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

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

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:

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

Theorem 4.1 (Bernstein's inequality)

For all $\tau \geq 0$,

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

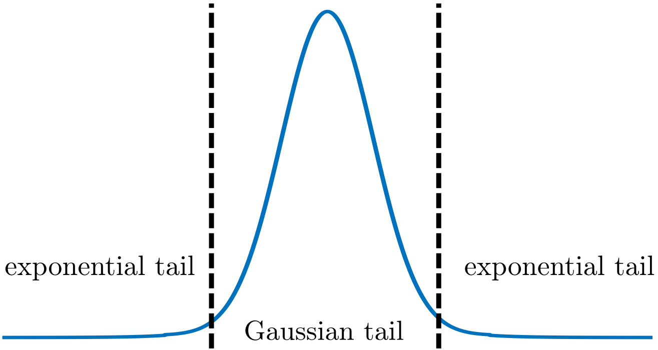
Tail behavior

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

- **moderate-deviation regime** (τ is small):
  — sub-Gaussian tail behavior $\exp(-\tau^2/2v)$
- **large-deviation regime** (τ 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 $\{\boldsymbol{X}_l \in \mathbb{R}^{d_1 \times d_2}\}$

- $\mathbb{E}[\boldsymbol{X}_l] = \boldsymbol{0}$
- $\|\boldsymbol{X}_l\| \le B$ 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 4.2 (Matrix Bernstein inequality)

For all $\tau \ge 0$,

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

Matrix Bernstein inequality

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

- $\mathbb{E}[\boldsymbol{X}_l] = \boldsymbol{0}$
- $\|\boldsymbol{X}_l\| \leq B$ 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 4.2 (Matrix Bernstein inequality)

For all $\tau \geq 0$,

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

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

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

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

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

Then we can define

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

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

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:** $\mathrm{e}^{\boldsymbol{A}} := \boldsymbol{I} + \sum_{k=1}^{\infty} \frac{1}{k!} \boldsymbol{A}^k$
  - 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 $\mathbf{0} \preceq \boldsymbol{A} \preceq \boldsymbol{H}$, then $\log \boldsymbol{A} \preceq \log(\boldsymbol{H})$ (does not hold for matrix exponential)

Matrix moments and cumulants

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

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

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

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

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

— *expectations may not exist for all* θ

Matrix Laplace transform method

Matrix Laplace transform

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

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

This can be generalized to the matrix case

Matrix Laplace transform

Lemma 4.3

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
- similar result holds for minimum eigenvalue

Proof of Lemma 4.3

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:

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

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

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

$$\boldsymbol{\Xi}_{\sum_l \boldsymbol{X}_l}(\theta) = \sum_l \boldsymbol{\Xi}_{\boldsymbol{X}_l}(\theta) \quad ?$$

— *Nope; But...*

Subadditivity of matrix CGF

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

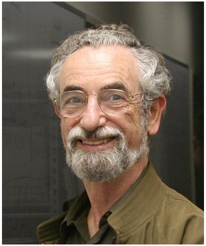
Lemma 4.4

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

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

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

Lieb's Theorem



Elliott Lieb

Theorem 4.5 (Lieb '73)

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

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

is concave on positive-definite cone

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

$$\mathbb{E}\big[\operatorname{tr}\exp(\boldsymbol{H} + \boldsymbol{X})\big] \leq \operatorname{tr}\exp\big(\boldsymbol{H} + \log \mathbb{E}\big[\mathrm{e}^{\boldsymbol{X}}\big]\big) \qquad (4.2)$$

Proof sketch of Lieb's Theorem

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

Lemma 4.6

For any $\boldsymbol{M} \succeq \boldsymbol{0}$, one has

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

Proof of Lemma 4.4

$$\begin{aligned}
\mathbb{E}\left[\operatorname{tr} \mathrm{e}^{\theta \sum_l \boldsymbol{X}_l}\right] &= \mathbb{E}\left[\operatorname{tr} \exp\left(\theta \sum\nolimits_{l=1}^{n-1} \boldsymbol{X}_l + \theta \boldsymbol{X}_n\right)\right] \\
&\leq \mathbb{E}\Big[\operatorname{tr} \exp\Big(\theta \sum\nolimits_{l=1}^{n-1} \boldsymbol{X}_l + \log \mathbb{E}\big[\mathrm{e}^{\theta \boldsymbol{X}_n}\big]\Big)\Big] \qquad \text{(by (4.2))} \\
&\leq \mathbb{E}\Big[\operatorname{tr} \exp\Big(\theta \sum\nolimits_{l=1}^{n-2} \boldsymbol{X}_l + \log \mathbb{E}\big[\mathrm{e}^{\theta \boldsymbol{X}_{n-1}}\big] + \log \mathbb{E}\big[\mathrm{e}^{\theta \boldsymbol{X}_n}\big]\Big)\Big] \\
&\leq \cdots \\
&\leq \operatorname{tr} \exp\Big(\sum\nolimits_{l=1}^{n} \log \mathbb{E}\big[\mathrm{e}^{\theta \boldsymbol{X}_l}\big]\Big)
\end{aligned}$$

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

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

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

Matrix Bernstein inequality

Matrix CGF

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

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 $\{\boldsymbol{X}_l \in \mathbb{R}^{d\times d}\}$

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

Theorem 4.8 (Matrix Bernstein inequality: symmetric case)

For all $\tau \geq 0$,

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

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

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

Proof of Theorem 4.8

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

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

$$\overset{\text{Lemma 4.9}}{\leq} \inf_{0<\theta<3/B} \frac{\operatorname{tr}\exp\big(g(\theta)\sum_{i=1}^n \mathbb{E}[\boldsymbol{X}_i^2]\big)}{e^{\theta t}}$$

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

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

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

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

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

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

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

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

Finish by observing $\log$ is monotone

Appendix: asymptotic notation

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

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

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

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

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

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

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

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