Lecture 12: Concentration of sums of independent random matrices

School of Mathematical Sciences, Xiamen University

1. Matrix calculus

- Let us focus on $n \times n$ symmetric matrices.

  $\mathbf{A} \succeq \mathbf{B}$ (of course, $\mathbf{B} \preceq \mathbf{A}$) means $\mathbf{A} - \mathbf{B}$ is positive semidefinite.

  This defines a *partial order* on the set of $n \times n$ symmetric matrices.

- $\|\mathbf{A}\|_2$ is the smallest non-negative number M such that

$$\|\mathbf{A}\mathbf{x}\|_2 \leqslant M\|\mathbf{x}\|_2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

- Functions of matrices

  Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a function and $\mathbf{X}$ be an $n \times n$ symmetric matrix. We define

$$f(\mathbf{X}) := \sum_{i=1}^{n} f(\lambda_i)\mathbf{u}_i\mathbf{u}_i^\top$$

  where λ_i are the eigenvalues of $\mathbf{X}$ and $\mathbf{u}_i$ are the corresponding eigenvectors satisfying

$$\mathbf{X} = \sum_{i=1}^{n} \lambda_i\mathbf{u}_i\mathbf{u}_i^\top.$$

2. Matrix Bernstein's inequality

Theorem 1 (Matrix Bernstein's inequality)

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that

$$\|\boldsymbol{X}_i\|_2 \leq K$$

almost surely for all i. Then, for every $t \geq 0$ we have

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N} \boldsymbol{X}_i\right\|_2 \geqslant t\right\} \leqslant 2n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Here

$$\sigma^2 = \left\|\sum_{i=1}^{N} \mathbb{E}\boldsymbol{X}_i^2\right\|_2$$

is the norm of the sum of the "matrix variance".

- The scalar case, where $n = 1$, is the classical Bernstein's inequality for random variables.
- A remarkable feature of matrix Bernstein's inequality, which makes it especially powerful, is that it does not require any independence of the entries (or the rows or columns) of $\boldsymbol{X}_i$; all is needed is that the random matrices $\boldsymbol{X}_i$ be independent from each other.
- The proof of Theorem 1 is based on bounding the moment generating function (MGF) $\mathbb{E}\exp(\lambda \boldsymbol{S})$ of the sum $\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{X}_i$.
- Golden–Thompson inequality: For any symmetric $\mathbf{X}$ and $\mathbf{Y}$,

$$\operatorname{tr}\left(\mathrm{e}^{\mathbf{X}+\mathbf{Y}}\right) \leqslant \operatorname{tr}\left(\mathrm{e}^{\mathbf{X}}\mathrm{e}^{\mathbf{Y}}\right).$$

- Lieb's lemma: For given $n \times n$ symmetric $\mathbf{H}$, the function

$$f(\mathbf{X}) = -\operatorname{tr}\exp(\mathbf{H} + \log \mathbf{X})$$

is convex on the space of $n \times n$ SPD matrices.

- Jensen's inequality:

  For any convex function f and a random matrix $\boldsymbol{X}$, it holds

$$f(\mathbb{E}\boldsymbol{X}) \leq \mathbb{E}f(\boldsymbol{X}).$$

- Lieb's inequality for random matrices

  Let $\mathbf{H}$ be a fixed $n \times n$ symmetric matrix and $\boldsymbol{Z}$ be an $n \times n$ symmetric random matrix. Then

$$\mathbb{E}\operatorname{tr}\exp(\mathbf{H}+\boldsymbol{Z}) \leqslant \operatorname{tr}\exp\left(\mathbf{H}+\log\mathbb{E}\mathrm{e}^{\boldsymbol{Z}}\right).$$

- MGF of a sum of independent random matrices

  Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ be independent $n \times n$ symmetric random matrices. Then the sum $\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{X}_i$ satisfies

$$\mathbb{E}\operatorname{tr}\exp(\lambda\boldsymbol{S}) \leq \operatorname{tr}\exp\left[\sum_{i=1}^{N}\log\mathbb{E}\mathrm{e}^{\lambda\boldsymbol{X}_i}\right].$$

Lemma 2

Let $\boldsymbol{X}$ *be an* $n \times n$ *symmetric random matrix. Assume that*

$$\mathbb{E}\boldsymbol{X} = \mathbf{0}$$

and

$$\|\boldsymbol{X}\|_2 \leq K$$

almost surely. Then, for all

$$0 < \lambda < \frac{3}{K}$$

we have

$$\mathbb{E}\exp(\lambda \boldsymbol{X}) \preceq \exp\left(g(\lambda)\mathbb{E}\boldsymbol{X}^2\right)$$

where

$$g(\lambda) = \frac{\lambda^2/2}{1 - \lambda K/3}.$$

Proof. First, check for $0 < \lambda < 3/K$ and $|x| \le K$, it holds

$$\mathrm{e}^{\lambda x} \le 1 + \lambda x + g(\lambda) x^2.$$

Then extend it to matrix version: for

$$0 < \lambda < 3/K \quad \text{and} \quad \|\boldsymbol{X}\|_2 \le K,$$

it holds

$$\mathrm{e}^{\lambda \boldsymbol{X}} \preceq \mathbf{I} + \lambda \boldsymbol{X} + g(\lambda) \boldsymbol{X}^2.$$

Finally, take the expectation and recall that

$$\mathbb{E}\boldsymbol{X} = \mathbf{0}$$

to obtain

$$\mathbb{E}\exp(\lambda \boldsymbol{X}) \preceq \mathbf{I} + g(\lambda)\mathbb{E}\boldsymbol{X}^2 \preceq \exp\left(\mathrm{g}(\lambda)\mathbb{E}\boldsymbol{X}^2\right),$$

where the last inequality is from the matrix version of the scalar inequality

$$1 + z \le e^z \quad \text{for all} \quad z \in \mathbb{R}. \qquad \square$$

Proof of Matrix Bernstein's inequality (Theorem 1).

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(\boldsymbol{S}) \geq t\} &= \mathbb{P}\{\exp(\lambda \cdot \lambda_{\max}(\boldsymbol{S})) \geq \exp(\lambda t)\} \\
&\leq \mathrm{e}^{-\lambda t}\mathbb{E}\mathrm{e}^{\lambda \cdot \lambda_{\max}(\boldsymbol{S})} \\
&= \mathrm{e}^{-\lambda t}\mathbb{E}\lambda_{\max}(\mathrm{e}^{\lambda \boldsymbol{S}}) \quad \text{(check!)} \\
&\leq \mathrm{e}^{-\lambda t}\mathbb{E}\mathrm{tr}(\mathrm{e}^{\lambda \boldsymbol{S}}) \\
&\leq \mathrm{e}^{-\lambda t}\mathrm{tr}\exp\left[\textstyle\sum_{i=1}^{N}\log\mathbb{E}\mathrm{e}^{\lambda \boldsymbol{X}_i}\right] \\
&\leq \mathrm{e}^{-\lambda t}\mathrm{tr}\exp\left[g(\lambda)\mathbf{Z}\right]
\end{aligned}$$

where $\mathbf{Z} := \sum_{i=1}^{N}\mathbb{E}\boldsymbol{X}_i^2$. Consider a special $\lambda = t/(\sigma^2 + Kt/3)$. With this value of λ, we conclude

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{S}) \geqslant t\right\} \leqslant n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Next, repeat the argument for $-\boldsymbol{S}$ to reinstate the absolute value. □

Corollary 3

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|\boldsymbol{X}_i\|_2 \leq K$ almost surely for all i. Then

$$\mathbb{E} \left\| \sum_{i=1}^{N} \boldsymbol{X}_i \right\|_2 \lesssim \sigma \sqrt{\log n} + K \log n$$

where $\sigma = \left\| \sum_{i=1}^{N} \mathbb{E} \boldsymbol{X}_i^2 \right\|_2^{1/2}$

Proof. The link from tail bounds to expectation is provided by the basic identity

$$\mathbb{E} Z = \int_0^\infty \mathbb{P}\{Z > t\} \, \mathrm{d}t$$

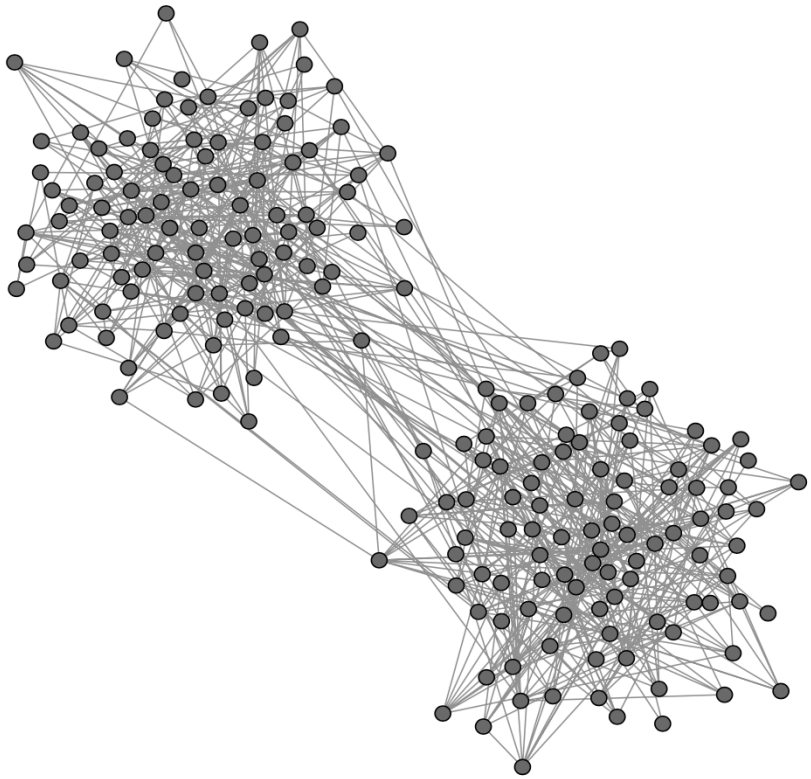
which is valid for any non-negative random variable Z. Integrating the tail bound given by matrix Bernstein's inequality, you will arrive at the expectation bound we claimed. □

3. Community recovery in networks

- A network can be mathematically represented by a graph, a set of n vertices with edges connecting some of them.
- For simplicity, we will consider undirected graphs where the edges do not have arrows.
- Real world networks often tend to have clusters, or communities - subsets of vertices that are connected by unusually many edges. (For example, a friendship network where communities form around some common interests.)
- An important problem in data science is to recover communities from a given network.
- Spectral clustering: one of the simplest methods for community recovery.

3.1 Stochastic block model

- Partition a set of n vertices into two subsets ("communities") with $n/2$ vertices each, and connect every pair of vertices independently with probability p if they belong to the same community and $q < p$ if not. The resulting random graph is said to follow the stochastic block model $G(n, p, q)$.



A network according to stochastic block model

$G(n, p, q)$

$n = 200$

$p = 1/20$

$q = 1/200$

3.2 Spectral clustering

- Suppose we are shown one instance of a random graph generated according to a stochastic block model $G(n, p, q)$. How can we find which vertices belong to which community?

- The spectral clustering algorithm we are going to explain will do precisely this. It will be based on the spectrum of the adjacency matrix $\boldsymbol{A}$ of the graph, which is the $n \times n$ symmetric matrix whose entries $\boldsymbol{A}_{ij}$ equal 1 if the vertices i and j are connected by an edge, and 0 otherwise.

- The adjacency matrix $\boldsymbol{A}$ is a random matrix. Let us compute its expectation first. This is easy, since the entires of $\boldsymbol{A}$ are Bernoulli random variables. If i and j belong to the same community then $\mathbb{E}\boldsymbol{A}_{ij} = p$ and otherwise $\mathbb{E}\boldsymbol{A}_{ij} = q$.

- Thus $\mathbb{E}\boldsymbol{A}$ has block structure: for example, if $n = 4$ then $\mathbb{E}\boldsymbol{A}$ looks like this:

$$\mathbb{E}\boldsymbol{A} = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right]$$

  For illustration purposes, we grouped the vertices from each community together.

- $\mathbb{E}\boldsymbol{A}$ has rank 2, and the non-zero eigenvalues and the corresponding eigenvectors are

$$\lambda_1 = \left(\frac{p+q}{2}\right) n, \ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}; \quad \lambda_2 = \left(\frac{p-q}{2}\right) n, \ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}.$$

- The eigenvalues and eigenvectors of $\mathbb{E}\boldsymbol{A}$ tell us a lot about the community structure of the underlying graph.

  The first (larger) eigenvalue,

$$d := \left(\frac{p+q}{2}\right) n,$$

  is the *expected degree* of any vertex of the graph.

  The second eigenvalue

$$\left(\frac{p-q}{2}\right) n,$$

  tells us whether there is any community structure at all (which happens when $p \neq q$ and thus $\lambda_2 \neq 0$).

  The signs of the components of the second eigenvector $\mathbf{v}_2$ tell us exactly how to separate the vertices into the two communities.

- The problem is that we do not know $\mathbb{E}\boldsymbol{A}$. Instead, we know the adjacency matrix $\boldsymbol{A}$. So is it true that $\boldsymbol{A} \approx \mathbb{E}\boldsymbol{A}$?

Theorem 4 (Concentration of the stochastic block model)

Let $\boldsymbol{A}$ be the adjacency matrix of a $G(n,p,q)$ random graph. Then

$$\mathbb{E}\|\boldsymbol{A} - \mathbb{E}\boldsymbol{A}\|_2 \lesssim \sqrt{d\log n} + \log n.$$

Here $d = (p+q)n/2$ is the expected degree.

Proof. To use matrix Bernstein's inequality, let us break $\boldsymbol{A}$ into a sum of independent random matrices

$$\boldsymbol{A} = \sum_{i,j:i\leqslant j} \boldsymbol{X}_{ij}$$

where each matrix $\boldsymbol{X}_{ij}$ contains a pair of symmetric entries of $\boldsymbol{A}$, or one diagonal entry. Matrix Bernstein's inequality obviously applies for the sum

$$\boldsymbol{A} - \mathbb{E}\boldsymbol{A} = \sum_{i\leqslant j} (\boldsymbol{X}_{ij} - \mathbb{E}\boldsymbol{X}_{ij})\,.$$

Corollary 3 gives

$$\mathbb{E}\|\boldsymbol{A} - \mathbb{E}\boldsymbol{A}\|_2 \lesssim \sigma\sqrt{\log n} + K\log n$$

where

$$\sigma^2 = \left\|\sum_{i\leqslant j}\mathbb{E}\left(\boldsymbol{X}_{ij} - \mathbb{E}\boldsymbol{X}_{ij}\right)^2\right\|_2$$

and

$$K = \max_{i,j}\left\|\boldsymbol{X}_{i,j} - \mathbb{E}\boldsymbol{X}_{i,j}\right\|_2.$$

It is a good exercise to check that

$$\sigma^2 \lesssim d$$

and

$$K \leqslant 2.$$

This completes the proof. □

- How useful is Theorem 4 for community recovery? Suppose that the network is not too sparse, namely

$$d \gg \log n.$$

Then (with high probability)

$$\|\boldsymbol{A} - \mathbb{E}\boldsymbol{A}\|_2 \lesssim \sqrt{d \log n}$$

while

$$\|\mathbb{E}\boldsymbol{A}\|_2 = \lambda_1(\mathbb{E}\boldsymbol{A}) = d,$$

which implies that

$$\|\boldsymbol{A} - \mathbb{E}\boldsymbol{A}\|_2 \ll \|\mathbb{E}\boldsymbol{A}\|_2.$$

In other words, $\boldsymbol{A}$ nicely approximates $\mathbb{E}\boldsymbol{A}$: the relative error or approximation is small in the operator norm.

- Classical results from the perturbation theory for matrices state that if $\boldsymbol{A}$ and $\mathbb{E}\boldsymbol{A}$ are close, then their eigenvalues and eigenvectors must also be close. The relevant perturbation results are Weyl's inequality for eigenvalues and Davis–Kahan's inequality for eigenvectors. (See the bible NLA book MC)

- So we can use $\mathbf{v}_2(\boldsymbol{A})$ to approximately recover the communities. This method is called spectral clustering:

  Spectral Clustering Algorithm.

  Compute $\mathbf{v}_2(\boldsymbol{A})$, the eigenvector corresponding to the second largest eigenvalue of the adjacency matrix $\boldsymbol{A}$ of the network. Use the signs of the coefficients of $\mathbf{v}_2(\boldsymbol{A})$ to predict the community membership of the vertices.