Matrix Chernoff Bound

Aarshvi Gajjar

NYU Tandon

Introduction

Random Matrix

Let $\mathbf{X}$ be a random matrix of size $d \times d$. There are two different ways to think of a random matrix:

1. A matrix sampled according to a distribution on matrices
2. An array of scalar random variables

Matrix Chernoff Bound

- In words, matrix Chernoff bound says that the eigenvalues of a sum of independent, random, positive-semidefinite matrices have a uniform upper bound.
- An ideal tool for studying random submatrices.

Example

Subspace Embedding based on Leverage Score sampling.

Difficulties in Generalising to Matrices

- How are functions like $\exp, \log$ extended to matrices?
- Multiplication is not commutative in matrices
- Proof of scalar Chernoff relies on the convexity of e^x. How is convexity defined in the matrix world?
- ...

The scalar and matrix versions

Theorem (Scalar Chernoff)

Let $X_1, \ldots, X_k$ *be independent real valued random variables with* $0 \le X_i \le R$. *Let* $\mu_{\min} \le \sum_{i=1}^{k} \mathbb{E}[X_i] \le \mu_{\max}$. *Then, for all* $\delta \ge 0$,

$$\Pr\left\{\sum_{i=1}^{k} X_i \ge (1+\delta)\mu_{\max}\right\} \le \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \le e^{-\frac{\delta^2 \mu_{\max}}{2+\delta}}$$

Theorem (Matrix Chernoff: Tropp, 2011)

Let $\mathbf{X}_1, \ldots, \mathbf{X}_k$ *be independent, random, symmetric, real matrices in* $\mathbb{R}^{d\times d}$ *with* $0 \preceq \mathbf{X}_i \preceq R \cdot I$ *and* $\mu_{\min} \cdot I \preceq \sum_{i=1}^{k} \mathbb{E}[\mathbf{X}_i] \preceq \mu_{\max} \cdot I$. *Then, for all* $\delta \in [0,1]$,

$$\Pr\left\{\lambda_{\max}\left(\sum_{i=1}^{k} \mathbf{X}_i\right) \ge (1+\delta)\mu_{\max}\right\} \le d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R}$$

Prerequisites from Matrix Analysis I

Definition (Spectral Mapping)

We extend $f : \mathbb{R} \to \mathbb{R}$ to a symmetric matrix, $\mathbf{X}$ by applying it to the eigenvalues of $\mathbf{X}$. i.e. if $\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^T$, then $f(\mathbf{X}) = \mathbf{U}f(\Sigma)\mathbf{U}^T$, where $f(\Sigma)_{ii} = f(\Sigma_{ii})$

- We also define the notion of monotonicity and concavity to spectral mapping.

Definition

$f : \mathbb{R} \to \mathbb{R}$ is:

1. **Operator Monotone** if $\mathbf{X} \preceq \mathbf{Y}$ implies that $f(\mathbf{X}) \preceq f(\mathbf{Y})$
2. **Operator Concave** if $f(\alpha\mathbf{X} + (1-\alpha)\mathbf{Y})) \succeq \alpha f(\mathbf{X}) + (1-\alpha)f(\mathbf{Y})$ for all $\alpha \in [0, 1]$. Example: $\log$, see Carlen, 2009.

Counter Example

f is monotone $\not\Rightarrow$ f is operator monotone! $f(x) = e^x$

Prerequisites from Matrix Analysis II

- We define an operator, $\odot$ which is a commutative version of matrix multiplication.

Definition

If $\mathbf{X}$ and $\mathbf{Y}$ are positive definite, then we define $\mathbf{X} \odot \mathbf{Y} = \exp(\log(\mathbf{X}) + \log(\mathbf{Y}))$

- Unlike matrix multiplication, this operation preserves positive definiteness.
- Note: If $\mathbf{X}$ and $\mathbf{Y}$ commute, then $\mathbf{X} \odot \mathbf{Y}$ is the usual multiplication $\mathbf{XY}$.
- To learn more, refer to Warmuth & Kuzmin, 2009.

We are ready to prove the theorem!

An intermediate result

Claim

$$\Pr\{\lambda_{\max}(\sum_{i=1}^{k} \mathbf{X}_i) \geq t\} \leq \inf_{\theta>0} e^{-\theta t} \cdot \textit{tr}\left(\mathbb{E}[e^{\theta \mathbf{X}_1}] \odot \mathbb{E}[e^{\theta \mathbf{X}_2}] \odot \ldots \odot \mathbb{E}[e^{\theta \mathbf{X}_k}]\right)$$

Proof.

- $\lambda_{\max}$ is a scalar and hence monotonicity and Markov's can be applied.

$$\Pr\{\lambda_{\max}(\sum_{i=1}^{k} \mathbf{X}_i) \geq t\} = \Pr\{e^{\lambda_{\max}(\sum_{i=1}^{k} \theta \mathbf{X}_i)} \geq e^{\theta t}\}, \quad \theta \geq 0$$

$$\leq e^{-\theta t} \cdot \mathbb{E}[e^{\lambda_{\max}(\sum_{i=1}^{k} \theta \mathbf{X}_i)}]$$

Proof continues..

Proof.

- Note that $\lambda_{\max}(e^{\mathbf{Y}}) = e^{\lambda_{\max}(\mathbf{Y})}$
- And $\lambda_{\max}(\mathbf{Y}) \leq \text{tr}(\mathbf{Y})$
- $\implies \exp(\lambda_{\max}(\sum_{i=1}^{k} \theta \mathbf{X}_i)) \leq \text{tr}(\exp(\sum_{i=1}^{k} \theta \mathbf{X}_i))$
- Taking Expectation and from the definition of $\odot$,

$$\begin{aligned} \mathbb{E}[\text{tr}(\exp(\sum_{i=1}^{k} \theta \mathbf{X}_i))] &= \mathbb{E}[\text{tr}(\exp(\sum_{i=1}^{k} \log(A_i)))], \quad \mathbf{A}_i = \exp(\theta \mathbf{X}_i) \\ &= \mathbb{E}[\text{tr}(\mathbf{A}_1) \odot \text{tr}(\mathbf{A}_2) \ldots \odot \text{tr}(\mathbf{A}_k)] \end{aligned}$$

- To take the expectation inside, we first use the result from Lieb which states that the map $\mathbf{X} \to \text{tr}(\mathbf{X} \odot \mathbf{Y})$ is concave followed by Jensen's Inequality and induction.

□

- We have currently shown:

$$\Pr\{\lambda_{\max}(\sum_{i=1}^{k} \mathbf{X}_i) \geq t\} \leq \inf_{\theta>0} e^{-\theta t} \cdot \mathsf{tr}\left(\mathbb{E}[e^{\theta \mathbf{X}_1}] \odot \mathbb{E}[e^{\theta \mathbf{X}_2}] \odot \ldots \odot \mathbb{E}[e^{\theta \mathbf{X}_k}]\right)$$

- We want to show:

$$\Pr\left\{\lambda_{\max}(\sum_{i=1}^{k} \mathbf{X}_i) \geq (1+\delta)\mu_{\max}\right\} \leq d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R}$$

The main proof

Proof.

- Continuing from the above claim, we have the following term in R.H.S.

$$\text{tr}(\mathbb{E}[e^{\theta \mathbf{X}_1} \odot \mathbb{E}[e^{\theta \mathbf{X}_2}] \ldots \odot \mathbb{E}[e^{\theta \mathbf{X}_k}])$$

- Next, we expand on the definition of $\odot$ and multiply and divide by k.

$$\text{tr}(\mathbb{E}[e^{\theta \mathbf{X}_1} \odot \mathbb{E}[e^{\theta \mathbf{X}_2}] \ldots \odot \mathbb{E}[e^{\theta \mathbf{X}_k}]) = \text{tr}(\exp(k \sum_{i=1}^{k} \frac{1}{k} \cdot \log(\mathbb{E}[e^{\theta \mathbf{X}_i}])))$$

- It has been shown before that the $\log$ is operator concave. Therefore,

$$\sum_{i=1}^{k} \frac{1}{k} \log(\mathbb{E}[e^{\theta \mathbf{X}_i}]) \preceq \log(\sum_{i=1}^{k} \frac{1}{k} \mathbb{E}[e^{\theta \mathbf{X}_i}]) \tag{1}$$

Proof Continues..

Proof.

- **Fact**: $\exp$ is not operator monotone but the composition $\text{tr}\exp$ is operator monotone. See Bhatia, 1997.
- From (1), $\uparrow$ and $\text{tr}(\mathbf{Y}) \leq d \cdot \lambda_{\max}(\mathbf{Y})$,

$$\text{tr}(\exp(k\sum_{i=1}^{k}\frac{1}{k}\cdot\log(\mathbb{E}[e^{\theta\mathbf{X}_i}]))) \leq d\cdot\lambda_{\max}(\exp(k\log(\sum_{i=1}^{k}\frac{1}{k}\mathbb{E}[e^{\theta\mathbf{X}_i}])))$$
$$= d\cdot\exp(k\log(\lambda_{\max}(\sum_{i=1}^{k}\frac{1}{k}\mathbb{E}[e^{\theta\mathbf{X}_i}])))$$

- The last equality holds because spectral mapping is only applied to eigenvalues.

Proof continues on..

Proof.

- **Fact**: It can be shown that $e^{\theta \mathbf{X}}$ is operator convex.
- $\implies$ if $0 \preceq \mathbf{X} \preceq 1$ then $\mathbb{E}[e^{\theta\cdot((1-\mathbf{X})\cdot 0+\mathbf{X}\cdot 1)}] \preceq I + (e^{\theta} - 1)\cdot \mathbb{E}[\mathbf{X}]$.
- Thus, the chain of inequalities follow:

$$
\begin{aligned}
d \cdot \exp(k \log(\lambda_{\max}(\sum_{i=1}^{k} \frac{1}{k}\mathbb{E}[e^{\theta \mathbf{X}_i}]))) &\leq d \cdot \exp(k \log \lambda_{\max}(I + (\exp(\theta) - 1) \sum_{i=1}^{k} \frac{1}{k}\mathbb{E}[\mathbf{X}_i]) \\
&= d \cdot \exp(k \log(1 + \frac{e^{\theta} - 1}{k}\lambda_{\max}(\sum_{i=1}^{k}\mathbb{E}[\mathbf{X}_i])) \\
&\leq d \cdot \exp(k \log(e^{\theta} - 1 \cdot \mu_{\max}))
\end{aligned}
$$

- Placing $t = (1+\delta)\mu_{\max}$ and $\theta = \ln(1+\delta)$ completes the proof.
- Q.E.D.

References

- Lecture notes by Nick Harvey
- An Introduction to Matrix Concentration Inequalities, Joel Tropp
- Sketching as a Tool for NLA, David Woodruff
- Matrix Analysis, R. Bhatia, Springer 1997
- Bayesian generalized probability calculus for density, M. K. Warmuth and D. Kuzmin, 2010
- Trace inequalities and quantum entropy: An introductor, E. Carlen, 2009