ORIE 7790 High Dimensional Probability and Statistics Lecture 7 - 02/11/2020

Lecture 7: Random Matrices I

Lecturer: Yudong Chen *Scribe: Xumei Xi*

References:

- M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Sections 5.4, 6.2.
- R. Vershynin, *High dimensional Probability*, Sections 7.2, 7.3.

1 Motivation

Consider the following matrix estimation problem. Let $Y^* \in \mathbb{R}^{n\times n}$ be an unknown low-rank matrix. Y is a noisy version of Y^*, with $\mathbb{E}[Y] = Y^*$. Our task is to produce an estimator $\hat{Y}$ by leveraging the low-rank structure of Y^*. To study the estimation error, we often need to control the quantity $\|Y - Y^*\|_{\text{op}}$. The question reduces to upper bounding $\|X\|_{\text{op}}$, where X is a random matrix with zero-mean.

We are going to introduce 3 approaches to bounding

$$\|X\|_{\text{op}} = \sup_{u,v\in\mathbb{S}^{n-1}} u^T X v.$$

1. From previous lectures, we know $\|X\|_{\text{op}}$ tends to concentrate around its mean $\mathbb{E}\left[\|X\|_{\text{op}}\right]$, because the operator norm is convex and 1-Lipschitz continuous. Then the next step is to bound the expectation of the supremum of an empirical process

$$\mathbb{E}\left[\|X\|_{\text{op}}\right] = \mathbb{E}\left[\sup_{u,v\in\mathbb{S}^{n-1}} u^T X v\right].$$

   This can be achieved by Gaussian comparison inequalities.

2. Using the ε-net argument, we can bound the supremum by discretizing on $\mathbb{S}^{n-1}$ and then invoking union bound.
3. If we write X as the sum of independent matrices, $X = \sum_{i=1}^m X^{(i)}$, there are matrix versions of concentration inequalities (Chernoff, Hoeffding, Berstein) that can help bound $\left\|\sum_{i=1}^m X^{(i)}\right\|_{\text{op}}$.

2 Gaussian Comparison Inequalities

Theorem 1 (Slepian's Inequality)**.** *Let $Z, Y \in \mathbb{R}^N$ be zero-mean Gaussian random vectors such that*

$$\mathbb{E}\left[Z_i^2\right] = \mathbb{E}\left[Y_i^2\right], \forall i \quad (1)$$

$$\mathbb{E}\left[Z_i Z_j\right] \ge \mathbb{E}\left[Y_i Y_j\right], \forall i, j. \quad (2)$$

Then we are guaranteed

$$\mathbb{E}\left[\max_i Z_i\right] \le \mathbb{E}\left[\max_i Y_i\right]. \quad (3)$$

Remark The theorem is basically saying that for zero-mean Gaussian processes, under the condition that variances are equal, high correlations reduce the expectation of maximum. Think of the extreme case where $Z_1 = Z_2 = \cdots = Z_N$. Then it is clear that the behavior of $\{Z_i\}$ is more controlled than $\{Y_i\}$, due to much higher correlations.

Proof For $\beta > 0$, we introduce $F_\beta(x) = \frac{1}{\beta}\log\sum_{i=1}^N e^{\beta x_i}$, which is commonly called the softmax function. Observe that

$$\max_i x_i \le F_\beta(x) \le \max_i x_i + \frac{\log N}{\beta}, \forall \beta > 0.$$

Additionally, F_β is differentiable and $F_\beta(x) \to \max_i x_i$ as $\beta \to +\infty$. So we can use the bound on F_β to control the maximum. Hence F_β really is, by its name, a "soft" version of the maximum.

We assume without loss of generality that Z, Y are independent. Define the Gaussian interpolation

$$X(t) = \sqrt{1-t}Z + \sqrt{t}Y, \qquad \forall t \in [0,1]$$

and consider the function $\phi(t) = \mathbb{E}\left[F_\beta(X(t))\right], \forall t \in [0,1]$. If we can show $\phi'(t) \ge 0, \forall t \in (0,1)$, then we can conclude that $\mathbb{E}\left[F_\beta(Y)\right] = \phi(1) \ge \phi(0) = \mathbb{E}\left[F_\beta(Z)\right]$.

In order to do that, we first use the chain rule to write down the first derivative

$$\phi'(t) = \sum_{j=1}^N \mathbb{E}\left[\frac{\partial F_\beta}{\partial x_j}(X(t))X_j'(t)\right].$$

Note that

$$\begin{aligned}
\mathbb{E}\left[X_i(t)X_j'(t)\right] &= \mathbb{E}\left[\left(\sqrt{1-t}Z_i + \sqrt{t}Y_i\right)\left(-\frac{1}{2\sqrt{1-t}}Z_j + \frac{1}{2\sqrt{t}}Y_j\right)\right] \\
&= \frac{1}{2}\left(\mathbb{E}\left[Y_iY_j\right] - \mathbb{E}\left[Z_iZ_j\right]\right), \qquad \text{by independence and zero-meanness} \\
&\begin{cases} \le 0, & \forall i,j \\ = 0, & i = j, \qquad \text{by assumption (2).} \end{cases}
\end{aligned}$$

So we can write

$$X_i(t) = \alpha_{ij}X_j'(t) + W_{ij},$$

where W_{ij}'s are Gaussian, $W_j := (W_{1j}, \dots, W_{Nj})$ is independent of $X_j'(t)$, and $\alpha_{ij} \le 0, \alpha_{ii} = 0$. [1]

Since F_β is twice differentiable, we may perform Taylor expansion

$$\frac{\partial F_\beta}{\partial x_j}(X(t)) = \frac{\partial F_\beta}{\partial x_j}(W_j) + \sum_{i=1}^N \frac{\partial^2 F_\beta}{\partial x_j \partial x_i}(U)\alpha_{ij}X_j'(t),$$

where $U \in \mathbb{R}^N$ is between $X(t)$ and W_j. Taking expectations gives us

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial F_\beta}{\partial x_j}(X(t))X_j'(t)\right] &= \mathbb{E}\left[\frac{\partial F_\beta}{\partial x_j}(W_j)X_j'(t)\right] + \sum_{i=1}^N \mathbb{E}\left[\frac{\partial^2 F_\beta}{\partial x_j \partial x_i}(U)\alpha_{ij}X_j'(t)^2\right] \\
&= \sum_{i=1}^N \mathbb{E}\left[\frac{\partial^2 F_\beta}{\partial x_j \partial x_i}(U)\alpha_{ij}X_j'(t)^2\right] \qquad \text{because } W_j \perp X_t'(t) and \mathbb{E}\left[X_j'(t)\right] = 0 \\
&\ge 0,
\end{aligned}$$

where the last inequality holds because the soft-max function satisfies $\frac{\partial^2 F_\beta}{\partial x_j \partial x_i}(x) \le 0, \forall x, \forall i \ne j$. Thus we have $\phi'(t) \ge 0, \forall t \in (0,1)$, which yields $\mathbb{E}\left[F_\beta(Z)\right] \le \mathbb{E}\left[F_\beta(Y)\right]$. Taking $\beta \to +\infty$, we get

$$\mathbb{E}\left[\max_i Z_i\right] \le \mathbb{E}\left[\max_i Y_i\right],$$

which completes the proof. □

---

[1] $X_i(t)$ can be seen as generated in this way because Gaussian distribution is determined by its mean and covariance.

Finally, there are some additional points worth mentioning.

- Note that our proof heavily relies on Gaussianity.
- Slepian's inequality holds for any N. In fact, it holds for comparing the expectation of the supremum over infinite sets.
- There is a stronger version called the Sudakov-Fernique theorem.

  Theorem 2 (Sudakov-Fernique)**.** *Let $Z, Y \in \mathbb{R}^N$ be zero-mean Gaussian random vectors. Suppose*
  $$\mathbb{E}\left[(Z_i - Z_j)^2\right] \le \mathbb{E}\left[(Y_i - Y_j)^2\right], \forall i, j. \tag{4}$$
  Then $\mathbb{E}\left[\max_i Z_i\right] \le \mathbb{E}\left[\max_i Y_i\right]$.

  It's easy to see that Slepian's inequality is just a corollary of the Sudakov-Fernique theorem.

3 Applications of Gaussian Comparison Inequalities

Next we return to the problem stated in the beginning.

3.1 Gaussian Matrices

First, we use the Slepian's inequality to bound $\|X\|_{\mathrm{op}}$. We assume $X \in \mathbb{R}^{n\times n}$, whose entries X_{ij}'s are i.i.d. standard normal. We next compare 2 Gaussian processes indexed by (u, v) with $u, v \in \mathbb{S}^{n-1}$,

$$Z_{uv} := u^T X v + \varepsilon = \left\langle X, uv^T \right\rangle + \varepsilon \qquad \text{where } \varepsilon \sim N(0,1) \text{ and } \varepsilon \text{ is independent of } X$$
$$Y_{uv} := g^T u + h^T v \qquad \text{where } g, h \sim N(0, I_n) \text{ and they are independent.}$$

It is easy to see that for all $u, v \in \mathbb{S}^{n-1}$

$$\begin{aligned}\mathbb{E}\left[Z_{uv}^2\right] &= \|u\|_2^2 \|v\|_2^2 + 1 = 2\\ \mathbb{E}\left[Y_{uv}^2\right] &= \|u\|_2^2 + \|v\|_2^2 = 2.\end{aligned}$$

Furthermore, for any $u, v, \tilde{u}, \tilde{v} \in \mathbb{S}^{n-1}$, we have

$$\begin{aligned}\mathbb{E}\left[(Z_{uv} - Z_{\tilde{u},\tilde{v}})^2\right] &= \mathbb{E}\left[\left\langle X, uv^T - \tilde{u}\tilde{v}^T \right\rangle^2\right]\\ &= \left\|uv^T - \tilde{u}\tilde{v}^T\right\|_F^2\\ &= \|\tilde{v}\|_2^2 \|u - \tilde{u}\|_2^2 + \|u\|_2^2 \|v - \tilde{v}\|_2^2 + 2\left(\|u\|_2^2 - \langle u, \tilde{u}\rangle\right)\left(\langle v, \tilde{v}\rangle - \|\tilde{v}\|_2^2\right)\\ &\le \|u - \tilde{u}\|_2^2 + \|v - \tilde{v}\|_2^2,\end{aligned}$$

where the last line can be justified by Cauchy-Schwarz inequality. For the other process, we have

$$\begin{aligned}\mathbb{E}\left[(Y_{uv} - Y_{\tilde{u},\tilde{v}})^2\right] &= \mathbb{E}\left[\left(g^T(u - \tilde{u}) + h^T(v - \tilde{v})\right)^2\right]\\ &= \|u - \tilde{u}\|_2^2 + \|v - \tilde{v}\|_2^2.\end{aligned}$$

Consequently, $\mathbb{E}\left[(Z_{uv} - Z_{\tilde{u},\tilde{v}})^2\right] \le \mathbb{E}\left[(Y_{uv} - Y_{\tilde{u},\tilde{v}})^2\right]$. Hence

$$\begin{aligned}\mathbb{E}\left[Z_{uv} Z_{\tilde{u}\tilde{v}}\right] &= \frac{1}{2}\left(\mathbb{E}\left[Z_{uv}^2\right] + \mathbb{E}\left[Z_{\tilde{u}\tilde{v}}^2\right] - \mathbb{E}\left[(Z_{uv} - Z_{\tilde{u}\tilde{v}})^2\right]\right)\\ &\ge \frac{1}{2}\left(\mathbb{E}\left[Y_{uv}^2\right] + \mathbb{E}\left[Y_{\tilde{u}\tilde{v}}^2\right] - \mathbb{E}\left[(Y_{uv} - Z_{\tilde{u}\tilde{v}})^2\right]\right) \qquad \text{by what we've proved}\\ &= \mathbb{E}\left[Y_{uv} Y_{\tilde{u}\tilde{v}}\right].\end{aligned}$$

Now that we've established the assumptions (1), (2) in Slepain's inequality, we can derive the bound

$$\begin{aligned}
\mathbb{E}\left[\sup_{u,v\in\mathbb{S}^{n-1}} u^T X v\right] &= \mathbb{E}\left[\sup_{u,v\in\mathbb{S}^{n-1}} u^T X v + \varepsilon\right] \\
&\le \mathbb{E}\left[\sup_{u,v\in\mathbb{S}^{n-1}} g^T u + h^T v\right] && \text{by Slepian's inequality} \\
&= \mathbb{E}\left[\|g\|_2 + \|h\|_2\right] \\
&\le \sqrt{\mathbb{E}\left[\|g\|_2^2\right]} + \sqrt{\mathbb{E}\left[\|h\|_2^2\right]} && \text{by Jensen's inequality used on concave function } \sqrt{\cdot} \\
&= 2\sqrt{n}.
\end{aligned}$$

Note that in $\mathbb{E}\left[\|X\|_{\text{op}}\right] \le 2\sqrt{n}$, the constant 2 is tight. It demonstrates Gaussian matrices like X are very well-behaved.

Recall from last lecture, we know

$$\mathbb{P}\left[\left|\|X\|_{\text{op}} - \mathbb{E}\left[\|X\|_{\text{op}}\right]\right| \ge t\right] \le e^{-t^2/4}.$$

Combing this concentration result with our bound on $\mathbb{E}\left[\|X\|_{\text{op}}\right]$, we eventually arrive at

$$\|X\|_{\text{op}} \le (2+\varepsilon)\sqrt{n}, \qquad \text{with probability} \ge 1 - e^{-\varepsilon^2 n/4}. \tag{5}$$

Remark If $X \in R^{n\times m}$, we have $\mathbb{E}\left[\|X\|_{\text{op}}\right] \le \sqrt{n} + \sqrt{m}$. The proof is similar. For Gaussian matrices with heterogeneous variances, refer to this paper: Ramon van Handel, *On the spectral norm of Gaussian random matrices*.[2]

3.2 Matrix Estimation

Recall our ground truth matrix $Y^* \in \mathbb{R}^{n\times n}$ with $\text{rank}(Y^*) \le r$. We observe a $Y = Y^* + E$, where the entries of E are i.i.d. $N(0,1)$. Then we can define our estimator, which is the best rank-r approximation of Y,

$$\hat{Y} = \underset{Z:\text{rank}(Z)\le r}{\arg\min} \|Y - Z\|_{\text{op}}.$$

We first bound the estimation error in spectral norm:

$$\begin{aligned}
\left\|\hat{Y} - Y^*\right\|_{\text{op}} &\le \left\|\hat{Y} - Y\right\|_{\text{op}} + \|Y^* - Y\|_{\text{op}} \\
&\le 2\|Y^* - Y\|_{\text{op}} && \text{by optimality of } \hat{Y} \\
&= 2\|E\|_{\text{op}} \\
&\le 6\sqrt{n}, && \text{with probability } \ge 1 - e^{-n/4}.
\end{aligned}$$

where the last inequality follows from plugging in $\varepsilon = 1$ in (5). Thus

$$\begin{aligned}
\frac{1}{n^2}\left\|\hat{Y} - Y^*\right\|_F^2 &\le \frac{1}{n^2} 2r \left\|\hat{Y} - Y^*\right\|_{\text{op}}^2 && \text{because } \text{rank}(\hat{Y} - Y^*) \le 2r \\
&\lesssim \frac{r}{n}.
\end{aligned}$$

We see that r is considerably less than n, the estimation error is quite small.

[2] `https://www.ams.org/journals/tran/2017-369-11/S0002-9947-2017-06922-1/`