Deck 7: Matrix concentration: Matrix Bernstein

Math 7870: Topics in Randomized Numerical Linear Algebra

Spring 2026

Akil Narayan

Matrix concentration inequalities

A review:

- $\boldsymbol{X}_j$, $j \in [n]$ are $m \times m$, independent, positive semi-definite matrices. $\boldsymbol{S}_n = \sum_{j\in[n]} \boldsymbol{X}_j$.
- We have master bounds, e.g., $\Pr(\lambda_{\min}(\boldsymbol{S}_n) \leqslant t) \leqslant \inf_{s<0} e^{-st} \operatorname{exp\,tr}\left(\sum_{j\in[n]} \boldsymbol{K}_{\boldsymbol{X}_j}(s)\right)$
- We've established that tr exp is a monotone function on spd matrices.
- We used this to derive a Chernoff bound for spd matrices, e.g., if $\|\boldsymbol{X}_j\|_2 \leqslant B$ with probability 1, then the master bound delivers for an absolute constant $c > 0$:

$$\Pr(\lambda_{\min}(\boldsymbol{S}_n) \leqslant (1-\epsilon)\lambda_{\min}(\mathbb{E}\boldsymbol{S}_n)) \leqslant m \exp\left(-c\frac{\lambda_{\min}(\mathbb{E}\boldsymbol{S}_n)}{B}\epsilon^2\right),$$

We will now work toward generalizing this result.

A Legendre-Fenchel perspective

We can identify a convex duality argument similar to the scalar case.
The benefit here is that we can allow indefinite (Hermitian) matrices.

Theorem 1. *Suppose $\boldsymbol{X}_j$, $j \in [n]$ is a sequence of $m \times m$, independent, Hermitian random matrices. Suppose that, for some deterministic spd matrices $\{\boldsymbol{A}_j\}_{j\in[n]}$, we have,*

$$\boldsymbol{M}_{\boldsymbol{X}_j}(s) \preceq \exp\left(f(s)\boldsymbol{A}_j\right), \qquad f:(0,\infty) \to [0,\infty],$$

Then:

$$\Pr(\lambda_{\max}(\boldsymbol{S}_n) \geqslant t) \leqslant m \exp\left(-rf^*(t/r)\right), \qquad r := \lambda_{\max}\Big(\sum_{j\in[n]} \boldsymbol{A}_j\Big),$$

where f^* is the convex conjugate of f, and $\boldsymbol{S}_n = \sum_{j\in[n]} \boldsymbol{X}_j$.

Compare this to the scalar concentration.
(Here, $f^*(t/r)$ is the rate function, and $r \sim n$ and $t \leftarrow nt$ for the iid case.)

The proof

We "just" have to use the master bounds:

$$\begin{aligned}
\Pr(\lambda_{\max}(\boldsymbol{S}_n) \geqslant t) &\leqslant \inf_{s>0} e^{-st} \operatorname{tr} \exp\left(\sum_{j\in[n]} \boldsymbol{K}_{\boldsymbol{X}_j}(s)\right) \\
&\leqslant \inf_{s>0} e^{-st} \operatorname{tr} \exp\left(f(s)\sum_j \boldsymbol{A}_j\right) \\
&\leqslant m \inf_{s>0} e^{-st} \lambda_{\max} \exp\left(f(s)\sum_j \boldsymbol{A}_j\right) \\
&= m \inf_{s>0} \exp\left(-st + f(s)r\right) \\
&= m \exp\left(-\sup_{s>0}(st - rf(s))\right).
\end{aligned}$$

For functions f for which f^* is computable, this is a reasonable bound.

However, the assumption $\boldsymbol{M}_{\boldsymbol{X}_j(s)} \preceq \exp(f(x)\boldsymbol{A}_j)$ is not very transparent.

On to Matrix Bernstein-type estimates

We'll now consider how to come up with a more practical bound under the assumption that the random summands are bounded.

As is (by now?) usual, the exercise boils down estimating the matrix CGF.

Lemma. *Assume $\boldsymbol{X}$ is a random Hermitian matrix, and that $\mathbb{E}\boldsymbol{X} = \mathbf{0}$ and $\lambda_{\max}(\boldsymbol{X}) \leqslant B$ with probability 1.*

Then for $s \in (0, 3/L)$:

$$\boldsymbol{K}_{\boldsymbol{X}}(s) \preceq g(s)\mathbb{E}\boldsymbol{X}^2, \qquad g(s) := \frac{1}{2}\frac{s^2}{(1 - sL/3)}$$

The proof relies on some technical estimates of g and of a univariate function f. In particular for some $s \in (0, 3/L)$:

$$f(x) := \frac{e^{sx} - sx - 1}{x^2}$$
$$x \leqslant B \quad \text{implies} \quad f(x) \leqslant f(B)$$
$$f(x) \leqslant g(x)$$

Finishing the proof

The rest of the proof uses some familiar properties of the Loewner ordering and operator monotone functions:

$$\exp(s\boldsymbol{X}) = \boldsymbol{I} + s\boldsymbol{X} + (e^{s\boldsymbol{X}} - s\boldsymbol{X} - \boldsymbol{I}) = \boldsymbol{I} + s\boldsymbol{X} + \boldsymbol{X} f(\boldsymbol{X}) \boldsymbol{X}.$$

Since $\lambda_{\max}(\boldsymbol{X}) \leqslant B$, then $\boldsymbol{X} \preceq B\boldsymbol{I}$, so $f(\boldsymbol{X}) \preceq \boldsymbol{I} f(B)$. Therefore:

$$\begin{aligned} \exp(s\boldsymbol{X}) = \boldsymbol{I} + s\boldsymbol{X} + \boldsymbol{X} f(\boldsymbol{X}) \boldsymbol{X} &\preceq \boldsymbol{I} + s\boldsymbol{X} + f(B) \boldsymbol{X} \boldsymbol{I} \boldsymbol{X} \\ &= \boldsymbol{I} + s\boldsymbol{X} + f(B) \boldsymbol{X}^2 \\ &\preceq \boldsymbol{I} + s\boldsymbol{X} + g(B) \boldsymbol{X}^2 \end{aligned}$$

Take expectations on both sides, use the monotone operator property of log.

Toward Hermitian matrix Bernstein

Armed with this inequality, we can prove a matrix Bernstein inequality.

The proof pieces together several results: Suppose $\boldsymbol{X}_j$ are iid Hermitian, centered, and that $\lambda_{\max}(\boldsymbol{X}_j) \leqslant B$ with probability 1 for every $j \in [n]$.

Then, e.g., for the upper probability bound:

$$\begin{aligned}
\Pr(\lambda_{\max}(\boldsymbol{S}_n) \geqslant t) &\leqslant \inf_{s>0} e^{-st} \operatorname{tr} \exp\left(\sum_{j\in[n]} \boldsymbol{K}_{\boldsymbol{X}_j}(s)\right) \\
&\leqslant \inf_{0<s<3/B} e^{-st} \operatorname{tr} \exp\left(g(s) \sum_{j\in[n]} \mathbb{E}\boldsymbol{X}_j^2\right) \\
&\leqslant m \inf_{0<s<3/B} e^{-st} \exp\left(g(s)\lambda_{\max}\left(\sum_{j\in[n]} \mathbb{E}\boldsymbol{X}_j^2\right)\right) \\
&= m \inf_{0<s<3/B} e^{-st} \exp\left(g(s)v(\boldsymbol{S}_n)\right) \quad (v(\boldsymbol{S}_n) := \|\boldsymbol{E}\boldsymbol{S}_n^2\|_2)
\end{aligned}$$

We can then minimize over s.

Hermitian matrix Bernstein

Theorem 2 (Hermitian matrix Bernstein)**.** *Suppose $\boldsymbol{X}_j$ are iid Hermitian, centered, and that $\lambda_{\max}(\boldsymbol{X}_j) \leqslant B$ with probability 1 for every $j \in [n]$. Define,*

$$\boldsymbol{S}_n = \sum_{j \in [n]} \boldsymbol{X}_j, \qquad v(\boldsymbol{S}_n) = \|\mathbb{E}\boldsymbol{S}_n^2\|_2.$$

Then, e.g.,

$$\Pr(\lambda_{\max}(\boldsymbol{S}_n) \geqslant t) \leqslant m \exp\left(-\frac{1}{2}\frac{t^2}{v(\boldsymbol{S}_n) + Bt/3}\right),$$

$$\mathbb{E}\lambda_{\max}(\boldsymbol{S}_n) \leqslant \sqrt{2v(\boldsymbol{S}_n)\log m} + \frac{B}{3}\log m$$

Bounded rectangular matrix Bernstein, I

With a matrix dilation, we can extend this to non-Hermitian matrices $\boldsymbol{X}_j$.

Let

$$\boldsymbol{Y}_j = \mathcal{D}(\boldsymbol{X}_j) = \begin{pmatrix} \mathbf{0} & \boldsymbol{X}_j \\ \boldsymbol{X}_j^* & \mathbf{0} \end{pmatrix}$$

With

$$\boldsymbol{S}_n = \sum_j \boldsymbol{X}_j, \qquad \boldsymbol{T}_n = \sum_j \boldsymbol{Y}_j$$

From the properties of matrix dilation:

$$\boldsymbol{T}_n = \mathcal{D}(\boldsymbol{S}_n), \qquad \lambda_{\max}(\boldsymbol{T}_n) = \|\boldsymbol{S}_n\|_2.$$

The generic strategy is then to bound the deviation of $\lambda_{\max}(\boldsymbol{T}_n)$ from $\lambda_{\max}(\mathbb{E}\boldsymbol{T}_n)$ using Hermitian Bernstein, and then translate quantities involving $\boldsymbol{Y}_j$ into those depending on $\boldsymbol{X}_j$.

Bounded rectangular matrix Bernstein, II

Theorem 3. *Let $\boldsymbol{X}_j$, $j \in [n]$ be a sequence of $m_1 \times m_2$ independent, centered, random matrices. Assume,*

$$\|\boldsymbol{X}_j\|_2 \leqslant B, \quad \textit{with probability 1 for all } j \in [n].$$

Define the variance statistic

$$v(\boldsymbol{S}_n) := \max\left\{\|\mathbb{E}\boldsymbol{S}_n\boldsymbol{S}_n^*\|_2\,, \|\mathbb{E}\boldsymbol{S}_n^*\boldsymbol{S}_n\|_2\right\}.$$

Then:

$$\Pr(\|\boldsymbol{S}_n\|_2 \geqslant t) \leqslant (m_1 + m_2)\exp\left(-\frac{1}{2}\frac{t^2}{(v(\boldsymbol{S}_n) + Bt/3)}\right), \qquad t > 0$$

$$\mathbb{E}\|\boldsymbol{S}_n\|_2 \leqslant \sqrt{2v(\boldsymbol{S}_n)\log(m_1 + m_2)} + \frac{B}{3}\log(m_1 + m_2).$$

NB: $d_1 + d_2$ is the size of the dilation $\mathcal{D}(\boldsymbol{S}_n)$.
Since $\mathbb{E}\boldsymbol{S}_n = \boldsymbol{0}$, the above quantifies deviation of $\boldsymbol{S}_n$ from its mean.

Subexponential matrices?

Can we extend these beyond matrices with bounded norms?

In the scalar case: suppose X is an exponential random variable with rate parameter $\lambda > 0$. By direct computation:

$$f_X(x) \sim e^{-\lambda x}, \qquad \mathbb{E}X = 1/\lambda, \qquad \mathbb{E}X^p = \frac{p!}{\lambda^p} \sim c^p p!$$

A subexponential random variable colloquiually is one with tails that are sufficiently light so the can be dominated by an exponential.

Equivalently, a random variable is subexponential if its moments grow at most as fast as an exponential random variable.

The more common moment bound definition for scalars is not the factorial expression above. Rather, by Stirling's approximation, $p! \sim \sqrt{p}\frac{p^p}{e^p}$, so

$$(\mathbb{E}X^p)^{1/p} \lesssim cp,$$

for some constant c.
However, for the matrix setting, a version of the factorial bound is the more convenient definition.

Subexponential matrix Bernstein

Theorem 4 (Subexponential Matrix Bernstein). *Let $\boldsymbol{X}_j$, $j \in [n]$ be independent, $m \times m$ self-adjoint matrices, and assume,*

$$\mathbb{E}\boldsymbol{X}_j = \mathbf{0}, \qquad \mathbb{E}\boldsymbol{X}_j^k \preceq \frac{1}{2} p!\, R^{p-2} \boldsymbol{A}_j^2,$$

for all $j \in [n]$, integer $p \geqslant 2$. Then by defining,

$$\sigma^2 = \left\| \sum_{j \in [n]} \boldsymbol{A}_j^2 \right\|_2,$$

we have

$$\Pr\left(\lambda_{\max}(\boldsymbol{S}_n) \geqslant t\right) \leqslant m \exp\left(-\frac{1}{2} \frac{t^2}{(\sigma^2 + Rt)}\right).$$

The bound above satisfies:

$$\Pr\left(\lambda_{\max}(\boldsymbol{S}_n) \geqslant t\right) \leqslant m \begin{cases} \exp(-t^2/4\sigma^2), & \text{when } t \leqslant \sigma^2/R \\ \exp(-t/4R), & \text{when } t \geqslant \sigma^2/R \end{cases}$$

Subexponential matrix Bernstein CGF bound

To prove the subexponential matrix Bernstein, the crucial estimate to prove is a bound on the matrix CGF.

Lemma (Subexponential matrix Bernstein CGF)**.** *Suppose* $\boldsymbol{X}$ *is a random centered, self-adjoint matrix satisfying* $\mathbb{E}\boldsymbol{X}^p \preceq \frac{p!}{2}\boldsymbol{A}^2$ *for all integers* $p \geqslant 2$*. Then,*

$$\boldsymbol{K}_{\boldsymbol{X}}(s) \preceq \frac{s^2}{2(1-s)}\boldsymbol{A}^2, \qquad s \in (0,1).$$

Proof: for $0 < s < 1$,

$$\begin{aligned}
\boldsymbol{M}_{\boldsymbol{X}}(s) = \sum_{p=0}^{\infty} \frac{s^p \mathbb{E}\boldsymbol{X}^p}{p!} &= \boldsymbol{I} + s\mathbb{E}\boldsymbol{X} + \sum_{p=2}^{\infty} \frac{s^p \mathbb{E}\boldsymbol{X}^p}{p!} \\
&\preceq \boldsymbol{I} + \frac{1}{2}\boldsymbol{A}^2 \sum_{p=2}^{\infty} s^p \\
&= \boldsymbol{I} + \frac{s^2}{2(1-s)}\boldsymbol{A}^2 \preceq \exp\left(\frac{s^2}{2(1-s)}\boldsymbol{A}^2\right).
\end{aligned}$$

The rest of the subexponential matrix Bernstein proof is the same as before.

Generalizations

Here are some generalizations:

- Rectangular subexponential matrix Bernstein
- Hermitian matrix Azuma's inequality
  (This requires somewhat different machinery than in the scalar case.)
- Matrix Hoeffding
  (This turns out to be a consequence of the proof of matrix Azuma)
- Rectangular matrix Azuma
- Hermitian matrix McDiarmid

Application setup: Random matrix approximation

Suppose $\boldsymbol{A} \in \mathbb{C}^{m_1 \times m_2}$ is a deterministic matrix. We consider the template procedure of approximating $\boldsymbol{A}$ by sums of iid copies of a random matrix $\boldsymbol{X}$. I.e., we consider

$$\boldsymbol{S}_n = \sum_{j \in [n]} \frac{1}{n} \boldsymbol{X}_j, \qquad \boldsymbol{X}_j \overset{\text{iid}}{\sim} \boldsymbol{X}, \qquad \mathbb{E}\boldsymbol{X} = \boldsymbol{A}.$$

Lemma (Matrix approximation via sampling)**.** *With the above setup, assume* $\|\boldsymbol{X}\|_2 \leqslant B$ *with probability 1, and define,*

$$v_2(\boldsymbol{X}) = \max\left\{\|\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^*)\|_2, \|\boldsymbol{X}^*\boldsymbol{X}\|_2\right\}.$$

Then:

$$\mathbb{E}\,\|\boldsymbol{S}_n - \boldsymbol{A}\|_2 \leqslant \sqrt{\frac{v_2(\boldsymbol{X})\log(m_1+m_2)}{n}} + \frac{2B\log(m_1+m_2)}{3n},$$

$$\Pr(\|\boldsymbol{S}_n - \boldsymbol{A}\|_2 \geqslant t) \leqslant (m_1+m_2)\exp\left(-\frac{1}{2}\frac{nt^2}{v_2(\boldsymbol{X}) + 2Bt/3}\right), \quad t > 0.$$

Proof steps: (i) Write $\boldsymbol{S}_n - \boldsymbol{A}$ as an iid sum of centered summands, (ii) compute uniform bounds on the centered summand ($2B/n$), (iii) compute the variance statistic v of the centered summands ($v_2(\boldsymbol{X})/n$). Apply (bounded) matrix Bernstein.

Application I: Randomized matrix multiplication, I

Let's tie up a loose end: randomized matrix multiplcation.

Recall: Let $\boldsymbol{A}, \boldsymbol{B}$ be two matrices of sizes $\boldsymbol{A} \in \mathbb{C}^{m_1 \times k}$, $\boldsymbol{B} \in \mathbb{C}^{k \times m_2}$.

If $\boldsymbol{a}_j$, $\boldsymbol{b}_j$, $j \in [k]$, are the rows and columns of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, then we define a rank-1 matrix $\boldsymbol{X}$ as:

$$\boldsymbol{X} = \frac{1}{p_j} \boldsymbol{a}_j \boldsymbol{b}_j^* \quad \text{w/ probability } p_j,\ j \in [k].$$

Then $\mathbb{E}\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}$.

Also recall: the choice,

$$p_j = \frac{\|\boldsymbol{a}_j\|_2 \|\boldsymbol{b}_j\|_2}{\sum_{q \in [k]} \|\boldsymbol{a}_q\|_2 \|\boldsymbol{b}_q\|_2},$$

minimizes a variance. For computational convenience, we'll choose the slightly different distribution:

$$p_j = \frac{\|\boldsymbol{a}_j\|_2^2 + \|\boldsymbol{b}_j\|_2^2}{\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2}$$

(The latter p_j are comparable to the former.)

Application I: Randomized matrix multiplication, II

Our estimator is $\boldsymbol{S}_n = \sum_{j\in[k]}(\frac{1}{n}\boldsymbol{X}_j)$, with $\boldsymbol{X}_j \overset{\text{iid}}{\sim} \boldsymbol{X}$.

Some computations reveal:

$$\|\boldsymbol{X}\|_2 \leqslant \frac{1}{2}\left(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2\right) = B.$$
$$v_2(\boldsymbol{X}) \leqslant (\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2)\max\{\|\boldsymbol{A}\|_2^2, \|\boldsymbol{B}\|_2^2\}.$$

Then, e.g., in expectation we obtain,

$$\frac{1}{\|\boldsymbol{A}\|_2^2 + \|\boldsymbol{B}\|_2^2}\mathbb{E}\,\|\boldsymbol{S}_n - \boldsymbol{A}\boldsymbol{B}\|_2 \lesssim \sqrt{\frac{2\tau}{n}} + \frac{\tau}{3n}, \qquad \tau = (\operatorname{srank}(\boldsymbol{A}) + \operatorname{srank}(\boldsymbol{B}))\log(m_1 + m_2),$$

where, srank is the *stable* or numerical rank, defined as,

$$\operatorname{srank}(\boldsymbol{A}) = \frac{\|\boldsymbol{A}\|_F^2}{\|\boldsymbol{A}\|_2^2} = \frac{\sum_j \sigma_j(\boldsymbol{A})^2}{\sigma_1^2(\boldsymbol{A})} \in [1, \min\{k, m_1\}].$$

(The expected statements hold in probability: sub-Gaussian tails for $\|\boldsymbol{S}_n - \boldsymbol{A}\boldsymbol{B}\|_2$ close to 0, sub-exponential tails when it's far from 0.)

Application II: Random sparsification of a matrix, I

We can now consider a slightly more complicated, and useful, numerical linear algebra application: matrix sparsification.

Let $\boldsymbol{A} \in \mathbb{C}^{m_1 \times m_2}$ be a deterministic matrix. We seek to compute an approximation to $\boldsymbol{A}$ via sparsification, i.e., if we write,

$$\boldsymbol{A} = \sum_{(i,j) \in [m_1] \times [m_2]} A_{i,j} \boldsymbol{e}_i \boldsymbol{e}_j^*.$$

Given a set $S \subset [m_1] \times [m_2]$, with $|S| = n$, we consider,

$$\boldsymbol{A}_S = \sum_S A_{i,j} \boldsymbol{e}_i \boldsymbol{e}_j^*, \qquad \text{Is } \|\boldsymbol{A} - \boldsymbol{A}_S\|_2 \sim \text{“small”?}$$

If $n = |S| \ll m_1 m_2$, this is a sparsification of $\boldsymbol{A}$.
We'll construct S randomly.

Note that best possible fixed-rank approximation in the 2-norm is given by the truncated SVD, and the trailing singular values provide an estimate of error.

Therefore it's reasonable to believe that the trailing singular values play a role in estimating error. In particular, we'll use the stable rank $\operatorname{srank}(\boldsymbol{A})$.

Application II: Random sparsification of a matrix, II

We construct the subset S randomly by selecting indices. This is similar to how we constructed matrix multiplication. Let

$$\boldsymbol{X} = \frac{1}{p_{i,j}} A_{i,j} \boldsymbol{e}_i \boldsymbol{e}_j^*, \qquad (i,j) = (k,\ell) \quad \text{with probability } p_{k,\ell},$$

where we choose the probabilities to be of the form,

$$2p_{i,j} = \frac{|A_{i,j}|^2}{\|\boldsymbol{A}\|_F^2} + \frac{|A_{i,j}|}{\|\boldsymbol{A}\|_1}, \qquad (i,j) \in [m_1] \times [m_2].$$

The estimator is $\boldsymbol{S}_n = \frac{1}{n} \sum_{j \in [n]} \boldsymbol{X}_j$, with $\boldsymbol{X}_j \overset{\text{iid}}{\sim} \boldsymbol{X}$.

We follow a similar trajectory:

$$\mathbb{E}\boldsymbol{X} = \boldsymbol{A}, \qquad B = \max_{i,j} \|\boldsymbol{X}\|_2 \leqslant 2\|\boldsymbol{A}\|_1, \qquad v_2(\boldsymbol{X}) \leqslant 2\max\{m_1, m_2\}.$$

Then, e.g.,

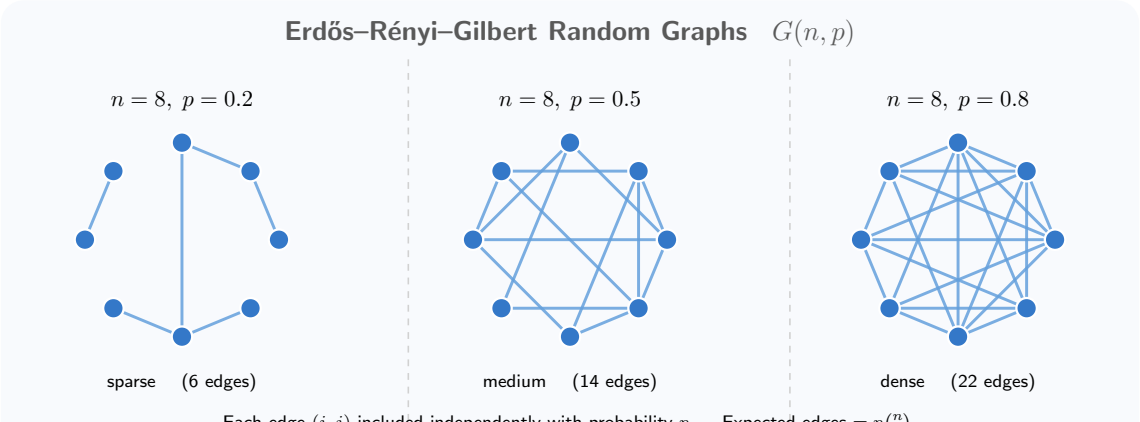
$$\frac{1}{\|\boldsymbol{A}\|_2} \mathbb{E}\|\boldsymbol{A} - \boldsymbol{S}_n\|_2 \lesssim \frac{\|\boldsymbol{A}\|_F}{\|\boldsymbol{A}\|_2} \sqrt{\frac{\max\{m_1, m_2\} \log(m_1 + m_2)}{n}}$$

For example, this imples that $n \gtrsim \operatorname{srank}(\boldsymbol{A})/\epsilon^2$ guarantees a relative expected error of 4ϵ.

Application III: Connectedness of a random graph, I

An Erdös-Rényi-Gilbert (ERG) graph is one where, given a fixed number of vertices, every edge is present according to an iid Bernoulli draw.

Given $n \in \mathbb{N}$ vertices, and a density parameter $p \in [0, 1]$, $G(n, p)$ is the ERG model where a graph is constructed over the n vertices by sampling an independent Bernoulli(p) random variable on each edge to determine if it's present or not.



Application III: Connectedness of a random graph, II

There are $\sim n^2/2$ possible edges over n vertices.

It's well-known that if $p \sim (\log n)/n$, then the probability that the graph is connected is high. (A graph is connected if edges are present to allow you to travel to one vertex from any other one.)

We can show this result using our matrix concentration results.

We'll need some background from (spectral) graph theory.

Given a graph G with vertices $[n]$, the *degree* d_j of vertex $j \in [n]$ is the number of vertices directly connected to it. (Equivalently, the number of edges connected to j.)

The **degree matrix** of G is the diagonal matrix:

$$\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_n).$$

Application III: Connectedness of a random graph, III

With the vertices $[n]$ of G, the **adjacency matrix** $\boldsymbol{A}$ of G encodes the edges:

$$\boldsymbol{A} \in \mathbb{R}^{n\times n}, \qquad (A)_{i,j} = \begin{cases} 0, & i \text{ not connected to } j \\ 1, & i \text{ connected to } j \end{cases}$$

The matrix $\boldsymbol{A}$ is Hermitian (like $\boldsymbol{D}$).

Finally, we need the (unnormalized) *Graph Laplacian*:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \succeq \boldsymbol{0}.$$

Since $\boldsymbol{L}$ is spd, its smallest possible eigenvalue is 0.
It turns out: any connected component in G results in a unique eigenvector with eigenvalue 0.
I.e., G is fully connected iff $\lambda = 0$ is a simple eigenvalue, or if the *second* smallest eigenvalue of $\boldsymbol{L}$ is strictly positive.

Application III: Connectedness of a random graph, IV

By explicit computation, the pair $(0, \mathbf{1})$ is *always* an eigenpair of any graph Laplacian.

With this knowledge, we can consider the graph Laplacian matrix corresponding to a compression that's orthogonal to $\mathbf{1}$.

$$\boldsymbol{S} := \boldsymbol{Q}^* \boldsymbol{L} \boldsymbol{Q}, \qquad \boldsymbol{Q}^* \boldsymbol{Q} = \boldsymbol{I}_{n-1}, \qquad \boldsymbol{Q}^* \mathbf{1} = \mathbf{0}.$$

The smallest eigenvalue of $\boldsymbol{S}$ is equal to the second smallest eigenvalue of $\boldsymbol{L}$.

What's the probability that the smallest eigenvalue of $\boldsymbol{S}$ is larger than 0?

To connect this to random matrices, suppose that G is from the $G(n, p)$ ERG model.
In this model, given a unique pair $\{i, j\}$ of distinct vertices:

- An edge connecting $\{i, j\}$ is activated according to the outcome of an indepedent RV $\gamma_{i,j} \sim \text{Bernoulli}(p)$.
- Both d_i and d_j are incremented by 1
- The (i, j) and (j, i) entries of the adjacency matrix take value 1.

Therefore:

$$\boldsymbol{L} = \sum_{i=1}^{n} \sum_{j<i} \gamma_{i,j} \left(\boldsymbol{e}_i^* \boldsymbol{e}_i + \boldsymbol{e}_j^* \boldsymbol{e}_j - \boldsymbol{e}_i^* \boldsymbol{e}_j - \boldsymbol{e}_j^* \boldsymbol{e}_i \right).$$

Application III: Connectedness of a random graph, V

$$\boldsymbol{S} := \boldsymbol{Q}^* \boldsymbol{L} \boldsymbol{Q} = \sum_{i \in [n]} \sum_{j<i} \boldsymbol{X}_{i,j}, \qquad \boldsymbol{L} = \sum_{i=1}^{n} \sum_{j<i} \gamma_{i,j} \left(\boldsymbol{e}_i^* \boldsymbol{e}_i + \boldsymbol{e}_j^* \boldsymbol{e}_j - \boldsymbol{e}_i^* \boldsymbol{e}_j - \boldsymbol{e}_j^* \boldsymbol{e}_i \right).$$

Therefore, $\boldsymbol{S}$ is an iid sum of $n(n-1)/2$ random (spd) matrices. One can compute:

$$\mathbb{E}\boldsymbol{S} = pn\boldsymbol{I}_{n-1}, \qquad \lambda_{\min}(\mathbb{E}\boldsymbol{S}) = pn, \qquad \|\boldsymbol{X}_{i,j}\|_2 \leqslant 2.$$

This is enough to employ the multiplicative spd Chernoff estimate:

$$\Pr(\lambda_{\min}(\boldsymbol{S}) \leqslant tpn) \leqslant (n-1) e^{-pn/2} \underbrace{\left(\frac{e}{t}\right)^{tpn/2}}_{\sim 1 \text{ for } t \ll 1}$$

We want this probability to be small, which translates into $p \gtrsim \frac{2\log(n-1)}{n}$.
(This ensures that the second smallest eigenvalue of $\boldsymbol{L}$ is positive with large probability.)

Application IV: Random features, I

Let $\{\boldsymbol{p}_j\}_{j\in[N]} \subset \mathbb{R}^P$ be a set of N points in P-dimensional space. (Formally $P = \infty$ is allowable.)

These points correspond to observed data, and the goal is to develop pairwise comparisons for downstream tasks (classification, identification, compression, etc.)

The *kernel trick* embeds such pairwise comparisons as scalars through a *kernel function*:

$$A_{i,j} = K(\boldsymbol{p}_i, \boldsymbol{p}_j), \qquad K : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{C}$$

The function K is symmetric (e.g., $K(\boldsymbol{p}, \boldsymbol{q}) \sim \exp(-\|\boldsymbol{p} - \boldsymbol{q}\|_2^2)$), so that $\boldsymbol{A}$ is Hermitian.

The problem: Evaluating $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is expensive, requiring say $\mathcal{O}(N^2P)$.

(P-dependent complexity arises when evaluating function over P inputs.)

Application IV: Random features, II

A particular class of kernels of interest are translation-invariant kernels on $\mathbb{R}^P$, which are given by,

$$K(\boldsymbol{p}, \boldsymbol{q}) = \phi(\boldsymbol{p} - \boldsymbol{q}), \qquad \phi : \mathbb{R}^P \to \mathbb{C}.$$

Through Bochner's Theorem, these translation-invariant kernels can be represented as,

$$\phi(\boldsymbol{p} - \boldsymbol{q}) = c \int \exp\left(i \langle \boldsymbol{p} - \boldsymbol{q}, \boldsymbol{w} \rangle\right) d\mu(\boldsymbol{w}),$$

for a constant c and a measure μ on $\mathbb{R}^P$. Written another way, if $\boldsymbol{w}$ is a random vector whose density function is $d\mu(\boldsymbol{w})$, then

$$A_{i,j} = \mathbb{E}\left(\psi(\boldsymbol{p}_i, \boldsymbol{w})\psi(\boldsymbol{p}_j, \boldsymbol{w})^*\right), \qquad \psi(\boldsymbol{p}, \boldsymbol{w}) = \sqrt{c} \exp\left(i \langle \boldsymbol{p}, \boldsymbol{w} \rangle\right).$$

Application IV: Random features, III

To demonstrate that this characterization has concrete applicability:

$$K(\boldsymbol{p}, \boldsymbol{q}) = \exp(-a\|\boldsymbol{p} - \boldsymbol{q}\|_2^2/2) \quad \Longrightarrow \quad \psi(\boldsymbol{p}, (\boldsymbol{w}, U)) = \sqrt{2}\cos\left(\langle \boldsymbol{p}, \boldsymbol{w}\rangle + U\right),$$

where $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \alpha \boldsymbol{I})$ and $U \sim \text{uniform}([0, 2\pi))$.

Hence, we know that $\boldsymbol{A}$ is the expectation of a random variable $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*, \qquad x_j = \psi(\boldsymbol{p}_j; \boldsymbol{w}), \qquad \mathbb{E}\boldsymbol{X} = \boldsymbol{A}.$$

Therefore, we develop a familiar strategy and perform some computations:

$$\boldsymbol{S}_n = \frac{1}{n}\sum_{j\in[n]} \boldsymbol{X}_j, \qquad \boldsymbol{X}_j \overset{\text{iid}}{\sim} \boldsymbol{X}, \qquad \|\boldsymbol{X}\|_2 \leqslant CN, \qquad v_2(\boldsymbol{X}) \leqslant CN\|\boldsymbol{A}\|_2,$$

where $|\psi(\boldsymbol{p}, \boldsymbol{w})| \leqslant C$.

These imply that $\mathbb{E}\|\boldsymbol{S}_n - \boldsymbol{A}\|_2 \lesssim \sqrt{CN\|\boldsymbol{A}\|_2 \log(2N)}$.
This translates into $n \gtrsim C\,\text{idim}(\boldsymbol{A})/\epsilon^2$ implies $\frac{1}{\|\boldsymbol{A}\|_2}\mathbb{E}\|\boldsymbol{S}_n - \boldsymbol{A}\|_2 \lesssim \epsilon$.