CS 839 Probability and Learning in High Dimension Lecture 1 - 01/26/2022

Lecture 1: Introduction and Matrix Bernstein Inequality

Lecturer: Yudong Chen *Scribe: Weijie Chen*

1 Course Introduction

We Can learn hidden structures, efficiently, from complex noisy data.

A few probabilistic/statistical tools can take us very far.

1.1 What is this course about?

Topic course on the **probability** and **statistical** tools for **high-dimensional** data analysis

- High-dimensional data: It means the data contains many data points, many features or many parameters. Sometimes the number of parameters may be much larger than the number of data points.
- We care about *information* rather than data. The information often exhibits as low-dimensional structure, for example, linearity, sparsity, low-rank, clusters, manifold.
- Probabilistic analysis. We are interested in Data generated from probabilistic/statistical/generative models. This is a (strong) **assumption**, but often worthwhile. In the average case, we have the performance guarantees. We will compare algorithms and analyze the performance limit.

1.2 Main Themes of This Course

This course may interplay between

- **Statistical considerations**. In this part, we will consider estimation accuracy, sample size, model flexibility, and robustness to noise. And we will ask that what can be learned from the data? It is related to the *Information theory*.
- **Computational considerations**. In this part, we will consider fast algorithms (at least poly time) and low storage/communication cost. And we will ask that what can be learned from the data in 1 hour? It is related to the *Optimization theory*

Hence, we will emphasize the connections between **convex/non-convex optimization**, **information theory**, **matrix analysis**.

1.3 Tentative Topics

The following topics may be covered by the course.

- Matrix concentration
- Spectral methods
- Convex relaxation methods
- Matrix and tensor estimation
- Randomized linear algebra

- Nonparametric statistical estimation
- Reinforcement learning and sample complexity Statistical methods based on non-convex optimization
- Information-theoretic lower bounds
- Uniform laws, localization and overparametrization

1.4 Notations Used in This Course

- $f(a) \lesssim g(a)$ or $f(a) = O(g(a))$ means
$$f(a) \leqslant Cg(a) \quad \forall a$$
for a *universal* positive constant C (say C=128) that is independent of any problem parameter (e.g., sample size, dimension, variance, etc.)
- $f(a) \gtrsim g(a)$ or $f(a) = \Omega(g(a))$ means
$$f(a) \geqslant Cg(a) \quad \forall a$$
- $f(a) \asymp g(a)$ or $f(a) = \Theta(g(a))$ means
$$C_1 g(a) \leqslant f(a) \leqslant C_2 g(a) \quad \forall a$$
- $f(a) = o(g(a))$ means
$$\lim_{a \to \infty} \left| \frac{f(a)}{g(a)} \right| = 0$$
- $f(a) = \omega(g(a))$ means $g(a) = o(f(a))$

2 Matrix Bernstein, Spectral Algorithm and Matrix Completion

In this section, we will provide an example of the type of problems and techniques studied in this course.

In particular, we will use the Matrix Bernstein Inequality as a probabilistic tool to analyze a Spectral Algorithm for the Matrix Completion problem. Using this powerful tool, we will show that a simple algorithm and analysis lead to near-optimal performance guarantees.

2.1 Matrix Bernstein Inequality

The following theorem is a generalization of the standard Bernstein inequality to matrices.

Theorem 1 (Matrix Bernstein Inequality)**.** *Suppose $X_1, ..., X_n \in \mathbb{R}^{d \times d}$ are independent, zero-mean,$||x_i||_{op} \leqslant b$. Here $|| \cdot ||_{op}$ is the operator/spectral norm (largest singular value).*

$$\max \left\{ \left\| \sum_i \mathbb{E} X_i^T X_i \right\|_{op}, \left\| \sum_i \mathbb{E} X_i X_i^T \right\|_{op} \right\} \leqslant \sigma^2$$

This $\max \cdot$ *is called "matrix variance". Then we have*

$$\mathbb{P} \left\{ \left\| \sum_i X_i \right\|_{op} \geqslant t \right\} \leqslant 2d \exp \left(-c \min \left\{ \frac{t^2}{\sigma^2}, \frac{t}{b} \right\} \right)$$

Remark

- Note the dimension factor d. It is sometimes sub-optimal, othertimes unavoidable.
- If $d = 1$, we can have the standard Bernstein's inequality.
- We will see the proof in the next lecture.
- The parameter c is the universal constant.

2.2 Application: Spectral algorithm for matrix completion

Let $Y^* \in \mathbb{R}^{d\times d}$ be an unknown rank-r matrix. $|Y^*_{ij} \leqslant 1, \forall i, j$. We observe, independently across i, j,

$$Y_{ij} = \begin{cases} Y^*_{ij} & w.p. \quad p \\ 0 & w.p. \quad 1-p \end{cases}$$

Here p is the observation probability and $p \ll 1$. Our goal is to estimate Y^* given Y.

Note that

$$\mathbb{E}\left[\frac{1}{p}Y_{ij}\right] = p \cdot \frac{1}{p} \cdot Y^*_{ij} + (1-p) \cdot \frac{1}{p} \cdot 0 = Y^*_{ij}$$

for each (i, j), hence $\mathbb{E}\left[\frac{1}{p}Y\right] = Y^*$.

Our **estimator** is

$$\begin{aligned} \hat{Y} &:= \text{best rank-}r \text{ approximation } \frac{1}{p}Y \\ &= \underset{z:\mathrm{rank}(Z)\leqslant r}{\arg\min} \left\| Z - \frac{1}{p}Y \right\|_F \quad \text{given by rank-}r \text{ SVD of } \frac{1}{p}Y \end{aligned}$$

2.2.1 Analysis of $\hat{Y}$

We start with the inequality

$$\begin{aligned} \left\|\hat{Y} - Y^*\right\|_{\mathrm{op}} &\leqslant \left\|\hat{Y} - \frac{1}{p}Y\right\|_{\mathrm{op}} + \left\|\frac{1}{p}Y - Y^*\right\|_{\mathrm{op}} \\ &\leqslant 2\left\|\frac{1}{p}Y - Y^*\right\|_{\mathrm{op}}. \end{aligned}$$

Since $\hat{Y}$ is the best rank-r approximation, we have

$$\begin{aligned} \frac{1}{d^2}\left\|\hat{Y} - Y^*\right\|_F^2 &\leqslant \frac{2r}{d^2}\left\|\hat{Y} - Y^*\right\|_{\mathrm{op}}^2 \quad \mathrm{rank}(\hat{Y} - Y^*) \leqslant 2r \\ &\leqslant \frac{8r}{d^2}\left\|\frac{1}{p}Y - Y^*\right\|_{\mathrm{op}}^2 \end{aligned} \tag{1}$$

Note that $\frac{1}{p}Y - Y^*$ is a zero-mean random matrix. We can control the last RHS using Matrix Bernstein, as done in the following lemma.

Lemma 1. *We have*

$$\left\|\frac{1}{p}Y - Y^*\right\|_{op} \leqslant c_1\left(\sqrt{\frac{d\log d}{p}} + \frac{\log d}{p}\right) \quad w.p. \geqslant 1 - \frac{2}{n^{c_2}}$$

*Note here c*1*,c*2 *are universal constants*

Proof Write $\frac{1}{p}Y - Y^*$ as the sum of independent random matrices:

$$\frac{1}{p}Y - Y^* = \sum_{i,j} X^{(i,j)},$$

where

$$X_{(ij)} \triangleq \left(\frac{1}{p}Y - Y^*\right) e_i e_j^T \in \mathbb{R}^{d\times d}.$$

Here e_i is the i-th standard basis vector $\mathbb{R}^d$. That is, $e_i = [0, 0, 0, ..., 1, ..., 0]$, where the element 1 is in the i-th position.

Observe that

$$\begin{aligned} &\mathbb{E}X(i,j) = 0 \\ &\left\|X^{(i,j)}\right\|_o p = \left|\frac{1}{p}Y_{i,j} - Y^*_{i,j}\right| \leqslant \frac{1}{p} \quad \to \boldsymbol{b} \end{aligned}$$

Moreover, we have

$$\begin{aligned} \mathbb{E}{X^{(i,j)}}^T X^{(i,j)} &= e_j e_j^T \mathbb{E}\left[\left(\frac{1}{p}Y_{ij} - Y^*\right)^2\right] \\ &= e_j e_j^T \left(p(\frac{1}{p} - 1)^2 {Y^*_{ij}}^2 + (1-p){Y^*_{ij}}^2\right) \\ &\overset{\text{(i)}}{\preccurlyeq} \frac{1}{p} e_j e_j^T \end{aligned}$$

Here $A \preccurlyeq B$ means $B - A$ is the positive definite matrix (p.s.d.). The inequality (i) holds since

$$\left(p(\frac{1}{p} - 1)^2 {Y^*_{ij}}^2 + (1-p){Y^*_{ij}}^2\right) \leqslant \frac{1}{p}$$

because $Y^*_{ij} \in [-1, 1]$. It follows that

$$\sum_{i,j} \mathbb{E}{X^{(i,j)}}^T X^{(i,j)} \preccurlyeq \frac{d}{p} I_{d\times d},$$

whence

$$\begin{aligned} &\left\|\mathbb{E}{X^{(i,j)}}^T X^{(i,j)}\right\|_{op} \leqslant \frac{d}{p} \quad \to \boldsymbol{\sigma^2} \\ \text{and similarly} \quad &\left\|\mathbb{E}X^{(i,j)}{X^{(i,j)}}^T\right\|_{op} \leqslant \frac{d}{p}. \end{aligned}$$

Applying matrix Bernstein (Theorem 1)with above $\boldsymbol{b}$ and $\boldsymbol{\sigma^2}$, we have

$$\begin{aligned} \mathbb{P}\left(\left\|\frac{1}{p}Y - Y^*\right\|_{op} \geqslant t\right) =& \mathbb{P}\left(\left\|\sum_{i,j} X^{(i,j)}\right\|_{op} \geqslant t\right) \\ \leqslant& 2d \exp\left(-c \min\left\{\frac{t^2}{\sigma^2}, \frac{t}{b}\right\}\right) \\ =& 2d \exp\left(-c \min\left\{\frac{t^2 p}{d}, tp\right\}\right). \end{aligned}$$

Taking $t = c\left(\sqrt{\frac{d\log d}{p}} + \frac{\log d}{p}\right)$, we get

$$\mathbb{P}\left(\left\|\frac{1}{p}Y - Y^*\right\|_{op} \geqslant t\right) \leqslant \frac{2}{d},$$

thereby proving Lemma 1. □

Combining Lemma 1 with equation (1), we obtain that with high probability, the following error bound holds:

$$\frac{1}{d^2}\left\|\hat{Y} - Y^*\right\|_F^2 \lesssim \frac{r\log d}{dp} + \frac{r\log^2 d}{d^2p^2}.$$

Let us parse the above error bound. Suppose that $p = \frac{r\log d}{d\epsilon^2}$, where $\epsilon \in [0,1]$. Then

$$\frac{1}{d^2}\left\|\hat{Y} - Y^*\right\|_F^2 \lesssim \epsilon^2 + \frac{\epsilon^4}{r} \leqslant 2\epsilon^2$$

- The observation probabilitiy p can be as small as $\frac{r\log d}{d} \ll 1$ if $r \ll d$.
- Total number of observed entries: $pd^2 \approx rd\log d \ll d^2$. In this case, only a small fraction of the entries of Y^* are observed.
- Expected number of observed entries per column/row is $pd \approx r\log d \ll d$

Remark

- The above analysis and error bound can be generalized to the noisy observation setting.
- This error bound is very hard to beat, even with sophisticated algorithms.
- In fact, this bound is un-improvable in noisy setting, up to constant and log factors.

2.3 Research on Matrix Concentration

The matrix Bernstein inequality (Theorem 1) is an example of the so-called *matrix concentration inequalities*, which generalize standard concentration inequalities (e.g., Hoeffding, Bernstein, etc.) for scalar random variables to random matrices.

Matrix concentration is an active field of research, and new results are still combining out at the time of this course. For example, one latest development is the following:

- "Universality and sharp matrix concentration inequalities", by Tatiana Brailovskaya and Ramon van Handel. `https://arxiv.org/abs/2201.05142`
- This paper was posted to arXiv on Jan 13, 2022.
- The results therein remove the dimension factor in Theorem 1 under quite general settings. (For the application in matrix completion problems, there're other ways to remove this factor.)
- In fact, this paper contains more general results.