

## **Matrix concentration inequalities**



Cong Ma

University of Chicago, Autumn 2021

# Concentration inequalities

---

Let  $X_1, X_2, \dots, X_n$  be i.i.d. random variables, law of large numbers tells us that

$$\frac{1}{n} \sum_{l=1}^n X_l - \mathbb{E} \left[ \frac{1}{n} \sum_{l=1}^n X_l \right] \rightarrow 0, \quad \text{as } n \rightarrow \infty$$

## Key message:

sum of independent random variables *concentrate* around its mean

— *how fast does it concentrate?*

# Bernstein's inequality

---

Consider a sequence of independent random variables  $\{X_l\} \in \mathbb{R}$

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

$$v := \mathbb{E}[(\sum_l X_l)^2] = \sum_{l=1}^n \mathbb{E}[X_l^2]$$

## Theorem 4.1 (Bernstein's inequality)

For all  $\tau \geq 0$ ,

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

# Tail behavior

---

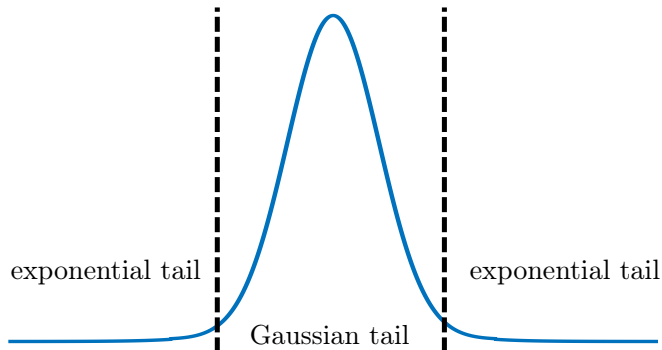
$$\mathbb{P}\left\{\left|\sum_l X_l\right| \geq \tau\right\} \leq 2 \exp\left(\frac{-\tau^2/2}{v + B\tau/3}\right)$$

- **moderate-deviation regime** ( $\tau$  is small):
  - sub-Gaussian tail behavior  $\exp(-\tau^2/2v)$
- **large-deviation regime** ( $\tau$  is large):
  - sub-exponential tail behavior  $\exp(-3\tau/2B)$  (slower decay)
- **user-friendly form** (exercise): with prob.  $1 - O(n^{-10})$

$$\left|\sum_l X_l\right| \lesssim \sqrt{v \log n} + B \log n$$

## Tail behavior (cont.)

---



**There are exponential concentration inequalities for spectral norm of sum of independent random matrices**

# 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 4.2 (Matrix Bernstein inequality)

For all  $\tau \geq 0$ ,

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

# 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 4.2 (Matrix Bernstein inequality)

For all  $\tau \geq 0$ ,

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

User-friendly form: with probability at least  $1 - O((d_1 + d_2)^{-10})$

$$\left\| \sum_l \mathbf{X}_l \right\| \lesssim \sqrt{v \log(d_1 + d_2)} + B \log(d_1 + d_2) \quad (4.1)$$



**This lecture: detailed introduction to matrix Bernstein**

*An introduction to matrix concentration inequalities*  
— Joel Tropp '15

# Outline

---

- Background on matrix functions
- Matrix Laplace transform method
- Matrix Bernstein inequality

## **Background on matrix functions**

# 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

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

— *align with our intuition about  $\mathbf{A}^k$*

# 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:**  $e^{\mathbf{A}} := \mathbf{I} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{A}^k$ 
  - 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  $\log \mathbf{A} \preceq \log(\mathbf{H})$  (does not hold for matrix exponential)

# 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):**

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

— *expectations may not exist for all  $\theta$*

## **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 4.3

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
- similar result holds for minimum eigenvalue

## Proof of Lemma 4.3

---

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:**

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

$$\text{tr } e^{\mathbf{X}_1 + \cdots + \mathbf{X}_n} \not\leq \text{tr } e^{\mathbf{X}_1} e^{\mathbf{X}_1} \cdots e^{\mathbf{X}_n} \quad \text{for } n \geq 3$$

## How about matrix CGF?

---

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

$$\Xi_X(\theta) = \log M_X(\theta) = \underbrace{\sum_{l=1}^n \log M_{X_l}(\theta)}_{\text{look at each } X_l \text{ separately}} = \sum_l \Xi_{X_l}(\theta)$$

In matrix case, can we hope for

$$\Xi_{\sum_l X_l}(\theta) = \sum_l \Xi_{X_l}(\theta) \quad ?$$

— *Nope; But...*

# Subadditivity of matrix CGF

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

## Lemma 4.4

*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 4.5 (Lieb '73)

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

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

*is concave on positive-definite cone*

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

$$\mathbb{E}[\operatorname{tr} \exp(\mathbf{H} + \mathbf{X})] \leq \operatorname{tr} \exp(\mathbf{H} + \log \mathbb{E}[e^{\mathbf{X}}]) \quad (4.2)$$

# Proof sketch of Lieb's Theorem

---

Main observation:  $\text{tr}(\cdot)$  admits a variational formula

## Lemma 4.6

For any  $M \succeq 0$ , one has

$$\text{tr} M = \sup_{T \succ 0} \text{tr} \left[ \underbrace{T \log M - T \log T + T}_{\text{relative entropy is } -T \log M + T \log T - T + M} \right]$$

## Proof of Lemma 4.4

---

$$\begin{aligned}\mathbb{E}[\mathrm{tr} e^{\theta \sum_l \mathbf{X}_l}] &= \mathbb{E}[\mathrm{tr} \exp(\theta \sum_{l=1}^{n-1} \mathbf{X}_l + \theta \mathbf{X}_n)] \\ &\leq \mathbb{E}\left[\mathrm{tr} \exp\left(\theta \sum_{l=1}^{n-1} \mathbf{X}_l + \log \mathbb{E}[e^{\theta \mathbf{X}_n}]\right)\right] \quad (\text{by (4.2)}) \\ &\leq \mathbb{E}\left[\mathrm{tr} \exp\left(\theta \sum_{l=1}^{n-2} \mathbf{X}_l + \log \mathbb{E}[e^{\theta \mathbf{X}_{n-1}}] + \log \mathbb{E}[e^{\theta \mathbf{X}_n}]\right)\right] \\ &\leq \dots \\ &\leq \mathrm{tr} \exp\left(\sum_{l=1}^n \log \mathbb{E}[e^{\theta \mathbf{X}_l}]\right)\end{aligned}$$



# Master bounds

---

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

## Theorem 4.7 (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 4.8 (Matrix Bernstein inequality: symmetric case)

For all  $\tau \geq 0$ ,

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

— left as exercise to prove extension to rectangular case

# Bounding matrix CGF

---

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

## Lemma 4.9

*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 4.8

---

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

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}\left(\sum_i \mathbf{X}_i\right) \geq t\right\} &\leq \inf_{\theta>0} \frac{\mathrm{tr} \exp\left(\sum_{i=1}^n \log \mathbb{E}[e^{\theta \mathbf{X}_i}]\right)}{e^{\theta t}} \\ &\stackrel{\text{Lemma 4.9}}{\leq} \inf_{0<\theta<3/B} \frac{\mathrm{tr} \exp\left(g(\theta) \sum_{i=1}^n \mathbb{E}[\mathbf{X}_i^2]\right)}{e^{\theta t}} \\ &\leq \inf_{0<\theta<3/B} \frac{d \exp(g(\theta)v)}{e^{\theta t}}\end{aligned}$$

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

## Proof of Lemma 4.9

---

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

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

Finish by observing  $\log$  is monotone

## Appendix: asymptotic notation

---

- $f(n) \lesssim g(n)$  or  $f(n) = O(g(n))$  means

$$\limsup_{n \rightarrow \infty} \frac{|f(n)|}{|g(n)|} \leq \text{const}$$

- $f(n) \gtrsim g(n)$  or  $f(n) = \Omega(g(n))$  means

$$\liminf_{n \rightarrow \infty} \frac{|f(n)|}{|g(n)|} \geq \text{const}$$

- $f(n) \asymp g(n)$  or  $f(n) = \Theta(g(n))$  means

$$f(n) \lesssim g(n) \quad \text{and} \quad f(n) \gtrsim g(n)$$

- $f(n) = o(g(n))$  means

$$\lim_{n \rightarrow \infty} \frac{|f(n)|}{|g(n)|} = 0$$