IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: P Rigollet
Scribe: Abhimanyu Dubey, Zachary Markovich

Lecture 10
Mar. 10, 2020

Goals: In the previous lecture, we introduced the Davis-Kahan $\sin(\theta)$ theorem, a result in perturbation theory that allows us to provide a bound on the 2-norm of the difference between the top eigenvector of a matrix and its perturbed version. In this lecture we will first provide a proof for the Davis-Kahan $\sin(\theta)$ theorem and subsequently we will provide two of its applications in matrix denoising and community detection.

1. DAVIS-KAHAN $\sin(\theta)$ THEOREM

The Davis-Kahan $\sin(\theta)$ theorem provides an important result in perturbation theory of matrices. We begin by restating the theorem.

Theorem: Let $A, \hat{A}$ be two symmetric $n \times n$ matrices, with eigen-decompositions given by:

$$A = \sum_{j=1}^{n} \lambda_j u_j u_j^\top \text{ and } \hat{A} = \sum_{j=1}^{n} \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top,$$

where, $\lambda_1 \geq \lambda_2 ... \geq \lambda_n$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 ... \geq \hat{\lambda}_n$ without loss of generality. Then,

$$|\sin(\measuredangle(u_1, \hat{u}_1))| \leq 2 \frac{\|A - \hat{A}\|_{\text{op}}}{\max\left(\lambda_1 - \lambda_2, \hat{\lambda}_1 - \hat{\lambda}_2\right)}.$$

Moreover,

$$\min_{\varepsilon \in \{\pm 1\}} |u_1 - \varepsilon \hat{u}_1|_2 \leq \sqrt{2} \, |\sin(\measuredangle(u_1, \hat{u}_1))| \, .$$

Proof. The proof proceeds by considering the eigendecomposition of A. For any $x \in \mathbb{R}^n$, such that $|x|_2 = 1, \sum_{j=1}^n \left(x^\top u_j\right)^2 = 1$. Therefore, since $\lambda_1 \geq \lambda_2 \geq ... \lambda_n$,

$$\begin{aligned}
x^\top A x = \sum_{j=1}^{n} \lambda_j (x^\top u_j)^2 &= \lambda_1 (x^\top u_1)^2 + \sum_{j \geq 1}^{n} \lambda_j (x^\top u_j)^2 \\
&\leq \lambda_1 (x^\top u_1)^2 + \lambda_2 \sum_{j \geq 2} (x^\top u_j)^2 && (\lambda_j \leq \lambda_2 \; \forall \, j \geq 2) \\
&= (\lambda_1 - \lambda_2)(x^\top u_1)^2 + \lambda_2 && \left(\textstyle\sum_{j=1}^n \left(x^\top u_j\right)^2 = 1\right) \\
&= \lambda_1 - (\lambda_1 - \lambda_2) \sin^2(\measuredangle(x, u_1)) \, . && (\cos(\measuredangle(x, u_1)) = x^\top u_1) \\
&= u_1^\top A u_1 - (\lambda_1 - \lambda_2) \sin^2(\measuredangle(x, u_1)) \, . && (\lambda_1 = u_1^\top A u_1)
\end{aligned}$$

Now, by taking $x = \hat{u}_1$,

$$\begin{aligned}
(\lambda_1 - \lambda_2) \sin^2 \left(\measuredangle(\hat{u}_1, u_1)\right) &\leq u_1^\top A u_1 - \hat{u}_1^\top A \hat{u}_1 \\
&= u_1^\top \hat{A} u_1 - \hat{u}_1^\top A \hat{u}_1 + u_1^\top (A - \hat{A}) u_1. \\
&\leq \hat{u}_1^\top \hat{A} \hat{u}_1 - \hat{u}_1^\top A \hat{u}_1 + u_1^\top (A - \hat{A}) u_1. \quad (x^\top A x \leq \hat{u}_1^\top \hat{A} \hat{u}_1 \ \forall |x|_2 = 1) \\
&= \langle \hat{A} - A, \hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top \rangle. \\
&\leq \|\hat{A} - A\|_{\text{op}} \|\hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top\|_1 && \text{(Hölder)} \\
&\leq \sqrt{2} \|\hat{A} - A\|_{\text{op}} \|\hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top\|_F && \text{(Cauchy-Schwarz)}
\end{aligned}$$

We can simplify the second term as follows.

$$\begin{aligned}
\|\hat{u}_1 \hat{u}_1^\top - u_1 u_t^\top\|_F^2 &= \underbrace{\text{Tr}\left(u_1 u_1^\top u_1 u_1^\top\right)}_{=1} + \underbrace{\text{Tr}\left(\hat{u}_1 \hat{u}_1^\top \hat{u}_1 \hat{u}_1^\top\right)}_{=1} - 2\,\text{Tr}\left(\hat{u}_1 \hat{u}_1^\top u_1 u_1^\top\right) \\
&= 2 - 2\left(\hat{u}_1^\top u_1\right)^2 \\
&= 2 \sin^2(\measuredangle(\hat{u}_1, u_1)). && (\cos(\measuredangle(\hat{u}_1, u_1)) = \hat{u}_1^\top u_1)
\end{aligned}$$

Replacing this in the original statement gives us:

$$(\lambda_1 - \lambda_2) \sin^2 \left(\measuredangle(\hat{u}_1, u_1)\right) \leq 2 \|A - \hat{A}\|_{\text{op}} \left|\sin(\measuredangle(\hat{u}_1, u_1))\right|.$$

An identical analysis can be done by considering the eigendecomposition of $\hat{A}$ (and replacing $x = u_1$ subsequently), which will provide:

$$(\hat{\lambda}_1 - \hat{\lambda}_2) \sin^2 \left(\measuredangle(\hat{u}_1, u_1)\right) \leq 2 \|A - \hat{A}\|_{\text{op}} \left|\sin(\measuredangle(\hat{u}_1, u_1))\right|.$$

Combining the two results,

$$\left|\sin\left(\measuredangle\left(u_1, \hat{u}_1\right)\right)\right| \leq 2 \frac{\|A - \hat{A}\|_{\text{op}}}{\max\left(\lambda_1 - \lambda_2, \hat{\lambda}_1 - \hat{\lambda}_2\right)}.$$

Typically, we would not like to use the eigenvalues of the perturbed matrix $\hat{A}$, since it is a random quantity in our applications and non-trivial to control. However, domain-specific assumptions can be made about the spectrum of A, leading to useful results. Now, for the second part of the proof.

$$\min_{\varepsilon \in \{\pm 1\}} |u_1 - \varepsilon \hat{u}_1|_2^2 = 2 - 2|u_1^\top \hat{u}_1| \overset{(a)}{\leq} 2 - 2(u_1^\top \hat{u}_1)^2 \overset{(b)}{\leq} 2 \sin^2(\measuredangle(u_1, \hat{u}_1)).$$

Here, (a) follows from the fact that $x \leq x^2 \ \forall x \leq 1$, and (b) follows from the relationship of the cosine and dot product. □

2. APPLICATIONS

2.1 Matrix Denoising

We will consider a sub-Gaussian matrix denoising problem similar to the previous lecture.

$$Y = \Theta^* + E$$

We will assume that all of $Y, \Theta^* \in \mathbb{R}^{n\times n}$ and $E = (e_{ij})_{i\geq j}^n \sim \mathsf{subG}(\sigma^2)$ are symmetric for simplicity. We can represent the matrices by their SVD:

$$\Theta^* = \sum \lambda_j u_j u_j^\top \quad \text{and} \quad Y = \sum \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top.$$

From the Davis-Kahan $\sin(\theta)$ theorem, we have:

$$\min_{\varepsilon\in\{\pm 1\}} |\hat{u}_1 - \varepsilon u_1|_2^2 \leq 2\frac{\|E\|_{\text{op}}}{\lambda_1 - \lambda_2}$$

We have construction, $u^\top E v \sim \mathsf{subG}(\sigma^2 |u|_2^2 |v|_2^2)$. We can therefore conclude that there exists a constant such that with probability at least 0.99,

$$\min_{\varepsilon\in\{\pm 1\}} |\hat{u}_1 - \varepsilon u_1|_2^2 \leq 2\frac{\|E\|_{\text{op}}}{\lambda_1 - \lambda_2} \lesssim \frac{\sigma\sqrt{n}}{\lambda_1 - \lambda_2}.$$

We see that the crucial quantity controlling the quality of approximation is the signal to noise ratio (SNR) given by $\frac{\lambda_1-\lambda_2}{\sigma}$.

2.2 Community Detection

Community detection is an important problem in social network analysis. We consider the stochastic block model in this example. The social network is determined by an undirected graph represented by an adjacency matrix $\tilde{A}$, where the n nodes represent people and an edge (i,j) denotes a friendship between persons i and j.

We assume persons belong to one of two groups. If two persons i,j belong to the same group then there is an edge (i,j) in $\tilde{A}$ with probability p. Alternatively, if they belong to different groups, there is an edge (i,j) with probability q. Therefore, $\tilde{A}_{ij} \sim \mathsf{Ber}(p)$ if i and j belong to the same community, and $\tilde{A}_{ij} \sim \mathsf{Ber}(q)$ otherwise. Self-edges, i.e., A_{ii}, can be modeled based on the application. In this setting, we will assume they are random variables, i.e., $A_{ii} \sim \mathsf{Ber}(p)$.

We assume that p and q are known *a priori*, however, relaxations can be made in case they are not. Each edge is assumed independent from the others (i.e., no network effects). The goal of the problem is to recover the community structure, given a realization $\tilde{A}$ of the network. From our formulation, we have:

$$\mathbb{E}[\tilde{A}_{ij}] = \begin{cases} p \text{ if } i \text{ and } j \text{ belong to the same community,} \\ q \text{ otherwise.} \end{cases} \tag{2.1}$$

To represent the matrices, assume that the first group is the first $n/2$ nodes. Then, we can write $\mathbb{E}[\tilde{A}]$ as:

$$\mathbb{E}[\tilde{A}] = \left[\begin{array}{ccc|ccc} p & \cdots & p & q & \cdots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \cdots & p & q & \cdots & q \\ \hline q & \cdots & q & p & \cdots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \cdots & q & p & \cdots & p \end{array}\right].$$

Let $E = \tilde{A} - \mathbb{E}[\tilde{A}]$. Then,

$$\tilde{A} = \mathbb{E}[\tilde{A}] + E.$$

We see that $\mathbb{E}[E_{ij}] = 0$ and $|E_{ij}| \leq 1 \; \forall i, j$. From Hoeffding's lemma, for vectors u and v,

$$u^\top E v \sim \mathsf{subG}(\|u\|_2^2 \cdot \|v\|_2^2).$$

For the community detection problem, we will analyse the centered matrix A:

$$A = \tilde{A} - \left(\frac{p-q}{2}\right) \mathbb{1}_n \mathbb{1}_n^\top.$$

Similar to $\mathbb{E}[\tilde{A}]$, we see that $\mathbb{E}[A]$ can be written as:

$$\mathbb{E}[A] = \left[\begin{array}{ccc|ccc} \frac{p-q}{2} & \dots & \frac{p-q}{2} & \frac{q-p}{2} & \dots & \frac{q-p}{2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{p-q}{2} & \dots & \frac{p-q}{2} & \frac{q-p}{2} & \dots & \frac{q-p}{2} \\ \hline \frac{q-p}{2} & \dots & \frac{q-p}{2} & \frac{p-q}{2} & \dots & \frac{p-q}{2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{q-p}{2} & \dots & \frac{q-p}{2} & \frac{p-q}{2} & \dots & \frac{p-q}{2} \end{array}\right] = \frac{p-q}{2} \left[\begin{array}{ccc|ccc} 1 & \dots & 1 & -1 & \dots & -1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & 1 & -1 & \dots & -1 \\ \hline -1 & \dots & -1 & 1 & \dots & 1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -1 & \dots & -1 & 1 & \dots & 1 \end{array}\right].$$

We see that $\mathbb{E}[A]$ is of rank 1 and it can be written as:

$$\mathbb{E}[A] = \frac{n(p-q)}{2} \left(\frac{u_1}{\sqrt{n}}\right) \left(\frac{u_1}{\sqrt{n}}\right)^\top, \text{ where } u_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}.$$

To relax the ordered-ness assumption of A, we can consider that we are provided with an alternate matrix $\Pi A \Pi^\top$, where Π is an $n \times n$ permutation of A. In that case, we can represent the resulting permutation of $\mathbb{E}[A]$ as follows:

$$\Pi \mathbb{E}[A] \Pi^\top = \frac{n(p-q)}{2} \left(\frac{\Pi u_1}{\sqrt{n}}\right) \left(\frac{\Pi u_1}{\sqrt{n}}\right)^\top.$$

To estimate u from the (centered) observed matrix, we consider the eigendecomposition of A, i.e.,

$$A = \tilde{A} - \left(\frac{p-q}{2}\right) \mathbf{1}_n \mathbf{1}_n^\top = \sum_{j=1}^n \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top.$$

From the Davis-Kahan $\sin(\theta)$ theorem, we can bound the difference in the first eigenvectors, in a fashion similar to the last example:

$$\min_{\varepsilon \in \{\pm 1\}} \left| \frac{\hat{u}_1}{\sqrt{n}} - \frac{\varepsilon u_1}{\sqrt{n}} \right|_2^2 \overset{(a)}{\leq} 2 \frac{\|A - \mathbb{E}[A]\|_{\text{op}}}{\lambda_1 - 0} \overset{(b)}{=} 4 \frac{\|E\|_{\text{op}}}{n(p-q)} \overset{(c)}{\lesssim} \frac{1}{(p-q)\sqrt{n}}.$$

Here, (a) follows from Davis-Kahan, (b) follows from the fact that $A-\mathbb{E}[A] = \tilde{A}-\mathbb{E}[\tilde{A}] = E$ and that $\lambda_1 = \frac{n(p-q)}{2}$ and (c) holds with probability at least 0.99, obtained from the bound on the operator norm for sub-Gaussian random matrices. Therefore, we can say that if $(p-q) \gg \frac{1}{\sqrt{n}}$,

$$\min_{\varepsilon\in\{\pm 1\}} \left| \frac{\hat{u}_1}{\sqrt{n}} - \frac{\varepsilon u_1}{\sqrt{n}} \right|_2^2 \ll 1.$$

A popular measure of performance is the *classification error*, i.e., the average number of times our prediction disagrees with the true assignment. This can be given by:

$$L(u_1, \hat{u}_1) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left\{\text{sign}(\hat{u}_1^{(j)}) \neq \text{sign}(u_1^{(j)})\right\}.$$

$L(u_1, \hat{u}_1)$ is non-convex. However, we can bound it by a convex relaxation that we can provide guarantees for, as follows.

$$\begin{aligned} L(u_1, \hat{u}_1) &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left\{\text{sign}(\hat{u}_1^{(j)}) \neq \text{sign}(u_1^{(j)})\right\} \\ &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left\{\hat{u}_1^{(j)} \cdot u_1^{(j)} < 0\right\} \\ &\leq \frac{1}{n}\sum_{j=1}^{n} \left(\hat{u}_1^{(j)} \cdot u_1^{(j)} - 1\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} \left(\hat{u}_1^{(j)} - u_1^{(j)}\right)^2 \\ &= \left\| \frac{\hat{u}_1}{\sqrt{n}} - \frac{u_1}{\sqrt{n}} \right\|_2^2. \end{aligned}$$

From the previous result, we can claim that $L(\hat{u}_1, u_1) \lesssim \frac{1}{(p-q)\sqrt{n}}$. Therefore, we can additionally conclude that as $\sqrt{n}(p-q) \to \infty$, the probability of error, $L(u_1, \hat{u}_1) \to 0$.

Additional Remarks. Stronger statements can be made regarding the recovery problem than the ones described in this lecture. [ABH16] prove lower bounds for *exact* recovery. Specifically, they state that, for constants α and β such that

$$p = \frac{\alpha \log(n)}{n} \text{ and } q = \frac{\beta \log(n)}{n},$$

they demonstrate that exact recovery of u_1 as $n \to \infty$ is possible when $\alpha + \beta - 2\sqrt{\alpha\beta} > 2$, and impossible otherwise. Exact recovery of u_1 implies that there exists an estimator $\hat{u}_1$ such that $\forall j$ simultaneously,

$$u_1^{(j)} \cdot \hat{u}_1^{(j)} > 0, \text{ w.p. } \to 1 \text{ as } n \to \infty.$$

Additionally, the authors provide an efficient algorithm for exact recovery. They consider the following optimization problem:

$$\max_{\|u\|_2=1, u^{(j)}\in\{\pm\frac{1}{\sqrt{n}}\}, \sum_j u^{(j)}=0} u^\top \left(A - \frac{p+q}{2}\mathbf{1}_n\mathbf{1}_n^\top\right) u.$$

Since this problem is NP-Hard, the authors solve an SDP relaxation, the solution of which is demonstrated to also be close to exact.

Alternately, if graph structure is very sparse (i.e., many isolated pairs) then subgroup recovery will be impossible. [MNS13] and [Mas14] provide a series of results for the sparse case. When, for constants a and b such that $p = \frac{a}{n}$ and $q = \frac{b}{n}$, recovery is impossible in general, but it is possible to do better than chance in certain cases. Specifically, when $(a-b)^2 \geq 2(a+b)$, there exists an estimator $\hat{u}_1$ and constant $\alpha > 0$ such that,

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left\{\hat{u}_1^{(j)} - u_1^{(j)}\right\} \to \frac{1}{2} - \alpha, \text{ as } n \to \infty.$$

Even using the techniques in our analysis, tighter bounds can be obtained by carefully analysing the concentration of the noise matrix. In our analysis, we assumed that the noise is $\mathsf{subG}(1)$, which can be tightened by considering a careful Bernstein concentration.

Summary: In this lecture, we first provided a proof for Davis-Kahan $\sin(\theta)$ theorem, that provides the following bound on the perturbations of top eigenvectors:

$$\min_{\varepsilon \in \{\pm 1\}} |u_1 - \varepsilon \hat{u}_1|_2 \leq \sqrt{2}\left|\sin\left(\measuredangle\left(u_1, \hat{u}_1\right)\right)\right| \leq 2\sqrt{2}\frac{\|A - \hat{A}\|_{\text{op}}}{\max\left(\lambda_1 - \lambda_2, \hat{\lambda}_1 - \hat{\lambda}_2\right)}.$$

Next, for the matrix denoising problem, and demonstrate that with high probability,

$$\min_{\varepsilon \in \{\pm 1\}} |\hat{u}_1 - \varepsilon u_1|_2^2 \leq 2\frac{\|E\|_{\text{op}}}{\lambda_1 - \lambda_2} \lesssim \frac{\sigma\sqrt{n}}{\lambda_1 - \lambda_2}.$$

Finally, we considered the community detection problem (stochastic block model), for which we bounded the misclassification error as follows.

$$L(u_1, \hat{u}_1) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left\{\text{sign}(\hat{u}_1^{(j)}) \neq \text{sign}(u_1^{(j)})\right\} \lesssim \frac{1}{(p-q)\sqrt{n}}.$$

We concluded the lecture with a short discussion on state-of-the-art results in the community detection problem.

References

[MNS13] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, Jan 2016.

[Mas14] Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.