x \leq e^x$. ■

14.4 Matrix Hoeffding Inequality

Theorem 14.6 *(Matrix Hoeffding) Let $X_1, \ldots, X_n$ be independent centered $d \times d$ symmetric random s.t. $X_i^2 \preceq A_i^2$ a.e., where $A_i \in \mathbb{S}^{++}$. Then for $S = \sum X_i$*

$$\mathbb{P}[\lambda_{max}(S) \geq t] \leq d exp\left\{\frac{-t^2}{8\sigma^2}\right\}$$

where $\sigma^2 = || \sum A_i^2 ||_{op}$