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

Lecturer: P Rigollet
Scribe: Abraham Shalom, Ziad Mansour

Lecture 9
Mar. 5, 2020

Goals: In the last lecture, we introduced the fundamental matrix estimation model and some basic matrix properties and inequalities. In this lecture, we introduce the singular value thresholding estimator for matrix denoising. We then show that, with a good threshold choice, this estimator's MSE will decay as the matrix dimensions grow. Finally, we introduce the Davis-Kahan-Sin(Theta) Theorem for use in matrix perturbation analysis.

1. MATRIX DENOISING

We consider the subGaussian matrix model:

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

where $Y \in \mathbb{R}^{m\times n}$ is the matrix of observed responses, and E is an $m \times n$ noise matrix such that

$$u^\top E v \sim \mathsf{subG}(\sigma^2)\,, \quad \forall u \in \mathbb{R}^m, v \in \mathbb{R}^n\,.$$

This is the case for example if $\forall\,, i, j \in [m] \times [n]$, each of entry of E denoted by $E_{i,j}$ is an independent $\mathsf{subG}(\sigma^2)$. Indeed for any $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$, it holds

$$\mathbb{E}[e^{su^\top Ev}] = \mathbb{E}[e^{\sum_{ij} u_i E_{ij} v_j}] = \prod_{i,j} \mathbb{E}[e^{su_i E_{ij} v_j}] \le \prod_{i,j} e^{\frac{s^2\sigma^2 u_i^2 v_j^2}{2}} = e^{\sum_{i,j}\frac{s^2\sigma^2 u_i^2 v_j^2}{2}} = e^{\frac{s^2\sigma^2 |u|_2^2 |v|_2^2}{2}}$$

Similar to the sub-Gaussian sequence model, we have a direct observation model where we observe the parameter of interest with additive noise. This enables us to use thresholding methods for estimating Θ^*. The analysis becomes interesting when we assume that Θ^* is low rank which is equivalent to sparsity in its unknown eigenbasis. Hence, we can consider the SVD of Θ^* and of Y:

$$\Theta^* = \sum_{j=1}^{m\wedge n} \lambda_j u_j v_j^\top$$

$$Y = \sum_{j=1}^{m\wedge n} \hat\lambda_j \hat u_j \hat v_j^\top$$

Hence, if we knew the u_j and v_j, we can estimate the λ_j's by hard thresholding. We claim that it is sufficient to estimate the eigenvectors of Θ^* by the eigenvectors of Y.

We define the Singular Value Thresholding Estimator, SVT, with threshold $2\tau \ge 0$ as:

$$\hat\Theta^{\text{SVT}} = \sum_j \hat\lambda_j \mathbb{1}(\hat\lambda_j > 2\tau) \hat u_j \hat v_j^\top$$

How to select τ? Recall in the Gaussian sequence model, we select τ so that it is larger than the maximum magnitude of the noise with probability $1-\delta$. We take a similar approach here except that the norm in which the magnitude of the noise is measured is adapted to the matrix case. Basically, we are trying to control the operator norm of the matrix E.

$$\text{Our Candidate : } \quad \tau \geq \max_j \lambda_j(E) = \|E\|_{\text{op}}$$

Recall from Lecture 5, the result on operator norms of $m \times n$ matrices: If $E_{i,j} \sim \mathsf{subG}(\sigma^2)$ are independent, then $\|E\|_{\text{op}} \leq \sigma(\sqrt{m}+\sqrt{n})$ with high probability. We generalize this result to the following lemma which will allow us to control the operator norm of the matrix E.

Lemma: Let E be an $m \times n$ random matrix defined as above. Then

$$\|E\|_{\text{op}} \leq 2\sigma\sqrt{5(m+n)} + 2\sigma\sqrt{2\log(1/\delta)}$$

with probability $1-\delta$.

Proof. Let $\mathcal{N}_m$ be a 1/4-net for the euclidean ball $\mathcal{B}_2(R^m)$ and $\mathcal{N}_n$ be a 1/4-net for the euclidean ball $\mathcal{B}_2(R^n)$. Then it follows from the results of lecture 5 that for $\varepsilon = 1/4$:

$$|\mathcal{N}_m| \leq (1+\frac{2}{\varepsilon})^m = 9^m$$

$$|\mathcal{N}_n| \leq (1+\frac{2}{\varepsilon})^n = 9^n$$

Now we can write the operator norm on E as:

$$\|E\|_{\text{op}} = \max_{x \in \mathcal{S}^{n-1}} |Ex|_2 = \max_{\substack{x \in \mathcal{S}^{n-1} \\ y \in \mathcal{S}^{m-1}}} y^\top Ex$$

On the operator norm, we use the triangular inequality and decompose each point x on the sphere $\mathcal{S}^{n-1}$ into a point z on the epsilon net $\mathcal{N}_n$ and a remainder term that can be upper bounded by 1/4. By applying the described decomposition on the maximum $x \in \mathcal{S}^{n-1}$ we can upper bound the operator norm of E by :

$$\|E\|_{\text{op}} \leq \max_{z \in \mathcal{N}_n} |Ez|_2 + \frac{1}{4}\|E\|_{\text{op}}$$

Now we examine the $|Ez|_2$ term:

$$\begin{aligned} |Ez|_2 &= \max_{y \in \mathcal{S}^{m-1}} y^\top Ez \\ &\leq \max_{w \in \mathcal{N}_m} w^\top Ez + \frac{1}{4}\|E\|_{\text{op}} \end{aligned}$$

Hence applying this bound to our $\|E\|_{\text{op}}$

$$\|E\|_{\text{op}} \leq \max_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} w^\top Ez + \frac{1}{4}\|E\|_{\text{op}} + \frac{1}{4}\|E\|_{\text{op}}$$

Rearranging the above inequality

$$\|E\|_{\text{op}} \leq 2 \max_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} w^\top E z$$

where $\forall w, z;\ w^\top E z \sim \mathsf{subG}(\sigma^2 |w|_2^2 |z|_2^2)$. Using the fact that $|w|_2^2 \leq 1$ and $|z|_2^2 \leq 1$, then $w^\top E z \sim \mathsf{subG}(\sigma^2)$. Using union bounds:

$$\mathbb{P}(\max_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} w^\top E z > \frac{t}{2}) \leq \sum_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} \mathbb{P}(w^\top E z > \frac{t}{2}) \leq \sum_{\substack{z \in \mathcal{N}_n \\ w \in \mathcal{N}_m}} e^{\frac{-t^2}{8\sigma^2}} \leq 9^{n+m} e^{\frac{-t^2}{8\sigma^2}} =: \delta$$

Solving for t we get

$$t \leq 2\sqrt{2}\sigma\sqrt{\log(9)(m+n)} + 2\sqrt{2}\sigma\sqrt{-\log(\delta)} \leq 2\sigma\sqrt{5(n+m)} + 2\sigma\sqrt{-2\log(\delta)}$$

Thus, we choose $\tau = 2\sigma\sqrt{5(n+m)} + 2\sigma\sqrt{-2\log(\delta)}$. □

Using the above lemma, we propose the following theorem:

Theorem: The SVT estimator $\hat{\Theta}^{\text{SVT}}$ with τ as above, satisfies (with probability $1-\delta$):

$$\frac{1}{m \cdot n} \left\| \hat{\Theta}^{\text{SVT}} - \Theta^* \right\|_F^2 \lesssim \frac{\sigma^2 \cdot \text{rank}(\Theta^*)}{m \cdot n}(m + n + \log\frac{1}{\delta})$$

Proof. To prove this, we first define a random set of indices S:

$$S = \{j \ : \ |\hat{\lambda}_j| > 2\tau\}$$

We then define the event $\mathcal{A} = \{\|E\|_{\text{op}} \leq \tau\}$ and recall that $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$. Now, we will make purely deterministic statements assuming that $\mathcal{A}$ holds.

By Weyl's inequality,

$$|\hat{\lambda}_j - \lambda_j| \leq \|Y - \Theta^*\|_{\text{op}} = \|E\|_{\text{op}} \leq \tau$$

Based off of the above in conjunction with triangle inequality, we can make statements about $j \in S$ and $j \in S^c$ as follows:

$$\begin{aligned} j \in S: &\quad |\hat{\lambda}_j| > 2\tau \implies |\hat{\lambda}_j| \geq |\hat{\lambda}_j| - |\hat{\lambda}_j - \lambda_j| > \tau \\ j \in S^c: &\quad |\hat{\lambda}_j| \leq 2\tau \implies |\lambda_j| \leq |\hat{\lambda}_j| + |\lambda_j - \hat{\lambda}_j| \leq 3\tau \end{aligned}$$

We introduce an oracle $\bar{\Theta}$ that knows the singular values and vectors, but not the support.

$$\bar{\Theta} = \sum_{j \in S} \lambda_j u_j v_j^\top$$

Next,

$$\left\| \hat{\Theta}^{\text{SVT}} - \bar{\Theta} \right\|_F^2 = \sum_j \lambda_j^2(\hat{\Theta}^{\text{SVT}} - \bar{\Theta}) \leq \left\| \hat{\Theta}^{\text{SVT}} - \bar{\Theta} \right\|_{\text{op}}^2 \text{rank}(\hat{\Theta}^{\text{SVT}} - \bar{\Theta}),$$

where we used the fact that $\lambda_j^2(\hat{\Theta}^{\text{SVT}} - \bar{\Theta}) \leq \left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_{\text{op}}^2$ for all j.

Note that the rank of the sum of two matrices is at most the sum of the two ranks. Since $\text{rank}(\hat{\Theta}^{\text{SVT}}) \vee \text{rank}(\bar{\Theta}) \leq |S|$, then $\text{rank}(\hat{\Theta}^{\text{SVT}} - \bar{\Theta}) \leq 2|S|$.

Now, we need to bound the operator norm in the above inequality. We first apply the triangle inequality, bringing in Θ^* and Y:

$$\left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_{\text{op}} \leq \left\|\hat{\Theta}^{\text{SVT}} - Y\right\|_{\text{op}} + \|Y - \Theta^*\|_{\text{op}} + \left\|\Theta^* - \bar{\Theta}\right\|_{\text{op}}$$

Now, we bound each of the individual terms that appear above. First, we have that, by construction,

$$\|Y - \Theta^*\|_{\text{op}} = \|E\|_{\text{op}} \leq \tau$$

We now work to bound $\left\|\hat{\Theta}^{\text{SVT}} - Y\right\|_{\text{op}}$, using the fact that $\forall j \in S^c: \ \hat{\lambda}_j \leq 2\tau$.

$$Y - \hat{\Theta}^{\text{SVT}} = \sum_{j \in S^c} \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top$$
$$\left\|\hat{\Theta}^{\text{SVT}} - Y\right\|_{\text{op}} = \max_{j \in S^c} \hat{\lambda}_j \leq 2\tau$$

We use similar process to bound $\left\|\Theta^* - \bar{\Theta}\right\|_{\text{op}}$, using $\forall j \in S^c: \ \lambda_j \leq 3\tau$.

$$\Theta^* - \bar{\Theta} = \sum_{j \in S^c} \lambda_j u_j v_j^\top$$
$$\left\|\Theta^* - \bar{\Theta}\right\|_{\text{op}} = \max_{j \in S^c} \lambda_j \leq 3\tau$$

Putting the above together, we get

$$\left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_{\text{op}} \leq 2\tau + \tau + 3\tau = 6\tau$$
$$\left\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\right\|_F^2 \leq 72\tau^2|S| = 72\sum_{j \in S} \tau^2$$

Now, we note that we are not interested in $\bar{\Theta}$ (from the oracle); instead, we want to compare our estimate to Θ^*. We therefore use the triangle inequality (again) to squeeze in the $\bar{\Theta}$.

$$\begin{aligned}\left\|\hat{\Theta}^{\text{SVT}} - \Theta^*\right\|_F^2 &\leq 2\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\|_F^2 + 2\left\|\bar{\Theta} - \Theta^*\right\|_F^2 \\ &\leq 144\sum_{j \in S} \tau^2 + 2\sum_{j \in S^c} \lambda_j^2\end{aligned}$$

Notice that the term in the left-hand sum, τ^2, is in fact equal to $\min(\tau^2, \lambda_j^2)$. In the right-hand sum, the term $\lambda_j^2 \leq \min(9\lambda_j^2, 9\tau^2)$. Thus, we add this to the above inequality:

$$\begin{aligned}
\left\|\hat{\Theta}^{\text{SVT}} - \Theta^*\right\|_F^2 &\leq 144 \sum_{j \in S} (\tau^2 \wedge \lambda_j^2) + 18 \sum_{j \in S^c} (\tau^2 \wedge \lambda_j^2) \\
&\leq 144 \sum_{j \in S} (\tau^2 \wedge \lambda_j^2) + 144 \sum_{j \in S^c} (\tau^2 \wedge \lambda_j^2) \\
&= 144 \sum_{j} (\tau^2 \wedge \lambda_j^2) \\
&\lesssim \tau^2 \text{rank}(\Theta^*) \\
&\lesssim (\sigma^2(m+n) + \sigma^2 \log \frac{1}{\delta}) \text{rank}(\Theta^*)
\end{aligned}$$

To finish the proof, we can simply multiply both sides of our inequality by $\frac{1}{m \cdot n}$ in order to get mean-squared error. This gives us a final bound of:

$$\frac{\left\|\hat{\Theta}^{\text{SVT}} - \Theta^*\right\|_F^2}{m \cdot n} \lesssim \frac{(\sigma^2(m+n) + \sigma^2 \log \frac{1}{\delta})}{m \cdot n} \text{rank}(\Theta^*)$$

□

Using this, we see that without knowing anything about $\hat{u}$, $\hat{v}$ and their closeness to u, v respectively we can get an MSE that is vanishing as $m, n \to \infty$. The question now becomes: can we get greedy and find any results indicating closeness of our estimates to the true singular vectors?

2. PERTURBATION THEORY

We give a brief introduction to Perturbation Theory, with the subject to be continued in the next lecture. The focus of perturbation theory is to understand how a matrix's spectrum changes if its entries are perturbed. We introduce an important result in perturbation theory.

Theorem (Davis-Kahan $\sin(\theta)$ Theorem): Let A, $\hat{A}$ by symmetric $n \times n$ matrices such that:

$$\begin{aligned}
A &= \sum_{j=1}^{n} \lambda_j u_j u_j^\top, \quad \lambda_1 \geq \lambda_2 \geq \ldots \\
\hat{A} &= \sum_{j=1}^{n} \hat{\lambda}_j \hat{u}_j \hat{u}_j^\top, \quad \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \\
|u_j|_2 &= |\hat{u}_j|_2 = 1
\end{aligned}$$

Then:

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

Moreover:

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

The proof to this Theorem will be given in the following lecture.

Summary: In this lecture, we develop tools for matrix denoising. We defined the Singular Value Thresholding Estimator (SVT) with threshold 2τ as

$$\hat{\Theta}^{\text{SVT}} = \sum_j \hat{\lambda}_j \mathbb{1}(\hat{\lambda}_j > 2\tau) \hat{u}_j \hat{v}_j^\top$$

We then established the following lemma that helps us control the operator norm of the noise matrix:
Let E be an $m \times n$ random matrix such that $u^\top E v \sim \mathsf{subG}(\sigma^2)$ for all $u \in \mathbb{R}^m, v \in \mathbb{R}^n$, then

$$\|E\|_{\text{op}} \lesssim \sigma\sqrt{m+n} + \sigma\sqrt{\log(1/\delta)}$$

with probability $1 - \delta$.

This, in turn, allows us to bound the MSE of the SVT estimator using the following theorem, and show that the average error goes to 0 as $m, n \to \infty$ (for $\text{rank}(\Theta^*) \ll m, n$).

The SVT estimator $\hat{\Theta}^{\text{SVT}}$ with τ as above, satisfies (with probability $1 - \delta$):

$$\frac{1}{m \cdot n} \left\| \hat{\Theta}^{\text{SVT}} - \Theta^* \right\|_F^2 \lesssim \frac{\sigma^2 \cdot \text{rank}(\Theta^*)}{m \cdot n} (m + n + \log \frac{1}{\delta})$$

Finally, we introduce the Davis-Kahan $\sin(\theta)$ theorem, which is an important result in determining the closeness of eigenspaces. This serves as the beginning of our studies into perturbation theory, which we will continue in the next lecture.