Randomized Matrix Computations Lecture 4

Daniel Kressner
Chair for Numerical Algorithms and HPC
Institute of Mathematics, EPFL
daniel.kressner@epfl.ch

EPFL

Matrix Concentration

- Eigenvalues of sums of independent matrices
- Matrix Bernstein
- Matrix Monte Carlo
- Matrix Chernoff (next week)

Literature:

Tropp'2020 Joel A. Tropp. *Randomized Algorithms for Matrix Computations*, Lecture notes, Caltech, 2020.

Tropp'2015 Joel A. Tropp. *An Introduction to Matrix Concentration Inequalities*, Foundations and Trends in Machine Learning, Vol. 8, No. 1–2 (2015).

pdf available on Moodle

Sums of independent matrices

For independent random matrices $X_1, \ldots, X_s \in \mathbb{R}^{n \times n}$, consider the sum

$$Y = X_1 + \cdots + X_s.$$

Goal: Control norm $\|Y\|_2$ of Y.

Often X_j are very simple random matrices. Two examples:

- Given a fixed $m \times n$ matrix A, one could choose the discrete distribution
$$\mathbb{P}\{X = mn \cdot a_{ij} e_i e_j^\top\} = \frac{1}{mn}.$$
Then $\mathbb{E}[X] = A$.
- Set $X_{ij} = x_{ij} e_i e_j^T$ where $x_{ij} \sim N(0,1)$ are independent for $i = 1, \ldots, m$, $j = 1, \ldots, n$. Then $Y = \sum_{ij} X_{ij}$ is a Gaussian random matrix.
- Set $X_{ij} = x_{ij} e_i e_j^T$ where x_{ij} are independent Rademacher rv. Then $Y = \sum_{ij} X_{ij}$ is a Rademacher random matrix.

Sums of symmetric independent matrices: MGF

We now consider *symmetric* independent random matrices $X_1, \dots, X_s \in \mathbb{R}^{n\times n}$:

$$Y = X_1 + \cdots + X_s.$$

The matrix mgf of Y is defined by taking the trace of the matrix exponential:

$$m_Y(\theta) = \mathbb{E}\Big[\operatorname{trace}\Big(\exp(\theta Y)\Big)\Big] = \mathbb{E}\Big[\operatorname{trace}\Big(\exp(\theta(X_1 + \cdots + X_s))\Big)\Big].$$

Big problem: The inequality

$$\begin{aligned} & \operatorname{trace}\Big(\exp(\theta(X_1 + \cdots + X_s))\Big) \\ \leq\ & \operatorname{trace}\Big(\exp(\theta X_1)\Big) \cdot \operatorname{trace}\Big(\exp(\theta X_2)\Big) \cdot \cdots \cdot \operatorname{trace}\Big(\exp(\theta X_s)\Big). \end{aligned}$$

only holds for $s = 2$ (Golden-Thompson inequality) but not for $s \geq 3$.

EFY: Find three matrices such that this inequality is violated (taking $n = 2$ suffices).

Sums of symmetric independent matrices: MGF

Way out: Lieb's concavity theorem. The set of symmetric positive semi-definite (spsd) matrices is convex :

$$A \geq 0, B \geq 0 \;\Rightarrow\; tA + (1-t)B \geq 0, \quad 0 \leq t \leq 1.$$

For fixed symmetric A, Elliott Lieb showed that the function

$$B \mapsto \mathrm{trace}\left(\exp(A + \log B)\right)$$

is convex on spsd matrices ($\log$ denotes the matrix logarithm).

This allows us to use Jensen's inequality!

$s = 2$:

$$\begin{aligned}
& \mathbb{E}_{(X_1,X_2)}\left[\mathrm{trace}\left(\exp(X_1 + X_2)\right)\right] \\
= \quad & \mathbb{E}_{X_1}\left[\mathbb{E}_{X_2}\left[\mathrm{trace}\left(\exp(X_1 + \log e^{X_2})\right) \mid X_1\right]\right] \\
\leq \quad & \mathbb{E}_{X_1}\left[\mathrm{trace}\left(\exp(X_1 + \log \mathbb{E}_{X_2} e^{X_2})\right)\right] \\
\leq \quad & \mathrm{trace}\left(\exp(\log \mathbb{E}_{X_1} e^{X_1} + \log \mathbb{E}_{X_2} e^{X_2})\right).
\end{aligned}$$

The argument extends to general s.

Sums of symmetric independent matrices: MGF

Lemma. Given symmetric independent random matrices $X_1, \ldots, X_s$, we have that

$$\mathbb{E}\Big[\operatorname{trace}\Big(\exp\Big(\sum_j X_j\Big)\Big)\Big] \le \operatorname{trace}\Big(\exp\Big(\sum_j \log \mathbb{E} e^{X_j}\Big)\Big).$$

Theorem. Let $Y = \sum_j X_j$ for symmetric independent random matrices $X_1, \ldots, X_s$. Then, for all $t \ge 0$,

$$\mathbb{P}\{\lambda_{\max}(Y) \ge t\} \le \inf_{\theta > 0} e^{-\theta t} \operatorname{trace}\Big(\exp\Big(\sum_j \log \mathbb{E} e^{\theta X_j}\Big)\Big).$$

Proof. By Markov's inequality, it holds for $\theta > 0$ that

$$\mathbb{P}\{\lambda_{\max}(Y) \ge t\} = \mathbb{P}\{e^{\theta \lambda_{\max}(Y)} \ge e^{\theta t}\} \le e^{-\theta t} \mathbb{E} e^{\theta \lambda_{\max}(Y)}.$$

Using that the largest eigenvalue of the exponential is the exponential of the largest eigenvalue, we get

$$\mathbb{E} e^{\theta \lambda_{\max}(Y)} = \mathbb{E} \lambda_{\max}\big(e^{\theta Y}\big) \le \mathbb{E} \operatorname{trace}\big(e^{\theta Y}\big).$$

The theorem now follows from applying the lemma. $\diamond$

Sums of symmetric independent matrices: MGF

EFY: Show that the theorem implies the bound

$$\mathbb{E}[\lambda_{\max}(Y)] \leq \inf_{\theta>0} \frac{1}{\theta} \log \operatorname{trace}\Big(\exp\Big(\sum_j \log \mathbb{E} e^{\theta X_j} \Big)\Big).$$

Matrix Bernstein

To prepare for matrix Bernstein, need to discuss operator monotonicity:

- For symmetric matrices A, B we write $A \le B$ iff $B - A$ is spsd.
- $f(\lambda) \le g(\lambda)$ for every eigenvalue of A implies $f(A) \le g(A)$.
- If $X \le Y$ holds (almost surely) for random matrices X, Y over the same probability space then $\mathbb{E}X \le \mathbb{E}Y$.
- A function $f : (0, \infty) \to \mathbb{R}$ is operator monotone if $A \le B$ implies $f(A) \le f(B)$ for all spd A, B.
- $f(x) = x^2$ is monotone but *not* operator monotone.
- EFY: Prove that $f(x) = -1/x$ is operator monotone.
- EFY: Using the integral representation

  $$\log a = \int_0^\infty \left[(1+x)^{-1} - (a+x)^{-1}\right] \mathrm{d}x, \quad a > 0,$$

  show that the logarithm is operator monotone.

Matrix Bernstein: MGF for bounded matrices

Lemma. Let X be a zero-mean random matrix such that $\|X\|_2 \le 1$ almost surely. Then

$$\log \mathbb{E} e^{\theta X} \le \frac{\theta^2/2}{1 - |\theta|/3} \cdot \mathbb{E} X^2 \quad \forall |\theta| < 3.$$

Proof. From Lecture 2, we already know that

$$e^{\theta x} \le 1 + \theta x + \frac{\theta^2 x^2/2}{1 - |\theta|/3}, \quad |x| \le 1, |\theta| < 3.$$

Hence,

$$e^{\theta X} \le I + \theta X + \frac{\theta^2/2}{1 - |\theta|/3} X^2.$$

Taking expected values on both sides gives

$$\mathbb{E} e^{\theta X} \le I + \frac{\theta^2/2}{1 - |\theta|/3} \mathbb{E} X^2 \le \exp\left(\frac{\theta^2/2}{1 - |\theta|/3} \mathbb{E} X^2 \right).$$

Using operator monotonicity, the result is obtained by taking the logarithm on both sides. ◇

Matrix Bernstein: Symmetric case

Theorem. Let $Y = \sum_j X_j$ for symmetric independent random matrices $X_1, \dots, X_s \in \mathbb{R}^{n\times n}$. Assume that

$$\mathbb{E}X_j = 0, \quad \lambda_{\max}(X_j) \le L, \quad j = 1, \dots, s,$$

and set $v(Y) = \|\mathbb{E}Y^2\|_2 = \left\| \sum_j \mathbb{E}X_j^2 \right\|_2$. Then

$$\mathbb{E}\lambda_{\max}(Y) \le \sqrt{2v(Y)\log n} + \frac{1}{3}L\log n$$

and

$$\mathbb{P}\{\lambda_{\max}(Y) \ge t\} \le n \cdot \exp\left(\frac{-t^2/2}{v(Y) + Lt/3}\right), \quad t \ge 0.$$

Example. For iid Rademacher ϵ_{ij}, $1 \le i \le j \le n$, consider matrices

$$X_{ij} = \epsilon_{ij}(e_i e_j^\top + e_j e_i^\top), \; i > j, \quad X_{ii} = \epsilon_{ij} e_i e_i^\top.$$

Then $v(Y) = n$ and, hence, $\mathbb{E}\lambda_{\max}(Y) \le \sqrt{2n\log n} + \frac{1}{3}\log n$.
Minor improvement in [Tropp'2015]: Term $\frac{1}{3}\log n$ can be dropped.

Matrix Bernstein: Symmetric case

Proof. Known bound (from Slide 6) $\rightsquigarrow$

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(Y) \geq t\} &\leq \inf_{\theta>0} e^{-\theta t} \operatorname{trace}\Big(\exp\Big(\sum_j \log \mathbb{E} e^{\theta X_j}\Big)\Big) \\
&\leq \inf_{0<\theta<3/L} e^{-\theta t} \operatorname{trace}\left(\exp\left(\frac{\theta^2/2}{1 - L\theta/3} \sum_j \mathbb{E} X_j^2\right)\right) \\
&= \inf_{0<\theta<3/L} e^{-\theta t} \operatorname{trace}\left(\exp\left(\frac{\theta^2/2}{1 - L\theta/3} \mathbb{E} Y^2\right)\right)
\end{aligned}$$

In the second inequality we used our mgf bound and: (1) Incorporate L by scaling. (2) EFY: The function $B \mapsto \operatorname{trace} \exp(B)$ is operator monotone.

Matrix Bernstein: Symmetric case

Proof ctd. Bounding the trace by $n \times \lambda_{\max} \rightsquigarrow$

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(Y) \geq t\} &\leq \inf_{0<\theta<3/L} n e^{-\theta t} \lambda_{\max}\Big(\exp\Big(\frac{\theta^2/2}{1 - L\theta/3} \mathbb{E} Y^2 \Big)\Big) \\
&\leq \inf_{0<\theta<3/L} n e^{-\theta t} \exp\Big(\frac{\theta^2/2}{1 - L\theta/3} \lambda_{\max}(\mathbb{E} Y^2) \Big) \\
&\leq \inf_{0<\theta<3/L} n e^{-\theta t} \exp\Big(\frac{\theta^2/2}{1 - L\theta/3} v(Y) \Big).
\end{aligned}$$

Choosing the (nearly optimal) value $\theta^* = \frac{t}{v(Y)+Lt/3}$ concludes the proof.

Bound on expected value proven in [Tropp'2015]. $\diamond$

Note: Factor n can be reduced to $4 \cdot \mathrm{intdim}(\mathbb{E} Y^2)$, where $\mathrm{indim}(B) = \mathrm{trace}(B)/\|B\|_2$.

Matrix Bernstein: General case

We now consider $Y = \sum_j X_j$ for general independent random matrices $X_1, \dots, X_s \in \mathbb{R}^{m \times n}$. This can be reduced to the symmetric case using the dilation

$$\mathcal{S}(B) = \begin{pmatrix} 0 & B \\ B^\top & 0 \end{pmatrix}.$$

We need the following simple facts:

- $S(B)$ is a symmetric $(m+n) \times (m+n)$ matrix.
- Dilation is linear $\rightsquigarrow \mathcal{S}(Y) = \sum_j \mathcal{S}(X_j)$.
- $\lambda_{\max}(\mathcal{S}(B)) = \|B\|_2$.
- $\mathcal{S}(B)^2 = \begin{pmatrix} BB^\top & 0 \\ 0 & B^\top B \end{pmatrix}$ and, hence, $\|\mathcal{S}(B)^2\|_2 = \max\{\|BB^\top\|_2, \|B^\top B\|_2\}$.

We can now apply symmetric Bernstein to $\mathcal{S}(Y)$ in order to obtain general Bernstein.

Matrix Bernstein: General case

Theorem (Matrix Bernstein inequality). Let $Y = \sum_j X_j$ for independent random matrices $X_1, \ldots, X_s \in \mathbb{R}^{m \times n}$. Assume that

$$\mathbb{E} X_j = 0, \quad \|X_j\|_2 \leq L, \quad j = 1, \ldots, s,$$

and set $\sigma^2 = \max\{\|\mathbb{E} Y^\top Y\|_2, \|\mathbb{E} Y Y^\top\|_2\}$. Then

$$\mathbb{E}\|Y\|_2 \leq \sqrt{2\sigma^2 \log(m+n)} + \frac{1}{3} L \log(m+n)$$

and

$$\mathbb{P}\{\|Y\|_2 \geq t\} \leq (m+n) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Lt/3}\right), \quad t \geq 0.$$

EFY. Derive bounds on the expected value and tail for the norm of a Rademacher matrix Y. Compare the obtained bounds with numerical experiments.

Matrix Monte Carlo

Abstract setting. Consider a fixed (unknown) matrix $B \in \mathbb{R}^{m \times n}$ we wish to approximate. Assume we have a decomposition

$$B = B_1 + \cdots B_d,$$

where each summand B_j is simple to compute.

Idea: Approximate B from sampling summands B_j.

Sampling is driven by (discrete) random matrix X with prob distribution

$$\mathbb{P}\{X = B_k/p_k\} = p_k, \quad k = 1, \ldots, d.$$

Scaling ensures that $\mathbb{E}[X] = B$.

Matrix Monte Carlo estimator:

$$\bar{X}_s = \frac{1}{s}\left(X_1 + \cdots + X_s\right), \quad X_j \sim X \text{ iid}.$$

Note that $\mathbb{E}[\bar{X}_s] = B$ and we expect/hope that $\bar{X}_s$ will concentrate around B as s increases.

Matrix Monte Carlo: Main result

Matrix Bernstein immediately gives:

Theorem. Let $B \in \mathbb{R}^{m\times n}$ be fixed. Consider random matrix X such that

$$\mathbb{E}X = B, \quad \|X\|_2 \le L, \quad \text{almost surely.}$$

Set $\sigma^2 = \max\{\|\mathbb{E}X^\top X\|_2, \|\mathbb{E}XX^\top\|_2\}$. Then $\bar{X}_s = (X_1 + \cdots + X_s)/s$ with $X_j \sim X$ iid satisfies the error bounds

$$\mathbb{E}\|\bar{X}_s - B\|_2 \le s^{-1/2}\sqrt{2\sigma^2 \log(m+n)} + (3s)^{-1} L \log(m+n)$$

and

$$\mathbb{P}\{\|\bar{X}_s - B\|_2 \ge t\} \le (m+n)\cdot \exp\left(-\frac{st^2}{\sigma^2 + 2Lt/3}\right), \quad t \ge 0.$$

Application: Approximate matrix multiplication

Idea: Approximate a matrix product of the form

$$B = CR^{\top}, \quad C \text{ is } m \times d,\ R \text{ is } n \times d \text{ with } d \gg \max\{m, n\},$$

by matrix Monte Carlo. Assume $\|C\|_2 = \|R\|_2 = 1$ in the following.

Partitioning $C = (c_1, \ldots, c_d)$, $R = (r_1, \ldots, r_d)$, allows us to rewrite

$$B = c_1 r_1^{\top} + c_2 r_2^{\top} + \cdots + c_d r_d^{\top} =: B_1 + B_2 + \cdots + B_d.$$

Define random matrix X by $\mathbb{P}\{X = c_k r_k^{\top} / p_k\} = p_k$ for sampling probabilities $p_1, \ldots, p_d$.

Matrix Monte Carlo estimator

$$\bar{X}_s = \frac{1}{s}(X_1 + \cdots + X_s), \quad X_j \sim X \text{ iid},$$

requires $\mathcal{O}(smn)$ operations, which is interesting when $s \ll d$.

Application: Approximate matrix multiplication

Uniform sampling: $p_k = 1/d$.

Obviously, the estimator will perform poorly if there are a few terms B_j that dominate the rest. Measured by coherence statistics:

$$\mu(C) := d \cdot \max_k \|c_k\|_2^2, \quad m \le \mu(C) \le d.$$

Recall that $\|C\|_2 = 1$. Incoherence (small coherence) is good!

Computing the ingredients of the theorem:

$$\begin{aligned}
\|X\|_2 &\le \max_k \|c_k r_k^\top / p_k\|_2 \le d \cdot \max_k \|c_k\|_2 \cdot \max_k \|r_k\|_2 \le \max\{\mu(C), \mu(R)\}, \\
\mathbb{E}[XX^\top] &= \sum_{k=1}^d p_k (c_k r_k^\top / p_k)(c_k r_k^\top / p_k)^\top = \sum_{k=1}^d p_k^{-1} \|r_k\|_2^2 c_k c_k^\top \\
&\le d \cdot \max \|r_k\|_2^2 \sum_{k=1}^d p_k^{-1} c_k c_k^\top = \mu(R) \cdot CC^\top.
\end{aligned}$$

This implies $\|\mathbb{E}[XX^\top]\|_2 \le \mu(R) \|CC^\top\|_2 = \mu(R)$. Applying the same arguments to $X^\top$ instead of X gives $\|\mathbb{E}[X^\top X]\|_2 \le \mu(C)$. In summary

$$\sigma^2 = \max\{\|\mathbb{E}X^\top X\|_2, \|\mathbb{E}XX^\top\|_2\} \le \max\{\mu(C), \mu(R)\}.$$

Application: Approximate matrix multiplication

The theorem now tells us that

$$\frac{\mathbb{E}\|\bar{X}_s - CR^\top\|_2}{\|C\|_2\|R\|_2} \leq 2\varepsilon$$

holds if

$$s \geq \max\left\{\varepsilon^{-2}, (3\varepsilon)^{-1}\right\} \cdot \max\{\mu(C), \mu(R)\} \cdot \log(m+n).$$

In favorable situations (incoherent factors), need $s \approx (m+n)\log(m+n)$ samples to attain fixed (but somewhat rough) accuracy.

Application: Approximate matrix multiplication

Importance sampling: Let us reconsider the second moment

$$\mathbb{E}[XX^\top] = \sum_{k=1}^{d} p_k^{-1}\|r_k\|_2^2 c_k c_k^\top$$

$p_k \sim \|r_k\|_2^2$ makes the unfavorable dependence of $\mathbb{E}[XX^\top]$ on "spiky" row norms disappear. Similarly, $p_k \sim \|c_k\|_2^2$ makes the unfavorable dependence of $\mathbb{E}[X^\top X]$ on "spiky" column norms disappear. Therefore, choose:

$$p_k = \frac{\|c_k\|_2^2 + \|r_k\|_2^2}{\|C\|_F^2 + \|R\|_F^2}.$$

With this choice, we have :

$$\|X\|_2 \leq \max_k \frac{1}{p_k}\|c_k r_k^\top\|_2 = (\|C\|_F^2 + \|R\|_F^2)\max_k \frac{\|c_k\|_2\|r_k\|_2}{\|c_k\|_2^2 + \|r_k\|_2^2} \leq \frac{\|C\|_F^2 + \|R\|_F^2}{2}$$

$$\|\mathbb{E}[XX^\top]\|_2 = \left\|\sum_{k=1}^{d} p_k^{-1}\|r_k\|_2^2 c_k c_k^\top\right\|_2 \leq \|C\|_F^2 + \|R\|_F^2.$$

Recall that $\|R\|_2 = \|C\|_2 = 1$.

Application: Approximate matrix multiplication

Defining the stable rank $\operatorname{srank}(C) := \|C\|_F^2 / \|C\|_2^2 \in [1, \operatorname{rank}(C)]$, we have

$$
\begin{aligned}
\|X\|_2 &\leq \frac{1}{2}(\operatorname{srank}(B) + \operatorname{srank}(C)), \\
\max\{\|\mathbb{E}[XX^\top]\|_2, \|\mathbb{E}[X^\top X]\|_2\} &\leq \operatorname{srank}(B) + \operatorname{srank}(C)
\end{aligned}
$$

The theorem now tells us that

$$
\frac{\mathbb{E}\|\bar{X}_s - CR^\top\|_2}{\|C\|_2 \|R\|_2} \leq 2\varepsilon
$$

holds if

$$
s \geq \max\left\{2\varepsilon^{-2}, (6\varepsilon)^{-1}\right\} \cdot (\operatorname{srank}(B) + \operatorname{srank}(C)) \cdot \log(m+n).
$$

This is never worse than (the bound for) uniform sampling. Becomes very effective for small stable ranks.

Matrix Chernoff: Setting and MGF

Now consider $Y = X_1 + \cdots + X_s$ for independent random symmetric positive semi-definite (SPSD) matrices $X_1, \ldots, X_s \in \mathbb{R}^{n\times n}$. Assume that

$$0 \le \lambda_{\min}(X_j), \quad \lambda_{\max}(X_j) \le L, \quad j = 1, \ldots, s.$$

Goals:

- Control norm/maximum eigenvalue of Y.
- Control invertibility/minimum eigenvalue of Y.

Lemma. Let X be a random symmetrix matrix s.t. $0 \preceq X \preceq I$. Then

$$\log \mathbb{E} e^{\theta X} \preceq (e^{\theta} - 1)\mathbb{E}X, \quad \forall \theta \in \mathbb{R}.$$

Proof. The convexity of $t \mapsto e^{\theta t}$ implies $e^{\theta t} \le 1 + (e^{\theta} - 1)t$ for $t \in [0, 1]$. Because the eigenvalues of X are in $[0, 1]$, this implies

$$e^{\theta X} \preceq I + (e^{\theta} - 1)X \quad \Rightarrow \quad \mathbb{E}e^{\theta X} \preceq I + (e^{\theta} - 1)\mathbb{E}X.$$

Taking the matrix logarithm (and using operator monotonicity) and using $\log(1 + t) \le t$ completes the proof. $\diamond$

Matrix Chernoff: Tail bounds

Theorem. Consider setting from previous slide and define

$$\mu_{\min} = \lambda_{\min}(\mathbb{E}Y), \quad \mu_{\max} = \lambda_{\max}(\mathbb{E}Y).$$

Then

$$\mathbb{P}\{\lambda_{\min}(Y) \leq (1-\epsilon)\mu_{\min}\} \leq n \cdot \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{\mu_{\min}/L}, \quad 0 < \epsilon \leq 1,$$

$$\mathbb{P}\{\lambda_{\max}(Y) \geq (1+\epsilon)\mu_{\min}\} \leq n \cdot \left(\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right)^{\mu_{\max}/L}, \quad 0 < \epsilon.$$

Corollary: If $X_1, \ldots, X_s$ are iid $\sim X$, $Y = (X_1 + \cdots + X_s)/s$, then

$$\mathbb{P}\{\lambda_{\min}(Y) \leq \mu_{\min}/2\} \leq n \cdot 0.86^{s \cdot \mu_{\min}/L}$$

with $\mu_{\min} = \lambda_{\min}(\mathbb{E}X)$, $\lambda_{\max}(X) \leq L$.

Conclusion: Smallest eigenvalue of Y stays nicely away from zero for $s \sim \log n$, assuming that $\mu_{\min}$, L do not depend on n. X only needs to be PD in expectation!

Matrix Chernoff: Tail bounds

Proof for $\lambda_{\max}$. Assume w.l.o.g. $L = 1$. Recall MGF bound $\log \mathbb{E} e^{\theta X} \leq g(\theta) \mathbb{E} X$ with $g(\theta) = e^{\theta} - 1$. Combined with Lemma from Slide 6, this gives

$$
\begin{aligned}
\mathbb{P}\{\lambda_{\max}(Y) \geq t\} &\leq \inf_{\theta > 0} e^{-\theta t} \operatorname{trace}\Big(\exp\Big(\sum_j \log \mathbb{E} e^{\theta X_j} \Big)\Big) \\
&\leq \inf_{\theta > 0} e^{-\theta t} \operatorname{trace}\Big(\exp\big(g(\theta) \mathbb{E} Y \big)\Big) \\
&\leq n \cdot \inf_{\theta > 0} e^{-\theta t} \lambda_{\max}\Big(\exp\big(g(\theta) \mathbb{E} Y \big)\Big) \\
&= n \cdot \inf_{\theta > 0} e^{-\theta t} \exp\big(g(\theta) \lambda_{\max}(\mathbb{E} Y) \big).
\end{aligned}
$$

Change of variables $t \mapsto (1 + \epsilon)\mu_{\max}$ and setting $\theta = \log(1 + \epsilon)$ complete proof of tail bound for $\lambda_{\max}$.

Matrix Chernoff: Tail bounds

Proof for $\lambda_{\min}$. Again combining Lemma from Slide 6 (applied to $-Y$) with mgf bound gives

$$\begin{aligned}
\mathbb{P}\{\lambda_{\min}(Y) \le t\} &\le \inf_{\theta<0} e^{-\theta t} \operatorname{trace}\Big(\exp\Big(\sum_j \log \mathbb{E} e^{\theta X_j}\Big)\Big) \\
&\le \inf_{\theta<0} e^{-\theta t} \operatorname{trace}\big(\exp\big(g(\theta)\mathbb{E}Y\big)\big) \\
&\le n \cdot \inf_{\theta>0} e^{-\theta t} \lambda_{\max}\big(\exp\big(g(\theta)\mathbb{E}Y\big)\big) \\
&= n \cdot \inf_{\theta>0} e^{-\theta t} \exp\big(g(\theta)\lambda_{\min}(\mathbb{E}Y)\big).
\end{aligned}$$

Here, we used that $g(\theta) = e^{\theta} - 1 < 0$ for $\theta < 0$. Hence, the largest eigenvalue of $g(\theta)\mathbb{E}Y$ corresponds to the smallest eigenvalue of $\mathbb{E}Y$. The proof is completed by change of variables $t \mapsto (1-\epsilon)\mu_{\min}$ and setting $\theta = \log(1-\epsilon)$. $\diamond$

Matrix Chernoff: Expectation bounds

EFY: Derive, with similar arguments, the following result by combining the bound of Slide 7 with the mgf bound.

Theorem. Consider setting from previous theorem. Then

$$\begin{aligned} \mathbb{E}\lambda_{\min}(Y) &\geq \frac{1-e^{-\theta}}{\theta}\mu_{\min} - \frac{1}{\theta} L \log n, \\ \mathbb{E}\lambda_{\max}(Y) &\leq \frac{e^{\theta}-1}{\theta}\mu_{\max} + \frac{1}{\theta} L \log n, \end{aligned}$$

hold for every $\theta > 0$.

Setting $\theta = 1$ in the bounds gives the simpified version

$$\begin{aligned} \mathbb{E}\lambda_{\min}(Y) &\geq 0.63\mu_{\min} - L \log n, \\ \mathbb{E}\lambda_{\max}(Y) &\leq 1.72\mu_{\max} + L \log n. \end{aligned}$$

Matrix Chernoff: Application to Erdős–Rényi graphs

Setting:

- Consider undirected graphs (V, E) with vertices $V = \{1, \dots, n\}$.
- *Adjacency matrix* A associated with (V, E) is a symmetric 0-1 matrix with $a_{ij} = 1$ if $\{i, j\} \in E$ and $a_{ij} = 0$. We use the convention that the diagonal entries are zero: $a_{ii} = 0$ for $i = 1, \dots, n$.

Define Graph Laplacian:

$$\Delta = D - A,$$

where D is diagonal matrix with diagonal entries $d_1 = \deg(1)$, ..., $d_n = \deg(n)$. Equivalently, $d = Ae$, where $e = [1, \dots, 1]^\top$.

Properties of Graph Laplacian:

- Δ is symmetric
- Because of $\Delta e = 0$, Δ has at least one zero eigenvalue.
- $\|\Delta\|_2 \le \max \deg(i)$ (EFY: Why?)
- Δ is SPSD (EFY: Why?)
- The number of zero eigenvalues of Δ equals the number of disconnected components of (V, E).
- (V, E) is connected iff second smallest eigenvalue of Δ is pos.

Matrix Chernoff: Application to Erdős–Rényi graphs

Erdős–Rényi graph $G(n,p)$: Between each pair of distinct vertices there is an edge with probability $p \in (0,1)$.

Entries of adjacency matrix:

$$a_{jk} = \begin{cases} \xi_{jk}, & 1 \le j < k \le n \\ \xi_{kj}, & 1 \le k < j \le n \\ 0, & j = k \end{cases} \qquad \xi_{jk} \sim \text{Bernoulli}(p) \text{ iid.}$$

Let E_{jk} be matrix with 1 at entry (j,k) and zero everywhere else. Then

$$A = \sum_{1 \le j < k \le n} \xi_{jk} (E_{jk} + E_{kj})$$

and

$$\Delta = \sum_{1 \le j < k \le n} \xi_{jk} (E_{jj} + E_{kk} - E_{jk} - E_{kj}).$$

Matrix Chernoff: Application to Erdős–Rényi graphs

Aim: Understand when Erdős–Rényi graphs are connected by applying matrix Chernoff.

Problem: Need to control second smallest eigenvalue of Δ, not smallest.

Solution: Deflate smallest eigenvalue. Let the columns of $R \in \mathbb{R}^{n\times(n-1)}$ contain any ONB of $\operatorname{span}(e)^\perp$. Then $R^\top R = I_{n-1}$ and $Re = 0$.

Second smallest eigenvalue of Δ = smallest eigenvalue of

$$Y = R^\top \Delta R = \sum_{1\leq j<k\leq n} \xi_{jk} R^\top (E_{jj} + E_{kk} - E_{jk} - E_{kj}) R =: \sum_{1\leq j<k\leq n} X_{jk}.$$

X_{jk} are independent, bounded SPSD matrices.

Quantities relevant for matrix Chernoff:

- $\|X_{ij}\|_2 \leq 2 =: L$ (EFY: Why?)
- $\mathbb{E}Y = p \sum_{1\leq j<k\leq n} R^\top (E_{jj} + E_{kk} - E_{jk} - E_{kj}) R =$
  $pR^\top\big((n-1)I - (ee^\top - I)\big)R = pnI_{n-1}$

Matrix Chernoff: Application to Erdős–Rényi graphs

By matrix Chernoff tail bound for $\lambda_{\min}(Y)$, we obtain that

$$\mathbb{P}\left\{\lambda_2(\Delta) \le \epsilon \cdot tn\right\} = \mathbb{P}\left\{\lambda_{\min}(Y) \le \epsilon \cdot tn\right\} \le (n-1)\left(\frac{e^{\epsilon-1}}{\epsilon^{\epsilon}}\right)^{pn/2}$$

For small ϵ, $e^{\epsilon-1}/\epsilon^{\epsilon} \approx e^{-1}$. Hence, second-smallest eigenvalue of Δ is unlikely to be zero when $\log(n-1) - pn/2 < 0$ or, equivalently,

$$p > \frac{2\log(n-1)}{n}.$$

EFY: Study the sharpness of this rule by a “phase transition diagram”.