

Matrix concentration inequalities



Yuxin Chen

Princeton University, Fall 2020

Recap: matrix Bernstein inequality

Consider a sequence of independent random matrices $\{\mathbf{X}_l \in \mathbb{R}^{d_1 \times d_2}\}$

- $\mathbb{E}[\mathbf{X}_l] = \mathbf{0}$
- $\|\mathbf{X}_l\| \leq B$ for each l
- variance statistic:

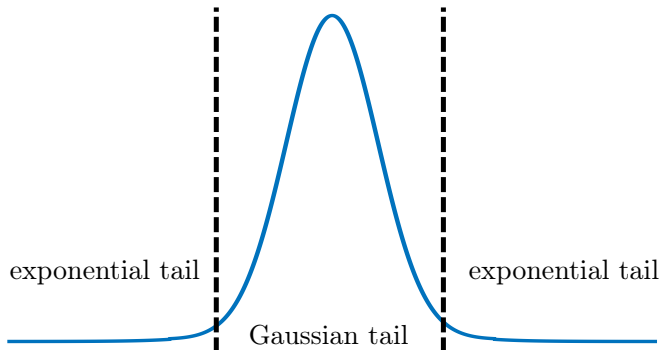
$$v := \max \left\{ \left\| \mathbb{E} \left[\sum_l \mathbf{X}_l \mathbf{X}_l^\top \right] \right\|, \left\| \mathbb{E} \left[\sum_l \mathbf{X}_l^\top \mathbf{X}_l \right] \right\| \right\}$$

Theorem 3.1 (Matrix Bernstein inequality)

For all $\tau \geq 0$,

$$\mathbb{P} \left\{ \left\| \sum_l \mathbf{X}_l \right\| \geq \tau \right\} \leq (d_1 + d_2) \exp \left(\frac{-\tau^2/2}{v + B\tau/3} \right)$$

Recap: matrix Bernstein inequality



This lecture: detailed introduction of matrix Bernstein

An introduction to matrix concentration inequalities
— Joel Tropp '15

Outline

- Matrix theory background
- Matrix Laplace transform method
- Matrix Bernstein inequality
- Application: random features

Matrix theory background

Matrix function

Suppose the eigendecomposition of a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is

$$\mathbf{A} = \mathbf{U} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \mathbf{U}^\top$$

Then we can define

$$f(\mathbf{A}) := \mathbf{U} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix} \mathbf{U}^\top$$

Examples of matrix functions

- Let $f(a) = c_0 + \sum_{k=1}^{\infty} c_k a^k$, then

$$f(\mathbf{A}) := c_0 \mathbf{I} + \sum_{k=1}^{\infty} c_k \mathbf{A}^k$$

- **matrix exponential:** $e^{\mathbf{A}} := \mathbf{I} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{A}^k$ (why?)
 - monotonicity: if $\mathbf{A} \preceq \mathbf{H}$, then $\text{tr } e^{\mathbf{A}} \leq \text{tr } e^{\mathbf{H}}$
- **matrix logarithm:** $\log(e^{\mathbf{A}}) := \mathbf{A}$
 - monotonicity: if $\mathbf{0} \preceq \mathbf{A} \preceq \mathbf{H}$, then $\log \mathbf{A} \preceq \log(\mathbf{H})$

Matrix moments and cumulants

Let \mathbf{X} be a random symmetric matrix. Then

- **matrix moment generating function (MGF):**

$$M_{\mathbf{X}}(\theta) := \mathbb{E}[e^{\theta \mathbf{X}}]$$

- **matrix cumulant generating function (CGF):**

$$\Xi_{\mathbf{X}}(\theta) := \log \mathbb{E}[e^{\theta \mathbf{X}}]$$

Matrix Laplace transform method

Matrix Laplace transform

A key step for a scalar random variable Y : by Markov's inequality,

$$\mathbb{P}\{Y \geq t\} \leq \inf_{\theta > 0} e^{-\theta t} \mathbb{E}[e^{\theta Y}]$$

This can be generalized to the matrix case

Matrix Laplace transform

Lemma 3.2

Let \mathbf{Y} be a random symmetric matrix. For all $t \in \mathbb{R}$,

$$\mathbb{P}\{\lambda_{\max}(\mathbf{Y}) \geq t\} \leq \inf_{\theta > 0} e^{-\theta t} \mathbb{E}[\text{tr } e^{\theta \mathbf{Y}}]$$

- can control the extreme eigenvalues of \mathbf{Y} via the trace of the matrix MGF

Proof of Lemma 3.2

For any $\theta > 0$,

$$\begin{aligned}\mathbb{P}\{\lambda_{\max}(\mathbf{Y}) \geq t\} &= \mathbb{P}\{e^{\theta\lambda_{\max}(\mathbf{Y})} \geq e^{\theta t}\} \\ &\leq \frac{\mathbb{E}[e^{\theta\lambda_{\max}(\mathbf{Y})}]}{e^{\theta t}} && \text{(Markov's inequality)} \\ &= \frac{\mathbb{E}[e^{\lambda_{\max}(\theta\mathbf{Y})}]}{e^{\theta t}} \\ &= \frac{\mathbb{E}[\lambda_{\max}(e^{\theta\mathbf{Y}})]}{e^{\theta t}} && (e^{\lambda_{\max}(\mathbf{Z})} = \lambda_{\max}(e^{\mathbf{Z}})) \\ &\leq \frac{\mathbb{E}[\text{tr } e^{\theta\mathbf{Y}}]}{e^{\theta t}}\end{aligned}$$

This completes the proof since it holds for any $\theta > 0$

Issues of the matrix MGF

The Laplace transform method is effective for controlling an independent sum when MGF decomposes

- in the scalar case where $X = X_1 + \cdots + X_n$ with independent $\{X_l\}$:

$$M_X(\theta) = \mathbb{E}[e^{\theta X_1 + \cdots + \theta X_n}] = \mathbb{E}[e^{\theta X_1}] \cdots \mathbb{E}[e^{\theta X_n}] = \underbrace{\prod_{l=1}^n M_{X_l}(\theta)}_{\text{look at each } X_l \text{ separately}}$$

Issues in the matrix settings:

$$e^{\mathbf{X}_1 + \mathbf{X}_2} \neq e^{\mathbf{X}_1} e^{\mathbf{X}_2} \quad \text{unless } \mathbf{X}_1 \text{ and } \mathbf{X}_2 \text{ commute}$$

$$\text{tr } e^{\mathbf{X}_1 + \cdots + \mathbf{X}_n} \not\leq \text{tr } e^{\mathbf{X}_1} e^{\mathbf{X}_1} \cdots e^{\mathbf{X}_n}$$

Subadditivity of the matrix CGF

Fortunately, the matrix CGF satisfies certain subadditivity rules, allowing us to decompose independent matrix components

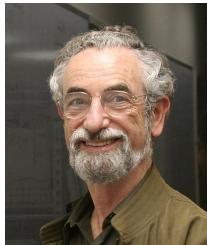
Lemma 3.3

Consider a finite sequence $\{\mathbf{X}_l\}_{1 \leq l \leq n}$ of independent random symmetric matrices. Then for any $\theta \in \mathbb{R}$,

$$\underbrace{\mathbb{E} \left[\text{tr} e^{\theta \sum_l \mathbf{X}_l} \right]}_{\text{tr exp} \left(\Xi_{\sum_l \mathbf{X}_l}(\theta) \right)} \leq \underbrace{\text{tr exp} \left(\sum_l \log \mathbb{E} [e^{\theta \mathbf{X}_l}] \right)}_{\text{tr exp} \left(\sum_l \Xi_{\mathbf{X}_l}(\theta) \right)}$$

- this is a deep result — based on Lieb's Theorem!

Lieb's Theorem



Elliott Lieb

Theorem 3.4 (Lieb '73)

Fix a symmetric matrix \mathbf{H} . Then

$$\mathbf{A} \mapsto \operatorname{tr} \exp(\mathbf{H} + \log \mathbf{A})$$

is concave on positive-semidefinite cone

Lieb's Theorem immediately implies (exercise: Jensen's inequality)

$$\mathbb{E}[\operatorname{tr} \exp(\mathbf{H} + \mathbf{X})] \leq \operatorname{tr} \exp(\mathbf{H} + \log \mathbb{E}[e^{\mathbf{X}}]) \quad (3.1)$$

Proof of Lemma 3.3

$$\begin{aligned}\mathbb{E}[\mathrm{tr} e^{\theta \sum_l \mathbf{X}_l}] &= \mathbb{E}[\mathrm{tr} \exp(\theta \sum_{l=1}^{n-1} \mathbf{X}_l + \theta \mathbf{X}_n)] \\ &\leq \mathbb{E}\left[\mathrm{tr} \exp\left(\theta \sum_{l=1}^{n-1} \mathbf{X}_l + \log \mathbb{E}[e^{\theta \mathbf{X}_n}]\right)\right] \quad (\text{by (3.1)}) \\ &\leq \mathbb{E}\left[\mathrm{tr} \exp\left(\theta \sum_{l=1}^{n-2} \mathbf{X}_l + \log \mathbb{E}[e^{\theta \mathbf{X}_{n-1}}] + \log \mathbb{E}[e^{\theta \mathbf{X}_n}]\right)\right] \\ &\leq \dots \\ &\leq \mathrm{tr} \exp\left(\sum_{l=1}^n \log \mathbb{E}[e^{\theta \mathbf{X}_l}]\right)\end{aligned}$$

Master bounds

Combining the Laplace transform method with the subadditivity of CGF yields:

Theorem 3.5 (Master bounds for sum of independent matrices)

Consider a finite sequence $\{\mathbf{X}_l\}$ of independent random symmetric matrices. Then

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_l \mathbf{X}_l \right) \geq t \right\} \leq \inf_{\theta > 0} \frac{\text{tr} \exp \left(\sum_l \log \mathbb{E}[e^{\theta \mathbf{X}_l}] \right)}{e^{\theta t}}$$

- this is a general result underlying the proofs of the matrix Bernstein inequality and beyond (e.g. matrix Chernoff)

Matrix Bernstein inequality

Matrix CGF

$$\mathbb{P} \left\{ \lambda_{\max} \left(\sum_l \mathbf{X}_l \right) \geq t \right\} \leq \inf_{\theta > 0} \frac{\text{tr} \exp \left(\sum_l \log \mathbb{E}[e^{\theta \mathbf{X}_l}] \right)}{e^{\theta t}}$$

To invoke the master bound, one needs to control the matrix CGF
main step for proving matrix Bernstein

Symmetric case

Consider a sequence of independent random symmetric matrices $\{\mathbf{X}_l \in \mathbb{R}^{d \times d}\}$

- $\mathbb{E}[\mathbf{X}_l] = \mathbf{0}$
- $\lambda_{\max}(\mathbf{X}_l) \leq B$ for each l
- variance statistic: $v := \|\mathbb{E}[\sum_l \mathbf{X}_l^2]\|$

Theorem 3.6 (Matrix Bernstein inequality: symmetric case)

For all $\tau \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_l \mathbf{X}_l\right) \geq \tau\right\} \leq d \exp\left(\frac{-\tau^2/2}{v + B\tau/3}\right)$$

Bounding matrix CGF

For bounded random matrices, one can control the matrix CGF as follows:

Lemma 3.7

Suppose $\mathbb{E}[\mathbf{X}] = \mathbf{0}$ and $\lambda_{\max}(\mathbf{X}) \leq B$. Then for $0 < \theta < 3/B$,

$$\log \mathbb{E}[e^{\theta \mathbf{X}}] \preceq \frac{\theta^2/2}{1 - \theta B/3} \mathbb{E}[\mathbf{X}^2]$$

Proof of Theorem 3.6

Let $g(\theta) := \frac{\theta^2/2}{1-\theta B/3}$, then it follows from the master bound that

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}\left(\sum_i \mathbf{X}_i\right) \geq t\right\} &\leq \inf_{\theta>0} \frac{\mathrm{tr} \exp\left(\sum_{i=1}^n \log \mathbb{E}[e^{\theta \mathbf{X}_i}]\right)}{e^{\theta t}} \\ &\stackrel{\text{Lemma 3.7}}{\leq} \inf_{0<\theta<3/B} \frac{\mathrm{tr} \exp\left(g(\theta) \sum_{i=1}^n \mathbb{E}[\mathbf{X}_i^2]\right)}{e^{\theta t}} \\ &\leq \inf_{0<\theta<3/B} \frac{d \exp(g(\theta)v)}{e^{\theta t}}\end{aligned}$$

Taking $\theta = \frac{t}{v+Bt/3}$ and simplifying the above expression, we establish matrix Bernstein

Proof of Lemma 3.7

Define $f(x) = \frac{e^{\theta x} - 1 - \theta x}{x^2}$, then for any \mathbf{X} with $\lambda_{\max}(\mathbf{X}) \leq B$:

$$\begin{aligned} e^{\theta \mathbf{X}} &= \mathbf{I} + \theta \mathbf{X} + (e^{\theta \mathbf{X}} - \mathbf{I} - \theta \mathbf{X}) = \mathbf{I} + \theta \mathbf{X} + \mathbf{X} \cdot f(\mathbf{X}) \cdot \mathbf{X} \\ &\preceq \mathbf{I} + \theta \mathbf{X} + f(B) \cdot \mathbf{X}^2 \end{aligned}$$

In addition, we note an elementary inequality: for any $0 < \theta < 3/B$,

$$\begin{aligned} f(B) &= \frac{e^{\theta B} - 1 - \theta B}{B^2} = \frac{1}{B^2} \sum_{k=2}^{\infty} \frac{(\theta B)^k}{k!} \leq \frac{\theta^2}{2} \sum_{k=2}^{\infty} \frac{(\theta B)^{k-2}}{3^{k-2}} = \frac{\theta^2/2}{1 - \theta B/3} \\ \implies e^{\theta \mathbf{X}} &\preceq \mathbf{I} + \theta \mathbf{X} + \frac{\theta^2/2}{1 - \theta B/3} \cdot \mathbf{X}^2 \end{aligned}$$

Since \mathbf{X} is zero-mean, one further has

$$\mathbb{E}[e^{\theta \mathbf{X}}] \preceq \mathbf{I} + \frac{\theta^2/2}{1 - \theta B/3} \mathbb{E}[\mathbf{X}^2] \preceq \exp\left(\frac{\theta^2/2}{1 - \theta B/3} \mathbb{E}[\mathbf{X}^2]\right)$$

Application: random features

Kernel trick

A modern idea in machine learning: replace the inner product by kernel evaluation (i.e. certain similarity measure)

Advantage: work beyond the Euclidean domain via task-specific similarity measures

Similarity measure

Define the similarity measure Φ

- $\Phi(\mathbf{x}, \mathbf{x}) = 1$
- $|\Phi(\mathbf{x}, \mathbf{y})| \leq 1$
- $\Phi(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{y}, \mathbf{x})$

Example: angular similarity

$$\Phi(\mathbf{x}, \mathbf{y}) = \frac{2}{\pi} \arcsin \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = 1 - \frac{2\angle(\mathbf{x}, \mathbf{y})}{\pi}$$

Kernel matrix

Consider N data points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d$. Then the kernel matrix $\mathbf{G} \in \mathbb{R}^{N \times N}$ is

$$G_{i,j} = \Phi(\mathbf{x}_i, \mathbf{x}_j) \quad 1 \leq i, j \leq N$$

- Kernel Φ is said to be positive semidefinite if $\mathbf{G} \succeq \mathbf{0}$ for any $\{\mathbf{x}_i\}$

Challenge: kernel matrices are usually large

- cost of constructing \mathbf{G} is $O(dN^2)$

Question: can we approximate \mathbf{G} more efficiently?

Random features

Introduce a random variable w and a feature map ψ such that

$$\Phi(x, y) = \mathbb{E}_w [\underbrace{\psi(x; w) \cdot \psi(y; w)}_{\text{decouple } x \text{ and } y}]$$

- **example (angular similarity)**

$$\underbrace{\Phi(x, y) = 1 - \frac{2\angle(x, y)}{\pi}}_{\text{Grothendieck's identity}} = \mathbb{E}_w [\text{sgn}\langle x, w \rangle \cdot \text{sgn}\langle y, w \rangle] \quad (3.2)$$

with w uniformly drawn from the unit sphere

Random features

Introduce a random variable w and a feature map ψ such that

$$\Phi(x, y) = \mathbb{E}_w \underbrace{[\psi(x; w) \cdot \psi(y; w)]}_{\text{decouple } x \text{ and } y}$$

- this results in a **random feature vector**

$$z = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} \psi(x_1; w) \\ \vdots \\ \psi(x_N; w) \end{bmatrix}$$

- $\underbrace{zz^\top}_{\text{rank 1}}$ is an **unbiased estimate** of G , i.e. $G = \mathbb{E}[zz^\top]$

Example

Angular similarity:

$$\begin{aligned}\Phi(\boldsymbol{x}, \boldsymbol{y}) &= 1 - \frac{2\angle(\boldsymbol{x}, \boldsymbol{y})}{\pi} \\ &= \mathbb{E}_{\boldsymbol{w}} [\text{sign}\langle \boldsymbol{x}, \boldsymbol{w} \rangle \text{sign}\langle \boldsymbol{y}, \boldsymbol{w} \rangle]\end{aligned}$$

where \boldsymbol{w} is uniformly drawn from the unit sphere

As a result, the random feature map is $\psi(\boldsymbol{x}, \boldsymbol{w}) = \text{sign}\langle \boldsymbol{x}, \boldsymbol{w} \rangle$

Random feature approximation

Generate n independent copies of $\mathbf{R} = \mathbf{z}\mathbf{z}^\top$, i.e. $\{\mathbf{R}_l\}_{1 \leq l \leq n}$

Estimator of the kernel matrix \mathbf{G} :

$$\hat{\mathbf{G}} = \frac{1}{n} \sum_{l=1}^n \mathbf{R}_l$$

Question: how many random features are needed to guarantee accurate estimation?

Statistical guarantees for random feature approximation

Consider the angular similarity example (3.2):

- To begin with,

$$\begin{aligned}\mathbb{E}[\mathbf{R}_l^2] &= \mathbb{E}[\mathbf{z}\mathbf{z}^\top \mathbf{z}\mathbf{z}^\top] = N\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = N\mathbf{G} \\ \implies v &= \left\| \frac{1}{n^2} \sum_{l=1}^n \mathbb{E}[\mathbf{R}_l^2] \right\| = \frac{N}{n} \|\mathbf{G}\|\end{aligned}$$

- Next, $\frac{1}{n} \|\mathbf{R}\| = \frac{1}{n} \|\mathbf{z}\|_2^2 = \frac{N}{n} =: B$
- Applying the matrix Bernstein inequality yields: with high prob.

$$\begin{aligned}\|\hat{\mathbf{G}} - \mathbf{G}\| &\lesssim \sqrt{v \log N} + B \log N \lesssim \sqrt{\frac{N}{n} \|\mathbf{G}\| \log N} + \frac{N}{n} \log N \\ &\lesssim \sqrt{\frac{N}{n} \underbrace{\|\mathbf{G}\|}_{\geq 1} \log N} \quad \text{(for sufficiently large } n\text{)}\end{aligned}$$

Sample complexity

Define the intrinsic dimension of \mathbf{G} as

$$\text{intdim}(\mathbf{G}) = \frac{\text{tr}\mathbf{G}}{\|\mathbf{G}\|} = \frac{N}{\|\mathbf{G}\|}$$

If $n \gtrsim \varepsilon^{-2} \text{intdim}(\mathbf{G}) \log N$, then we have

$$\frac{\|\hat{\mathbf{G}} - \mathbf{G}\|}{\|\mathbf{G}\|} \leq \varepsilon$$

Reference

- "*An introduction to matrix concentration inequalities*," J. Tropp, *Foundations and Trends in Machine Learning*, 2015.
- "*Convex trace functions and the Wigner-Yanase-Dyson conjecture*," E. Lieb, *Advances in Mathematics*, 1973.
- "*User-friendly tail bounds for sums of random matrices*," J. Tropp, *Foundations of computational mathematics*, 2012.
- "*Random features for large-scale kernel machines*," A. Rahimi, B. Recht, *Neural Information Processing Systems*, 2008.