36-755: Advanced Statistical Theory I **Fall 2017**

Lecture 8: September 27

Lecturer: Alessandro Rinaldo *Scribes: Xiaoyi Yang*

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

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

8.1 Continue on Matrix Bernstein Inequality

Theorem 8.1 *(Matrix Bernstein Inequality) $X_1, ..., X_n$ are independent, zero mean, $d * d$ and symmetry matrices, such that $||X_i||_{op} \leq C$, then we have*

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

where $\sigma^2 = ||\sum_{i=1}^{n} EX_i^2||_{op} = ||\sum_{i=1}^{n} Var(X_i)||_{op}$

8.1.1 Review the end of last lecture

Proof: The proof of Matrix Bernstein Inequality is following:

Step 1: Bound on moment generating function by Chernoff bound.

$$P(||\sum_{i=1}^{n} X_i||_{op} \geq t) \leq P(\lambda_{max}(\sum_{i=1}^{n} X_i) \geq t) \leq e^{-\lambda s} E[\lambda_{max}(e^{\lambda \sum_{i=1}^{n} X_i})] \leq e^{-\lambda \sum_{i=1}^{n} X_i} E[tr(e^{\lambda \sum_{i=1}^{n} X_i})]$$

Step 2: Apply monotonicity and Lieb's inequality.

Note: The tool we have used in this step and may be in later steps:

1. Operator monotonicity of log: If $0 \preceq A \preceq B$, then $log(A) \preceq log(B)$
2. Monotonicity of $tr(e)$: If $A \preceq B$, then $tr(e^A) \leq tr(e^B)$
3. Lieb's inequality

By these previous tools, we have

$$P(\lambda_{max}(\sum_{i=1}^{n} X_i) \geq t) \leq inf_{\lambda}\{e^{\lambda t} tr(\exp\{\sum_{i=1}^{n} logE[e^{\lambda X_i}]\})\}$$

8.1.2 Continue on the proof

Step 3: Bound $E[e^{\lambda X_i}]$.

Lemma 8.2 *Let a function* $g : (0, \infty) \to [0, \infty)$*, and* $A_1,...,A_n$ *be PSD matrix such that* $E[e^{\lambda X_i}] \leq \exp\{g(\lambda)A_i\}$*,* $\lambda > 0, \forall i$*. Then we have*

$$P(\lambda_{max}(\sum_{i=1}^{n} X_i) \geq t) \leq d * inf_\lambda \exp\{-\lambda t + g(\lambda) * \lambda_{max}(\sum_{i=1}^{n} A_i)\}$$

Proof: By log operator monotonicity, since $E[e^{\lambda X_i}] \leq \exp\{g(\lambda)A_i\}$, we have $log(E[e^{\lambda X_i}]) \leq g(\lambda)A_i$.

By Monotonicity of $tr(e)$, we have $tr(\exp\{\sum_{i=1}^{n} log(E[e^{\lambda X_i}])\}) \leq tr(\exp\{\sum_{i=1}^{n} g(\lambda)A_i\})$

Notice that $tr(\Sigma) \leq d * \lambda_{max}(\Sigma)$. Therefore, after extract all the constants terms, we have

$$\begin{aligned}
P(\lambda_{max}(\sum_{i=1}^{n} X_i) &\leq inf_\lambda\{e^{\lambda t} tr(\exp\{\sum_{i=1}^{n} logE[e^{\lambda X_i}]\})\} \\
&\leq inf_\lambda\{e^{\lambda t} tr(\exp\{\sum_{i=1}^{n} g(\lambda)A_i\})\} \\
&\leq inf_\lambda\{e^{\lambda t} * d * \lambda_{max}(\exp\{\sum_{i=1}^{n} g(\lambda)A_i\})\} \\
&\leq d * inf_\lambda \exp\{-\lambda t + g(\lambda) * \lambda_{max}(\sum_{i=1}^{n} A_i)\}
\end{aligned}$$

■

Then we have $P(\lambda_{max}(\sum_{i=1}^{n} X_i) \leq d * inf_\lambda \exp\{-\lambda t + g(\lambda) * ||\sum_{i=1}^{n} A_i||_{op}\}$ if we have $E[e^{\lambda X_i}] \leq \exp\{g(\lambda)A_i\}$.

Notice that if we assume that $||X_i||_{op} \leq 1$, thus let $C = 1$.

Lemma 8.3 $E[e^{\lambda X_i}] \leq \exp\{(e^\lambda - \lambda - 1)E[X_i^2]\}$

Proof: Define the function

$$f_\lambda(x) = \begin{cases} \frac{e^{\lambda x} - \lambda x - 1}{x^2} & \text{if } x \neq 0 \\ \frac{\lambda^2}{2} & \text{if } x = 0 \end{cases}$$

By taking the first derivative, we know that $f_\lambda(x)$ is increasing thus if $x \leq 1$, then $f_\lambda(x) \leq f_\lambda(1)$, thus $f_\lambda(X_i) \leq f(1) * I$.

By spectral theorem and the assumption of $||X_i||_{op}$, we have

$$e^{\lambda X_i} = I + \lambda X_i + X_i^T f_\lambda(X_i) X_i \leq I + \lambda X_i + f_\lambda(1) X_i^2$$

Take expectation for both side and use the fact that $1 + x \leq e^x$ and X_i is zero mean, we will have

$$E[e^{\lambda X}] \leq E[I + \lambda X_i + f_\lambda(1)X_i^2] = I + f_\lambda(1)E[X_i^2] \leq \exp\{f_\lambda(1)E[X_i^2]\} = \exp\{(e^\lambda - \lambda - 1)E[X_i^2]\}$$

■

Step 4: Apply two lemmas in the step 3 and warp the proof.

Let $A_i = E[X_i^2]$, and $g(\lambda) = e^\lambda - \lambda - 1$, then based on Lemma 8.3, we have

$$E[e^{\lambda X_i}] \leq \exp\{(e^\lambda - \lambda - 1)E[X_i^2]\} = \exp\{g(\lambda)A_i\}$$

Notice that $A_i = E[X_i^2]$ is PSD matrix thus we satisfied the condition of Lemma 8.2.

Apply Lemma 8.2 and notice that let $\sigma^2 = ||\sum_{i=1}^n E[X_i^2]||$, we have

$$P(\lambda_{max}(\sum_{i=1}^n X_i) \geq t) \leq d * inf_\lambda \exp\{-\lambda t + g(\lambda) * \lambda_{max}(\sum_{i=1}^n A_i)\} \leq d * inf_\lambda \exp\{-\lambda t + g(\lambda) * \sigma^2\}$$

Recall in step 3, we have the assumption $\lambda_{max}(\sum_{i=1}^n A_i) \leq 1$, but in the question we only have the condition $||X_i||_{op} \leq C$, therefore, we replace X_i as $\frac{X_i}{C}$ in the inequality and have

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

Minimize the RHS to achieve the narrowest bound by taking the derivative and set ot zero, we have $-t + (e^\lambda - 1) * \frac{\sigma^2}{C^2} = 0$, then solve the equation and $inf_\lambda = log(\frac{tC^2}{\sigma^2} + 1)$.

By plugging in the inf_λ, and let $t^* = tC$, then notice that $inf_\lambda = log(\frac{t^*C}{\sigma^2} + 1)$ now, and we will have

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

$$= d * \exp\{-\frac{t^*}{C} log(\frac{t^*C}{\sigma^2} + 1) + (\frac{t^*C}{\sigma^2} + 1 - log(\frac{t^*C}{\sigma^2} + 1) - 1)\frac{\sigma^2}{C^2}\}$$

$$= d * \exp\{\frac{t^*}{C} - (\frac{t^*}{C} + \frac{\sigma^2}{C^2}) log(\frac{t^*C}{\sigma^2} + 1)\}$$

$$= d * \exp\{\frac{\sigma^2}{C^2}[\frac{t^*C}{\sigma^2} - (\frac{t^*C}{\sigma^2} + 1) log(\frac{t^*C}{\sigma^2} + 1)]\}$$

If we let $h(u) = (1 + u)log(1 + u) - u$ and replace the t^* with t, then we have

$$P(\lambda_{max}(\sum_{i=1}^n X_i) \geq t) \leq d * \exp\{-\frac{\sigma^2}{C^2} h(u)\} \qquad \text{where } u = \frac{tC}{\sigma^2}$$

Notice that $h(u) \geq \frac{u^2}{2(1+\frac{u}{3})}$, then

$$P(\lambda_{max}(\sum_{i=1}^n X_i) \geq t) \leq d * \exp\{-\frac{\sigma^2}{C^2}\frac{u^2}{2(1+\frac{u}{3})}\} = d * \exp\{-\frac{\sigma^2}{C^2}\frac{(\frac{tC}{\sigma^2})^2}{2(1+\frac{\frac{tC}{\sigma^2}}{3})}\} = d * \exp\{-\frac{t^2}{2(\sigma^2 + \frac{tC}{3})}\}$$

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

Based on the value of t, we have

$$P(\lambda_{max}(\sum_{i=1}^n X_i) \geq t) \leq \begin{cases} d * \exp\{\frac{-3t^2}{8\sigma^2}\} & \text{if } t \leq \frac{\sigma^2}{C} \\ d * \exp\{\frac{-3t}{8C}\} & \text{if } t > \frac{\sigma^2}{C} \end{cases}$$

8.2 Extension and Remarks on Matrix Bernstein Inequality

1. There exists a Bounded in Expectation version, if all the assumption in Theorem 8.1 are satisfied, then

$$E[||\sum_{i=1}^{n} X_i||_{op} \leq C * [\sigma\sqrt{log(d)} + C * log(d)]]$$

where $\sigma = \sqrt{||\sum_{i=1}^{n} EX_i^2||}$

2. There exists a Bounded difference inequality version.

Theorem 8.4 *Let $x = (x_1, ...x_n)$ be independent random variables and there exists a function H such that $H : R^n \to R^{d*d}$. If there exists a sequence of matrix A_i such that $(H(x_1, ...x_i, ...x_n) - H(x_1, ...x_i', ...x_n)) \preceq A_i^2$ for all $i = 1, ...n$ and let $\sigma^2 = ||\sum_{i=1}^{n} A_i^2||$, then we have*

$$P(\lambda_{max}(H(x) - EH(x)) \geq t) \leq d * e^{\frac{-t^2}{8\sigma^2}}$$

3. Weakening the assumption is possible(proof in the book). The weakened Bernstein condition is

$$E[X_i^p] \leq \frac{p!}{2} C^{p-2} E[X_i^2], \qquad \text{for } p = 3, 4, ...$$

4. We define X are $d * d$, symmetry, zero mean is sub-Gaussian(Σ) matrix if $E[e^{\lambda X}] \leq \exp\{\frac{\lambda^2}{2}\Sigma\}$, for some PD matrix Σ and $\forall \lambda \in R$, or is sub-Exponential(V, α) if $E[e^{\lambda X}] \leq \exp\{\frac{\lambda^2}{2} V\}$, for some PD matrix V and $\forall |\lambda| \leq \frac{1}{\alpha}$.

If we insert this bound to the Matrix Bernstein Inequality, we will have a Hoeffding/Bernstein inequality which is

$$P(||\sum_{i=1}^{n} X_i|| \geq t) \leq 2d \exp\{-\frac{t^2}{2\sigma^2}\}$$

if $X_i \in SG(\Sigma_i)$ are independent and $\sigma^2 = ||\sum_{i=1}^{n} \Sigma_i||$

5. The theorem can be extended to no-symmetric or rectangular matrix with Jordan-WieLaudt theorem (Steward & Sum, 1990).

If B is a $d_1 * d_2$ matrix or a $d * d$ but not symmetric matrix, let A be the pilation of B such that $A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}$.

Then A will be a $(d_1+d_2)*(d_1+d_2)$ and symmetric. Since $A^2 = \begin{bmatrix} BB^T & 0 \\ 0 & B^T B \end{bmatrix}$, then A's non-zero eigenvalues are $\pm$ singular value of B and $||A|| = ||B||$. Then the matrix inequality can be re-written as,

$$P(\lambda_{max}(\sum_{i=1}^{n} B_i) \geq t) \leq (d_1 + d_2) * \exp\{-\frac{t^2}{2(\sigma^2 + \frac{tC}{3})}\}$$

where $\sigma^2 = max\{||\sum_{i=1}^{n} E[B_i B_i^T]||, ||\sum_{i=1}^{n} E[B_i^T B_i]||\}$

8.3 Application of Matrix Bernstein Inequality

8.3.1 Covariance Estimation

Theorem 8.5 *Let $X_1, ..., X_n$ are independent, zero mean vectors in R^d such that $||X_i||^2 \leq C_d, \forall i$. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$, then,*

$$P(||\hat{\Sigma} - \Sigma||_{op} \geq t) \leq 2d \exp\{-\frac{nt^2}{2C_d(||\Sigma|| + \frac{t}{3})}\}$$

Proof: Let $Q_i = X_i X_i^T - \Sigma$, then Q_i is symmetric and zero mean. Consider,

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

$$||\Sigma||_{op} = \max_{z \in S^{d-1}} z^T E[X_i X_i^T]^2 = \max_{z \in S^{d-1}} E[(z^T X_i)^2] \leq ||z||^2 ||X_i||^2 \leq 1 * C_d = C_d$$

Thus, $||Q_i||_{op} \leq 2C_d$. Since Q_i is zero mean, then,

$$EQ_i^2 = Var[Q_i] = E[(X_i X_i^T)^2] - \Sigma^2 \leq E[(X_i X_i^T)^2] = E[||X_i||^2 X_i X_i^T] = C_d * E[X_i X_i^T] = C_d \Sigma$$

Thus, $||EQ_i^2||_{op} \leq C_d ||\Sigma||_{op}$. Therefore, let $\sigma^2 = ||\sum_{i=1}^n EQ_i^2||_{op} = nC_d ||\Sigma||_{op}$ and $C = 2C_d$. By applying the Matrix Bernstein Inequality and the extension 8.3.5, we have

$$P(||\hat{\Sigma} - \Sigma||_{op} \geq t) = P(||\sum_{i=1}^n Q_i||_{op} \geq nt) \leq 2d * \exp\{-\frac{n^2 t^2}{2(\sigma^2 + \frac{ntC}{3})}\} = 2d * \exp\{-\frac{nt^2}{2C_d(||\Sigma|| + \frac{t}{3})}\}$$

■

If we assume that $||X_i|| \leq K\sqrt{E[||X_i||^2]} = K\sqrt{tr(\Sigma)} \leq K\sqrt{d||\Sigma||_{op}}$, then $C_d = K^2 d ||\Sigma||_{op}$. In this case, with high probability, we have,

$$\frac{||\hat{\Sigma} - \Sigma||_{op}}{||\Sigma||_{op}} \leq C * max\{\sqrt{\frac{d * log(d)}{n}}, \frac{d * log(d)}{n}\}$$

8.3.2 Random Graph

Let A be a $n * n$ symmetric matrix with 0 element on the diagonal and $A_{ij} \in \{0, 1\}$, for $i \neq j$. We can consider $A_{ij} \sim Bernoulli(p_{ij})$ independently. A is usually called the adjacency matrix of a graph on node $\{1, 2, ..., n\}$ such that i and j are connected if $A_{ij} = 1$, and we have $\begin{pmatrix} n \\ 2 \end{pmatrix}$ independent Bernoulli.

References

[JT2015] J. A. Tropp , "An introduction to matrix concentration inequalities," *Foundations and Trends in Machine Learning*, 8.1-2, 2015, pp. 1–230.

[SS1990] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[CR2011] F. Chung and M. Redcliffe , "On the spectra of general random graphs," *the electronic journal of combinatorics*, 18.1, 2011, pp. 215.