CS 839 Probability and Learning in High Dimension Lecture 3 - 02/02/2022

Lecture 3: Matrix Concentration I

Lecturer: Yudong Chen *Scribe: Zaidan Wu*

In this lecture,[1] we will restate Matrix Bernstein's inequality and introduce some properties of Matrix calculus as well as Lieb's theorem as background for the proof of this inequality.

1 Notation

A quick summary of the notation.

1. **Random variables or matrices:** X, Y
2. **Specific values:** b, σ, c

For two scalars a, b, $a \lesssim b$ means that there is a constant c and $a \leq c \cdot b$.

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

If A is a $d \times d$ symmetric matrix and its eigenvalues are sorting as $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$, then we denote $\lambda_i(A)$ be the i-th largest eigenvalue of A, namely $\lambda_i(A) = \lambda_i$.

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

2 Statement

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

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

Then for every $t \geq 0$, we have

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

Remark The upper bound in equation (1) implies that

$$\mathbb{P}\left[\left\| \sum_{i=1}^{n} X_i \right\|_{\mathrm{op}} \geq t \right] \leq (d_1 + d_2) \cdot \exp\left(-c \cdot \min\left[\frac{t^2}{\sigma^2}, \frac{t}{b} \right] \right),$$

[1] *Reading:* Section 6 in J. Tropp, An Introduction to Matrix Concentration Inequalities. `https://arxiv.org/abs/1501.01571`

where c is constant that should be independent of every parameter in the statement. This upper bound indicates that: if t is large such that $\frac{t^2}{\sigma^2} \geq \frac{t}{b}$, then the upper bound is proportional to $\exp\left(-\frac{t}{b}\right)$ which is called the sub-exponential tail, in which case the random variable $\left\|\sum_{i=1}^{n} X_i\right\|_{\text{op}}$ has a density whose tail is not heavier than exponential random variable. Otherwise, if t is small, the upper bound is proportional to $\exp\left(-\frac{t^2}{\sigma^2}\right)$ which is called the sub-Gaussian tail and the density is not heavier than normal random variable.

Remark This upper bound implies so-called user-friendly form of inequality, which is obtained by choosing an appropriate value of t such that with probability at least $1-(d_1+d_2)^{-10}$:

$$\left\|\sum_{i=1}^{n} X_i\right\|_{\text{op}} \lesssim \sqrt{\sigma^2 \log(d_1+d_2)} + b\log(d_1+d_2).$$

However, one drawback of this inequality is the logarithm factor depending on the dimensions of matrices.

The rest of this lecture is devoted to the proof the matrix Bernstein inequality.

3 Matrix Theory Background

The proof of scalar Bernstein's inequality uses the Laplace transform of random variable. In order to draw an analogy the proof to the matrix Bernstein's inequality, we define the functions of matrices as following.

Definition 1 (Matrix Functions)**.** *For symmetric matrix $A \in R^{d\times d}$, its eigen-decomposition is*

$$A = U\Lambda U^{\mathrm{T}},$$

where U is orthogonal matrix, that is $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I$ and Λ is diagonal matrix:

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}.$$

Then for a real value function f, we define

$$f(A) = Uf(\Lambda)U^{\mathrm{T}},$$

where

$$f(\Lambda) \triangleq \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix}.$$

Example 2. Consider function $f(a) = c_0 + c_1 a + c_2 a^2 + \dots$ and symmetric matrix A, since $A^k = U\Lambda^k U^{\mathrm{T}}$, we have

$$f(A) = c_0 + c_1 A + c_2 A^2 + \dots.$$

Example 3 (Matrix exponential)**.** Let $f(x) = e^x = 1 + \sum\limits_{k=1}^{\infty} \frac{1}{k!}x^k$, then exponential function of symmetric matrix $A \in \mathbb{R}^{d\times d}$ by following Definition 1 is

$$e^A = U \begin{bmatrix} e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_d} \end{bmatrix} U^{\mathrm{T}} = I + \sum_{k=1}^{\infty} \frac{1}{k!} A^k.$$

Monotonicity holds: $H \succeq A \Rightarrow \mathsf{tr}\left(e^H\right) \geq \mathsf{tr}\left(e^A\right)$ if , which is directly proven by the definition of positive-semidefinite ordering and matrix functions.

Example 4 (Matrix logarithm)**.** The logarithm of positive definite matrix is defined via setting $f(x) = \log(x)$. In addition, for any symmetric matrix A we have

$$\log(e^A) = A.$$

Matrix logarithm satisfies the *operator monotonicity* property: $H \succeq A \succeq 0 \Rightarrow \log(H) \succeq \log(A)$. The proof of this property requires some non-trivial work.

Definition 2 (Matrix MGF and CGF)**.** *For $\theta \in \mathbb{R}$ and random matrix X, the matrix moment generating function (MGF) is*

$$M_X(\theta) \triangleq \mathbb{E}\left[e^{\theta X}\right].$$

The cumulant generating function (CGF) is

$$K_X(\theta) \triangleq \log\left[\mathbb{E}\left[e^{\theta X}\right]\right].$$

For a scalar random variable Y and any $\theta > 0$, by Markov's inequality:

$$\mathbb{P}(Y \geq t) = \mathbb{P}\left(e^{\theta Y} \geq e^{\theta t}\right) \leq \inf_{\theta>0} e^{-\theta t} \cdot \mathbb{E}(e^{\theta Y}).$$

Similarly, Lemma 1 is a matrix version of this inequality.

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

$$\mathbb{P}\left(\lambda_{\max}(Y) \geq t\right) \leq \inf_{\theta>0} e^{-\theta t} \cdot \mathbb{E}\left[\mathsf{tr}\left(e^{\theta Y}\right)\right], \tag{2}$$

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

Proof Note that $e^{\theta\lambda_{\max}(Y)} = \lambda_{\max}\left(e^{\theta}Y\right) \leq \sum_i \lambda_i\left(e^{\theta}Y\right) = \mathsf{tr}\left(e^{\theta Y}\right)$ since eigenvalues of $e^{\theta Y}$ are positive. Thus for any $\theta > 0$, we have

$$\begin{aligned} &\mathbb{P}\left(\lambda_{\max}(Y) \geq t\right) \\ =&\mathbb{P}\left(e^{\theta\lambda_{\max}(Y)} \geq e^{\theta t}\right) \\ \leq&\mathbb{E}\left[e^{\theta\lambda_{\max}(Y)}\right] \cdot e^{-\theta t} \\ =&\mathbb{E}\left[\lambda_{\max}(e^{\theta Y})\right] \cdot e^{-\theta t} \\ \leq&\mathbb{E}\left[\mathsf{tr}\left(e^{\theta Y}\right)\right] \cdot e^{-\theta t}. \end{aligned}$$

Taking infimum over $\theta > 0$, the inequality is proven as desired. $\square$

4 Lieb's theorem

The MGF of sum of independent *scalar* random variables is equal to the product of MGF of random variables. That is, for independent scalar random variables $X_1, \ldots, X_n$,

$$\mathbb{E}\left(e^{\theta(X_1+\ldots X_n)}\right) = \mathbb{E}\left(e^{\theta X_1}\right)\cdots\mathbb{E}\left(e^{\theta X_n}\right). \tag{3}$$

However, for matrix version, we don't have this perfectly useful tool for our proof. In general, for two matrices Y_1, Y_2,

$$e^{Y_1+Y_2} \neq e^{Y_1}\cdot e^{Y_2}$$

unless Y_1 and Y_2 commute.

Fortunately, the Lieb's theorem helps us to overcome this challenge. Recall that a function f is concave iff for any x, y in the domain of f and any $\alpha \in [0,1]$, $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$.

Theorem 5 (Lieb's theorem)**.** *For a fixed symmetric matrix H, the function of A, $f(A)$ defined through*

$$f(A) \triangleq \mathsf{tr}\left(\exp\left(H + \log(A)\right)\right)$$

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

The Lieb's theorem is a deep result. See the referenced paper for a proof. In this lecture, we take this theorem for granted.

Using the Lieb's theorem, we can get a generalization of equation (3).

Lemma 2. *Suppose $X_1, \ldots, X_n$ are independent symmetric matrices, then*

$$\mathsf{tr}\exp\left(K_{\left(\sum_{i=1}^{N} X_i\right)}(\theta)\right) = \mathbb{E}\left[\mathsf{tr}\exp\left(\theta\sum_{i=1}^{n} X_i\right)\right] \leq \mathsf{tr}\exp\left(\sum_{i=1}^{n}\log\mathbb{E}\left[e^{\theta X_i}\right]\right) = \mathsf{tr}\exp\left(\sum_{i=1}^{n} K_{X_i}(\theta)\right) \tag{4}$$