| CS 839 Probability and Learning in High Dimension | Lecture 2 - 01/31/2022 |
|---|---|

# Lecture 2: Non-parametric Bradley-Terry Model

*Lecturer: Yudong Chen* *Scribe: Augustine(Runshi) Tang*

In this lecture, we will introduce another application of Spectral Algorithm and Matrix Bernstein Inequality, which is to prove a upper bound for a result of Spectral Algorithm used on the so called Non-parametric Bradley-Terry Model.[1]

## 1 Notation

A quick summary of the notation.

1. **Random variables:** $X, Y, U, V$
2. **Ranges/alphabets:** $\mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathcal{V}$
3. **Specific values:** $x, y, u, v$

For a vector $u \in \mathbb{R}^d$, we use $\|u\|_2$ to denote its $\ell_2$ norm and $\|u\|_\infty$ its $\ell_\infty$ norm. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we use $\|A\|_{\mathrm{F}}$ to denote its Frobenius norm and $\|A\|_{\mathrm{op}}$ its operator/spectral norm (i.e., the largest singular value of $A$). For two matrices $A, B$ of the same dimension, $\langle A, B\rangle := \mathsf{tr}(A^\top B)$ denotes their trace inner product. The trace inner product reduces to the usual inner product between vectors for when $A, B \in \mathbb{R}^{d \times 1}$.

## 2 Preliminaries

**Lemma 1** (Matrix Bernstein's inequality[2])**.** *Let $X_1, \ldots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \le K$ almost surely for all $i$. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N X_i\right\|_{op} \ge t\right\} \le 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

*Here $\sigma^2 = \left\|\sum_{i=1}^N \mathbb{E}X_i^2\right\|_{op}$ is the norm of the matrix variance of the sum. In particular, we can express this bound as the mixture of sub-gaussian and sub-exponential tail, just like in the scalar Bernstein's inequality:*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N X_i\right\|_{op} \ge t\right\} \le 2n \exp\left[-c \cdot \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right]$$

**Lemma 2** (Eckart-Young-Mirsky Theorem[3])**.** *Let*

$$D = U\Sigma V^\top \in \mathbb{R}^{m \times n}, \quad m \ge n$$

*be the singular value decomposition (SVD) of $D$ and partition $U, \Sigma =: \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$, and $V$ as follows:*

$$U =: \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \quad \Sigma =: \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \quad \text{and} \quad V =: \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$

[1] *Reference:* Chatterjee, Sourav. "Matrix estimation by universal singular value thresholding." The Annals of Statistics 43.1 (2015): 177-214. Section 2.7.

[2] *Reference:* Theorem 5.4.1 of HDP-book.

[3] *Reference:* https://en.wikipedia.org/wiki/Low-rank_approximation

*where $U_1$ is $m \times r, \Sigma_1$ is $r \times r$, and $V_1$ is $n \times r$, and $\sigma_1 \geq \cdots \geq \sigma_m$. Then the rank- $r$ matrix, obtained from the truncated singular value decomposition*

$$\widehat{D}^* = U_1 \Sigma_1 V_1^\top,$$

*satisfies*

$$\left\| D - \widehat{D}^* \right\|_{\mathrm{F}} = \min_{\operatorname{rank}(\widehat{D}) \leq r} \|D - \widehat{D}\|_{\mathrm{F}} = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_m^2}$$

*The minimizer $\widehat{D}^*$ is unique if and only if $\sigma_{r+1} \neq \sigma_r$.*

In words, $\widehat{D}^*$, given by the truncated SVD of $D$, is a best rank-$r$ approximation of $D$.

# 3 Non-parametric Bradley-Terry Model

Assume an ordered set $\Omega$ with $n$ elements $\omega_i$ and we will use '$\succ$' to denote the ordering. This ordering is unknown, but imagine the setting where one may arrange matches between pairs of items and observe the results of the matches. The results of the matches are random; if $\omega_i \succ \omega_j$, then $\omega_i$ has a higher chance of beating the opponent than $\omega_j$ does agains the same opponent.

The Non-parametric Bradley-Terry model formalized the above setting. If $\omega_i \succ \omega_j$, then $\mathbb{P}(\omega_i \text{ beats } \omega_k) \geq \mathbb{P}(\omega_j \text{ beats } \omega_k)$ for any $k$. Denote a matrix $Y^*$ with $Y_{ij}^* := \mathbb{P}(\omega_i \text{ beats } \omega_j)$. Further denote $Y \in \{0,1\}^{n \times n}$ as an observed random matrix, whose entries independently follow the distribution:

$$Y_{ij} = \begin{cases} 1, & \text{with probability } pY_{ij}^*, \\ 0, & \text{with probability } p(1 - Y_{ij}^*) \\ 0, & \text{with probability } 1 - p, \end{cases}$$

where $p \in [0,1]$. This model can be interpreted as follows: there are $n$ teams and we want to find a ranking among them. With probability $p$, a match is played between a pair of teams independently. We use $Y$ to denote the results of the matches. If $\omega_i$ beats $\omega_j$ in the match, then we write $Y_{ij} = 1$. If $\omega_j$ beats $\omega_i$, or if a match is not played between them, then we write $Y_{ij} = 0$. Our goal is to estimate $Y^*$ given that $Y$.[4]

We use a Spectral Algorithm. In particular, our estimator $\hat{Y}$ is given by the best rank-$r$ approximation of $\frac{Y}{p}$, namely

$$\hat{Y} := \arg \min_{\operatorname{rank}(D) \leq r} \left\| \frac{Y}{p} - D \right\|_{\mathrm{F}},$$

where we choose the rank as $r = \sqrt{np}$. It will be clear why we choose this value later in the proof of the error bound. Note that $\hat{Y}$ can be computed using the truncated SVD of $\frac{Y}{p}$, thanks to the Eckart-Young-Mirsky Theorem.

Our proof relies on two intermediate result. An immediate result from the Matrix Bernstein Inequality is:

**Lemma 3.** *When $p \geqslant (\log n)/n$,*

$$\left\| \frac{1}{p} Y - Y^* \right\|_{op} \lesssim \sqrt{\frac{n \log n}{p}}$$

*holds with high probability.*

The proof follows similar lines as in last lecture and is left as an exercise.

We also claim that:

[4]Given a good estimate of $Y^*$, one may further estimate the ranking. We will not discuss this problem in this lecture.

**Claim 1.** *There exists a rank-$r$ matrix $Z$ such that* $\begin{cases} \|Z - Y^*\|_F^2 \leqslant \frac{n^2}{r} \\ |Z_{ij}| \leqslant 1, \forall i, j \end{cases}$.

**Proof** For each $i = 1, \ldots, n$ define the number $s_i \triangleq \sum_{j=1}^n Y_{ij}^*$. For each $l = 1, \ldots, r$, define the index set $T_l \triangleq \left\{ i : s_i \in \left[ \frac{n(l-1)}{r}, \frac{nl}{r} \right) \right\}$ and let $k(l) \triangleq$ first element in $T_l$.

For each $l = 1, \ldots, r$ and all $i \in T_l$, set

$$Z_{i-} = Y_{k(l)-}^*.$$

This gives a matrix $Z \in [0,1]^{n \times n}$ with row vectors as $Z_{i-}$ and it has rank less than $r$.

For each $l = 1, \ldots, r$ and each $i \in T_l$:

(1) If $\omega_i \succ \omega_{k(l)}$:

$$\begin{aligned} \sum_{j=1}^n \left(Y_{ij}^* - Z_{ij}\right)^2 &= \sum_{j=1}^n \left(Y_{ij}^* - Y_{k(l)j}^*\right)^2 \\ &\leqslant \sum_{j=1}^n \left|Y_{ij}^* - Y_{k(l)j}^*\right| \\ &= \sum_{j=1}^n \left(Y_{ij}^* - Y_{k(l)j}^*\right) \\ &= S_i - S_{k(l)} \\ &\leqslant \frac{n}{r}. \end{aligned}$$

(2) if $\omega_{k(l)} \prec \omega_i$:

$$\begin{aligned} \sum_{j=1}^n \left(Y_{ij}^* - Z_{ij}\right)^2 &= \sum_{j=1}^n \left(Y_{ij}^* - Y_{k(l)j}^*\right)^2 \\ &\leqslant \sum_{j=1}^n \left|Y_{ij}^* - Y_{k(l)j}^*\right| \\ &= -\sum_{j=1}^n \left(Y_{ij}^* - Y_{k(l)j}^*\right) \\ &= S_{k(l)} - S_i \\ &\leqslant \frac{n}{r}. \end{aligned}$$

Sum up over all $i$, we have $\|Y^* - Z\|_F^2 \leqslant \sum_{i=1}^n \frac{n}{r} = \frac{n^2}{r}$. ∎

We are now ready to prove an error upper bound for the estimator $\hat{Y}$ from Spectral Algorithm.

Let $Z$ be the rank-$r$ matrix given by Claim 1. Since $\hat{Y}$ is best rank-$r$ approximation of $\frac{Y}{p}$, we have

$$\left\| \frac{1}{p} Y - Z \right\|_F^2 \geqslant \left\| \frac{1}{p} Y - \hat{Y} \right\|_F^2 = \left\| \frac{1}{p} Y - Z \right\|_F^2 + \left\| \hat{Y} - Z \right\|_F^2 + 2 \left\langle \frac{Y}{p} - Z, Z - \hat{Y} \right\rangle.$$

Rearranging terms gives

$$\begin{aligned}
\|Z-\hat{Y}\|_F^2 &\leqslant 2\left\langle \frac{Y}{p}-Z, \hat{Y}-Z\right\rangle \\
&= 2\left\langle \frac{Y}{p}-Y^*, \hat{Y}-Z\right\rangle + 2\left\langle Y^*-Z, \hat{Y}-Z\right\rangle \\
&\leqslant 2\left\|\frac{1}{p}Y-Y^*\right\|_{op}\|\hat{Y}-Z\|_* + 2\left\|Y^*-Z\right\|_F\|\hat{Y}-Z\|_F \\
&\leqslant 2\left\|\frac{1}{p}Y-Y^*\right\|_{op}\sqrt{2r}\|\hat{Y}-Z\|_F + 2\left\|Y^*-Z\right\|_F\|\hat{Y}-Z\|_F
\end{aligned}$$

So

$$\|Z-\hat{Y}\|_F \leqslant 2\sqrt{2r}\left\|\frac{1}{p}Y-Y^*\right\|_{op} + 2\left\|Y^*-Z\right\|_F$$

Thus

$$\begin{aligned}
\left\|\hat{Y}-Y^*\right\|_F &\leqslant \|Z-\hat{Y}\|_F + \|Z-Y^*\|_F \\
&\leqslant 2\sqrt{2r}\left\|\frac{1}{p}Y-Y^*\right\|_{op} + 3\left\|Z-Y^*\right\|_F \\
&\leqslant C\sqrt{r}\sqrt{\frac{n\log n}{p}} + \frac{3n}{\sqrt{r}} \quad \text{w.h.p. by Lemma 3 and Claim 1 above} \\
&\leqslant C\sqrt{r}\sqrt{\frac{n\log n}{p}} + \frac{3n\sqrt{\log n}}{\sqrt{r}},
\end{aligned}$$

where $C > 0$ is a constant from Lemma 3. Choosing $r = \sqrt{pn}$ (which approximately balances the two terms above), we obtain the following error bound for the spectral algorithm.

**Theorem 1.** *Under the above setting, we have, with high probability,*

$$\frac{1}{n^2}\left\|\hat{Y}-Y^*\right\|_F^2 \lesssim \frac{\log n}{\sqrt{np}}.$$

Note that, if $p \gtrsim \frac{\log^2 n}{n\varepsilon^2}$, then RHS $\leqslant \varepsilon$.