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

Lecture 10: October 3

Lecturer: Alessandro Rinaldo *Scribes: Riccardo Fogliato*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

10.1 From previous lecture

Theorem 10.1 *Let $(X_1, \dots, X_n) \overset{iid}{\sim} SG_d(\sigma^2)$, and $cov(X_i) = \Sigma\ \forall i = 1, \dots, n$. Then there exists some constant $C > 0$ such that*

$$\mathbb{P}\left(\left\|\hat{\Sigma} - \Sigma\right\|_{op} \geq C \max\left\{\sqrt{\frac{d + \log\frac{1}{\delta}}{n}}, \frac{d + \log\frac{1}{\delta}}{n}\right\}\right) \leq 1 - \delta.$$

Proof: Let $\{v_1, \dots, v_n\} \in \mathbb{S}^{d-1}$ be a minimal $\frac{1}{4}$ covering of $\mathbb{S}^{d-1}$. Then, letting $t \geq 0$, we obtain

$$\mathbb{P}(\|A\|_{op} \geq t) \leq \mathbb{P}(\max_j |v_j^T A v_j| \geq t/2) \leq \sum_{j=1}^n \mathbb{P}(|v_j^T A v_j| \geq t/2).$$

Next, for any $v \in \mathbb{S}^{d-1}$,

$$v^T A v = v^T(\hat{\Sigma_n} - \Sigma)v = v = v^T\left(\sum_{i=1}^n \frac{X_i X_i^T}{n} - \Sigma\right)v = \frac{1}{n}\sum_{i=1}^n \left(Z_i^2 - \mathbb{E}[Z_i^2]\right).$$

For each $j = 1, \dots, n$,

$$\mathbb{P}(|v_j^T A v_j| \geq t/2) \leq 2\exp\left\{-\frac{n}{2}\min\left\{\left(\frac{t}{22\sigma^2}\right)^2, \frac{t}{22\sigma^2}\right\}\right\}\}$$

therefore

$$\mathbb{P}\left(\|A\|_{op} \geq t\sigma^2\right) \leq 2 \cdot 9^d \exp\left\{-\frac{n}{2}\min\left\{\left(\frac{t}{22\sigma^2}\right)^2, \frac{t}{22\sigma^2}\right\}\right\}\}$$

for the RHS smaller than $\delta \in (0, 1)$, we obtain

$$\frac{t}{22} \geq \max\left\{\frac{2d\log 9}{n} + \frac{2}{n}\log\left(\frac{2}{\delta}\right), \sqrt{\frac{2d\log 9}{n} + \frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right\}$$

■

Quick extension
Assume $X_i = \Sigma^{\frac{1}{2}} Z_i$ where Z_i is positive definite (PD), $Z_i \in SG_d(1)$, and $V(Z_i) = I_d$; then $X_i \in SG_d(\|\Sigma\|_{op})$. Therefore

$$\left\| \sum_{i=1}^{n} \frac{X_i X_i^T}{n} - \Sigma \right\|_{op} = \left\| \Sigma^{\frac{1}{2}} \left(\sum_{i=1}^{n} \frac{Z_i Z_i^T}{n} - I_d \right) \Sigma^{\frac{1}{2}} \right\|_{op} \leq \left\| \sum_{i=1}^{n} \frac{Z_i Z_i^T}{n} - I_d \right\|_{op} \left\| \Sigma^{\frac{1}{2}} \right\|_{op}^2 .$$

Now the rate for $\left\| \hat{\Sigma}_n - \Sigma \right\|_{op}$ depends on $\|\Sigma\|_{op}$ instead of σ^2.

10.2 Matrix concentration inequalities

Theorem 10.2 (matrix Bernstein inequality) *Let $X_1, \ldots, X_n$ be mean-zero, independent, symmetric $d \times d$ random matrices such that $\|X_i\|_{op} \leq C$ a.e. for some $C > 0$. Then, $\forall t \geq 0$,*

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

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

Notice that for $d = 1$ we recover the usual Bernstein's inequality.
Matrix Bernstein inequality has many applications: randomised algorithms for fast SVD, sparsification and matrix subsampling, dimensionality reduction, combinatorial optimization.

Warm-up
Let A be a $d \times d$ symmetric matrix, and consider its SVD form $A = U \Lambda U^T = \sum_{i=1}^{n} \lambda_i U_i U_i^T$.
A few facts:

- if A is positive semi-definite (PDS), then $\lambda_j \geq 0$, $\forall j = 1, \ldots, d$;
- letting S^+ be the cone of PSD matrices, if $A \in S^+$, then $\alpha A \in S^+$ $\forall \alpha \geq 0$;
- if $B - A$ is PDS, then the PSD order is expressed as $A \preceq B$;
- let $f : \mathbb{R} \to \mathbb{R}$, then $f(A) = U f(\Lambda) U^T = \sum_{i=1}^{d} f(\lambda_i) u_i u_i^T$.

Remember that for two matrices A and B, $A \preceq B$ implies $\lambda_{A,i} \leq \lambda_{B,i}$ $\forall i = 1, \ldots, d$ only if they share the same eigenvectors. For instance, $A \preceq I_d \iff U \Lambda U^T \preceq U U^T$.

Examples:

- exponential: $\exp(A) = I + \sum_{i=1}^{\infty} \frac{A^i}{i!}$, which follows from the definition of function on a marix;
- exponential-logarithm: $\log(\exp(A)) = A$, ie logarithm is the inverse function of exponential. However, $\exp(\log(A)) = A$ only if $A \in S^+$;
- trace:$tr(A) = \sum_{i=1}^{d} A_i = \sum_{i=1}^{d} \lambda_i$;

- transfer function property: $f, g : I \to \mathbb{R}$ s.t. $f(x) \le g(x)\ \forall x \in I$; then $f(A) \preceq g(A)$;
- trace-exponential inequality: if $A \preceq B$, then $tr(\exp(A)) \le tr(\exp(B))$;
- logarithm is operator concave: if $0 \prec A \preceq B$, then $\log(A) \preceq \log(B)$.

Notice that $\exp(A+B) \neq \exp(A)\exp(B)$ if $AB \neq BA$.

Proof: Step I: bounding the MGF
For the symmetric $d \times d$ matrix A, $\|A\|_{op} = \max\{\lambda_{\max}(A), \lambda_{min}(A)\} = \max\{\lambda_{\max}(A), \lambda_{\max}(-A)\}$. Therefore it will be enough to bound $\lambda_{\max}$.
Set $S = \sum_{i=1}^n X_i$. Then, for $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(\lambda_{\max}(S) \ge t) &\le e^{-\lambda t}\mathbb{E}[e^{\lambda - \lambda_{\max}(S)}] \\
&= e^{-\lambda t}\mathbb{E}[\lambda_{\max}(\exp\{\lambda S\})] \\
&\le e^{-\lambda t}\mathbb{E}[tr(\exp\{\lambda S\})] \\
&= e^{-\lambda t}\mathbb{E}[tr(\exp\{\lambda \sum_{i=1}^n X_i\})]
\end{aligned}$$

Step II: Lieb's inequality
An useful fact: let B be symmetric; the function $A^+ \to tr(\exp\{B + \log(A)\})$ is concave on S^+. Therefore, letting $Y = \exp\{X\} \in S^+$, it follows that $\mathbb{E}[tr(\exp\{B + \log Y\})] \le tr(\exp\{B + \log \mathbb{E}Y\})$ by Jensen.
Back to the proof: we obtain

$$\begin{aligned}
\mathbb{E}[tr(\exp\{\lambda \sum_{i=1}^n X_i\})] &= \mathbb{E}[tr(\exp\{\lambda \sum_{i=1}^{n-1} X_i + \lambda X_n\})] \\
&= \mathbb{E}_{X_1,\dots,X_{n-1}}[\mathbb{E}_{X_n}[tr(\exp\{\lambda \sum_{i=1}^{n-1} X_i + \lambda X_n\})|X_n]] \\
&\le \mathbb{E}[tr(\exp\left\{\sum_{i=1}^{n-1} \lambda X_i + \log\left(\mathbb{E}_{X_n}[\exp\{\lambda X_n\}]\right)\right\} \\
&\le \dots \\
&\le e^{-\lambda t} tr(\exp\left\{\sum_{i=1}^n \log(\mathbb{E}[\exp\{\lambda X_i\}])\right\}.
\end{aligned}$$

Such a result is what Tropp calls the master tail bound tail bound:

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

■