Matrix concentration inequalities

Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

Recap: matrix Bernstein inequality

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

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

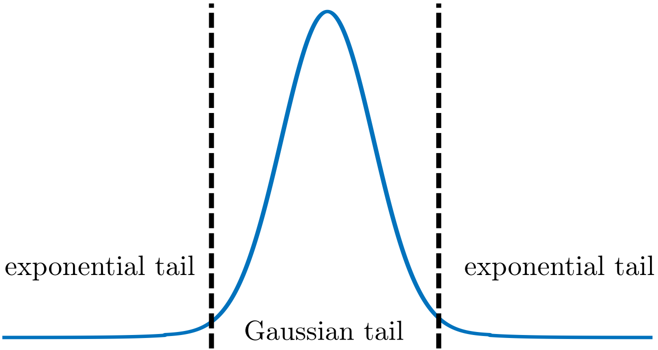
$$v := \max\left\{\left\|\mathbb{E}\left[\sum_l \boldsymbol{X}_l \boldsymbol{X}_l^\top\right]\right\|, \left\|\mathbb{E}\left[\sum_l \boldsymbol{X}_l^\top \boldsymbol{X}_l\right]\right\|\right\}$$

Theorem 3.1 (Matrix Bernstein inequality)

For all $\tau \geq 0$,

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

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 $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is

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

Then we can define

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

Examples of matrix functions

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

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

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

- **matrix logarithm:** $\log(\mathrm{e}^{\boldsymbol{A}}) := \boldsymbol{A}$
  - monotonicity: if $\boldsymbol{0} \preceq \boldsymbol{A} \preceq \boldsymbol{H}$, then $\log \boldsymbol{A} \preceq \log(\boldsymbol{H})$

Matrix moments and cumulants

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

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

$$\boldsymbol{M}_{\boldsymbol{X}}(\theta) := \mathbb{E}[e^{\theta \boldsymbol{X}}]$$

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

$$\boldsymbol{\Xi}_{\boldsymbol{X}}(\theta) := \log \mathbb{E}[e^{\theta \boldsymbol{X}}]$$

Matrix Laplace transform method

Matrix Laplace transform

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

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

This can be generalized to the matrix case

Matrix Laplace transform

Lemma 3.2

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

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

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

Proof of Lemma 3.2

For any $\theta > 0$,

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

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 $\{X_l\}$:

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

Issues in the matrix settings:

$$\mathrm{e}^{\boldsymbol{X}_1 + \boldsymbol{X}_2} \neq \mathrm{e}^{\boldsymbol{X}_1} \mathrm{e}^{\boldsymbol{X}_2} \qquad \text{unless } \boldsymbol{X}_1 \text{ and } \boldsymbol{X}_2 \text{ commute}$$

$$\operatorname{tr} \mathrm{e}^{\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n} \nleq \operatorname{tr} \mathrm{e}^{\boldsymbol{X}_1} \mathrm{e}^{\boldsymbol{X}_1} \cdots \mathrm{e}^{\boldsymbol{X}_n}$$

Subadditivity of the matrix CGF

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

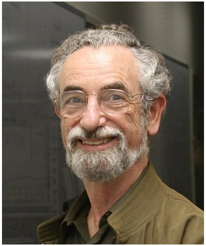
Lemma 3.3

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

$$\underbrace{\mathbb{E}\left[\operatorname{tr} e^{\theta \sum_l \boldsymbol{X}_l}\right]}_{\operatorname{tr}\exp\left(\boldsymbol{\Xi}_{\sum_l \boldsymbol{X}_l}(\theta)\right)} \le \underbrace{\operatorname{tr}\exp\left(\sum\nolimits_l \log \mathbb{E}\left[e^{\theta \boldsymbol{X}_l}\right]\right)}_{\operatorname{tr}\exp\left(\sum_l \boldsymbol{\Xi}_{\boldsymbol{X}_l}(\theta)\right)}$$

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

Lieb's Theorem



Elliott Lieb

Theorem 3.4 (Lieb '73)

Fix a symmetric matrix $\boldsymbol{H}$*. Then*

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

is concave on positive-semidefinite cone

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

$$\mathbb{E}\left[\operatorname{tr}\exp(\boldsymbol{H} + \boldsymbol{X})\right] \leq \operatorname{tr}\exp\left(\boldsymbol{H} + \log \mathbb{E}\left[e^{\boldsymbol{X}}\right]\right) \tag{3.1}$$

Proof of Lemma 3.3

$$
\begin{aligned}
\mathbb{E}\left[\operatorname{tr} e^{\theta \sum_l \boldsymbol{X}_l}\right] &= \mathbb{E}\left[\operatorname{tr} \exp\left(\theta \sum_{l=1}^{n-1} \boldsymbol{X}_l + \theta \boldsymbol{X}_n\right)\right] \\
&\leq \mathbb{E}\Big[\operatorname{tr} \exp\Big(\theta \sum_{l=1}^{n-1} \boldsymbol{X}_l + \log \mathbb{E}\big[e^{\theta \boldsymbol{X}_n}\big]\Big)\Big] \qquad \text{(by (3.1))} \\
&\leq \mathbb{E}\Big[\operatorname{tr} \exp\Big(\theta \sum_{l=1}^{n-2} \boldsymbol{X}_l + \log \mathbb{E}\big[e^{\theta \boldsymbol{X}_{n-1}}\big] + \log \mathbb{E}\big[e^{\theta \boldsymbol{X}_n}\big]\Big)\Big] \\
&\leq \cdots \\
&\leq \operatorname{tr} \exp\Big(\sum_{l=1}^{n} \log \mathbb{E}\big[e^{\theta \boldsymbol{X}_l}\big]\Big)
\end{aligned}
$$

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 $\{\boldsymbol{X}_l\}$ of independent random symmetric matrices. Then

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

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

Matrix Bernstein inequality

Matrix CGF

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

To invoke the master bound, one needs to $\underbrace{\text{control the matrix CGF}}_{\text{main step for proving matrix Bernstein}}$

Symmetric case

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

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

Theorem 3.6 (Matrix Bernstein inequality: symmetric case)

For all $\tau \geq 0$,

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

Bounding matrix CGF

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

Lemma 3.7

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

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

Proof of Theorem 3.6

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

$$\mathbb{P}\left\{\lambda_{\max}\big(\sum_i \boldsymbol{X}_i\big) \geq t\right\} \leq \inf_{\theta>0} \frac{\operatorname{tr}\exp\big(\sum_{i=1}^n \log \mathbb{E}[\mathrm{e}^{\theta \boldsymbol{X}_i}]\big)}{\mathrm{e}^{\theta t}}$$

$$\overset{\text{Lemma 3.7}}{\leq} \inf_{0<\theta<3/B} \frac{\operatorname{tr}\exp\left(g(\theta)\sum_{i=1}^n \mathbb{E}[\boldsymbol{X}_i^2]\right)}{\mathrm{e}^{\theta t}}$$

$$\leq \inf_{0<\theta<3/B} \frac{d\,\exp\left(g(\theta)v\right)}{\mathrm{e}^{\theta t}}$$

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 $\boldsymbol{X}$ with $\lambda_{\max}(\boldsymbol{X}) \leq B$:

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

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

$$f(B) = \frac{e^{\theta B} - 1 - \theta B}{B^2} = \frac{1}{B^2} \sum_{k=2}^{\infty} \frac{(\theta B)^k}{k!} \leq \frac{\theta^2}{2} \sum_{k=2}^{\infty} \frac{(\theta B)^{k-2}}{3^{k-2}} = \frac{\theta^2/2}{1 - \theta B/3}$$

$$\Longrightarrow \qquad e^{\theta \boldsymbol{X}} \preceq \boldsymbol{I} + \theta \boldsymbol{X} + \frac{\theta^2/2}{1 - \theta B/3} \cdot \boldsymbol{X}^2$$

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

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

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

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

Example: angular similarity

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

Kernel matrix

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

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

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

Challenge: kernel matrices are usually large

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

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

Random features

Introduce a random variable $\boldsymbol{w}$ and a feature map ψ such that

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

- **example (angular similarity)**

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

with $\boldsymbol{w}$ uniformly drawn from the unit sphere

Random features

Introduce a random variable $\boldsymbol{w}$ and a feature map ψ such that

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

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

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix} = \begin{bmatrix} \psi(\boldsymbol{x}_1; \boldsymbol{w}) \\ \vdots \\ \psi(\boldsymbol{x}_N; \boldsymbol{w}) \end{bmatrix}$$

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

Example

Angular similarity:

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

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

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

Random feature approximation

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

Estimator of the kernel matrix $\boldsymbol{G}$:

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

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,

$$\mathbb{E}[\boldsymbol{R}_l^2] = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{z}\boldsymbol{z}^\top] = N\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = N\boldsymbol{G}$$

$$\Longrightarrow \qquad v = \left\| \frac{1}{n^2} \sum\nolimits_{l=1}^{n} \mathbb{E}[\boldsymbol{R}_l^2] \right\| = \frac{N}{n} \|\boldsymbol{G}\|$$

- Next, $\frac{1}{n}\|\boldsymbol{R}\| = \frac{1}{n}\|\boldsymbol{z}\|_2^2 = \frac{N}{n} =: B$

- Applying the matrix Bernstein inequality yields: with high prob.

$$\|\hat{\boldsymbol{G}} - \boldsymbol{G}\| \lesssim \sqrt{v \log N} + B \log N \lesssim \sqrt{\frac{N}{n}\|\boldsymbol{G}\| \log N} + \frac{N}{n} \log N$$

$$\lesssim \sqrt{\frac{N}{n} \underbrace{\|\boldsymbol{G}\|}_{\geq 1} \log N} \qquad \text{(for sufficiently large } n\text{)}$$

Sample complexity

Define the intrinsic dimension of $\boldsymbol{G}$ as

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

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

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

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.