CS 839 Probability and Learning in High Dimension Lecture 7 - 02/16/2022

Lecture 7: Spectral Algorithms II

Lecturer: Yudong Chen *Scribe: Zhuoyan Xu*

In the last lecture, we introduced the singular perturbation theory and its application. In this lecture,[1] we will introduce eigen perturbation theorem for probability transition matrix and discuss its application to the parametric Bradley Terry model.

1 Parametric Bradley-Terry Model

In this section we recall the parametric Bradley-Terry Model.

Suppose we have n items to be ranked, and w_i denote the score for item i, where $w_i > 0, \forall i$. We have:

- item i is better than item $j \iff w_i > w_j$.
- $\mathbb{P}\{\text{item } i \text{ beats item } j\} = \frac{w_i}{w_i+w_j}$.

We observe the matrix $Y \in \{0,1\}^{n\times n}$, where

$$Y_{ij} = \begin{cases} 1 & \text{with probability } \frac{w_j}{w_i+w_j} \\ 0 & \text{with probability } \frac{w_i}{w_i+w_j} \end{cases}$$

Our goal is to estimate $w_1, \ldots, w_n$ given Y. (Note that multiplying all w_i's by the same constant does not change the probabilities above, so we can only hope to estimate w_i's up to a global scaling.)

Remark This model can be extended to partial observation setting: we observe Y_{ij} for a subset of all (i, j)'s.

Remark The Bradley-Terry model is closely related to the ELO rating systems used in many areas, such as Chess, Go, WOW, TopCoder, College Football, and so on.

2 Spectral Algorithm

In this section we apply spectral algorithm to this problem.

Consider a Markov Chain with state space $\{1, 2, \ldots, n\}$ and probability transition matrix $\hat{P} \in [0,1]^{n\times n}$, where

$$\hat{P}_{ij} = \begin{cases} \frac{1}{2n} Y_{ij} & i \neq j \\ 1 - \sum_{l:l\neq i} \frac{1}{2n} Y_{il} & i = j \end{cases} \tag{1}$$

[1] *Reading:*

- "Rank centrality: ranking from pairwise comparisons," S. Negahban, S. Oh, D. Shah, Operations Research, 2016. `https://arxiv.org/abs/1209.1688`
- "Spectral method and regularized MLE are both optimal for top-K ranking," Yuxin Chen, Jianqing Fan, Cong Ma, Kaizheng Wang, Annals of Statistics, 2019. `https://arxiv.org/abs/1707.09971`
- "Lecture Notes for ELE 520", Yuxin Chen, Princeton University.

We then compute stationary distribution $\hat{\pi}$, i.e., the top left eigenvector of $\hat{P}$ such that

$$\hat{\pi} = \hat{\pi}\hat{P}.$$

Remark Practical rating systems can often be viewed as dynamic/online algorithms for computing this eigen vector. For example, one may approximately compute the stationary distribution of $\hat{P}$ by simulating the Markov chain $\hat{P}$; this idea is closely related to Google's PageRank algorithm.

Remark Note that $\hat{P}$ defined above satisfies $P_{ii} \geq 1/2, \forall i$. This is called a *lazy* Markov chain.[2]

Define the population probability transition matrix $P \in [0,1]^{n \times n}$, whose (i,j) entry is

$$P_{ij} := \mathbb{E}[\hat{P}_{ij}] = \begin{cases} \frac{1}{2n} \frac{w_j}{w_i + w_j} & i \neq j \\ 1 - \sum_{\ell : \ell \neq i} \frac{1}{2n} \frac{w_\ell}{w_i + w_\ell} & i = j \end{cases}. \tag{2}$$

We observe that P satisfies the "detailed balance equation":

$$w_i P_{ij} = w_j P_{ji} \quad \forall i, j$$

In this case, it is easy to verify that w after normalization is the stationary distribution of P, i.e., the distribution

$$\pi \triangleq \frac{1}{\sum_i w_i} w$$

satisfies $\pi = \pi P$.

Since $P = \mathbb{E}[\hat{P}]$, intuitively we expect that their stationary distributions, π and $\hat{\pi}$, are close to each other. To make this intuition precise, we want to bound $\hat{P} - P$ and $\hat{\pi} - \pi$.

Note that the matrices P and $\hat{P}$ defined above are asymmetric in general.

2.1 Aside: Uniqueness of Stationary Distribution

One thing we need to consider is whether such the stationary distribution π is unique or not. In general, the stationary distribution might not be unique even when the detailed balance equation is satisfied. One example is when P is the identity matrix, in which case detailed balance equation $w_i P_{ij} = w_j P_{ji}$ holds for any $w \in \mathbb{R}^n$, hence any distribution is stationary for this P.

The stationary distribution is unique if we assume the finite Markov chain P is *irreducible*, meaning that every state can be reached from every other state in a finite number of steps. We give the definition below.

Definition 1 (Irreducible Markov Chain)**.** *A Markov chain with probability transition matrix P is irreducible if for all i, j, there exists a $t > 0$ such that $(P^t)_{ij} > 0$.*

Consider Parametric Bradley-Terry Model and recall the construction of *population* probability transition matrix P in equation (2). Since $w_i > 0, \forall i$ by assumption, P is irreducible (and aperiodic, since the underlying graph is connected with self-loops). On the other hand, to ensure the spectral algorithm is always well-defined, we need to argue that the *empirical* transition matrix $\hat{P}$ is also irreducible. To this end, one may add some regularization when constructing $\hat{P}$; see Section 3.3 in the Negahban-Oh-Shah 2016 paper for details. Alternatively, one may argue that $\hat{P}$ is irreducible provided that P has a sufficiently large eigen gap and $\hat{P} - P$ is small (which holds with high probability under our probabilistic model, as shown in Section 4).

All the results in the following sections hold when the relevant Markov chains have unique stationary distribution.

[2] A lazy reversible Markov chain P has non-negative eigenvalues, since one can write $P = \frac{1}{2}(Q + I)$ for some other reversible Markov chain Q and $\lambda_i(Q) \geq -1, \forall i$ by Perron-Frobenius theorem. Note that any Markov chain can be made lazy without substantially reducing its eigen gap or increasing its mixing time.

3 Eigen Perturbation Theory for Probability Transition Matrices

Developing a eigen perturbation theory for *general* asymmetric matrices is challenging:

1. Eigen values and eigen vectors may be complex valued.
2. Eigen vectors may not be orthogonal.
3. There is no simple relationship between singular values and eigen values.

In this section, we focus on a special class of asymmetric matrices: Probability Transition Matrices. We consider the following setting:

- A Markov Chain $\{X_t, t = 0, 1, \ldots\}$ with n states (items).
- Transition probability from i to j: $P_{ij} = \mathbb{P}\{X_{t+1} = j \mid X_t = i\}$.
- Note that P is a (row-)stochastic matrix: $\sum_j P_{ij} = 1, \forall i$.
- There exists a stationary distribution $\pi = \pi P$.
- We assume that P is reversible with respect to π, meaning that P and π satisfy the detailed balance equation $\pi_i P_{ij} = \pi_j P_{ji} \quad \forall i, j$. In this setting, all the eigenvalues of P are real.[3]

Given nonnegative vector π, we define two norms below:

Definition 2 (Weighted ℓ_2 norm)**.** *For a vector a, $\|a\|_\pi \triangleq \sqrt{\pi_1 a_1^2 + \cdots + \pi_n a_n^2}$.*

Definition 3 (Induced matrix norm)**.** *For a matrix A, $\|A\|_\pi \triangleq \sup_{\|x\|_\pi = 1} \left\|x^\top A\right\|_\pi$.*

Theorem 1. *Suppose that P and $\hat{P}$ are probability transition matrices with stationary distributions π and $\hat{\pi}$, respectively, and that P is reversible. Then:*

$$\|\hat{\pi} - \pi\|_\pi \leqslant \frac{\|\pi(\hat{P} - P)\|_\pi}{\underbrace{1 - \max\{\lambda_2(P), -\lambda_n(P)\}}_{\textit{spectral gap}} - \underbrace{\|\hat{P} - P\|_\pi}_{\textit{perturbation}}}$$

assuming the denominator is strictly positive.

4 Application to Parametric Bradley-Terry Model

This section we apply the previous theorem into Parametric Bradley-Terry Model. To keep things simple, we assume all the scores are on the same order: $\max_{i,j} \frac{w_i}{w_j} \leqslant C$ for a universal constant $C > 0$.[4] Then we can verify:

1. The distribution π is similar to the uniform distribution up to a constant: $\pi_i \asymp \frac{1}{n}, \forall_i$. (This notation means $c_1 \frac{1}{n} \leq \pi_i \leq c_2 \frac{1}{n}$ for universal constants $c_1, c_2 > 0$.)
2. The π-weighted norm and the induced matrix norm are similar to the usual ℓ_2 and spectral norm, respectively: $\|a\|_\pi \asymp \frac{1}{n}\|a\|_2, \forall a \in \mathbb{R}^n$ and $\|A\|_\pi \asymp \|A\|_{op}, \forall A \in \mathbb{R}^{n \times n}$.
3. The transition of the population Markov chain is nearly uniform up to a constant:

$$P_{ij} \triangleq \frac{1}{2n} \frac{w_j}{w_i + w_j} \asymp \frac{1}{n} \quad \forall i \neq j \tag{3}$$

[3]Proof: Let D be the diagonal matrix with $D_{ii} = \sqrt{\pi_i}$ for each i. Under the reversibility assumption, one may verify that the matrix $D^{-1}PD$ is symmetric and hence has real eigenvalues. This implies that P also has real eigenvalues.

[4]This assumption can be relaxed; see the Chen-Fan-Ma-Wang 2019 paper.

From Markov chain theory and the property (3) above, we can show that P has an eigen gap bounded away from zero:

$$1 - \max\{\lambda_2(P), -\lambda_n(P)\} \geqslant c > 0 \tag{4}$$

for some universal constant c.[5]

We further observe that the matrix $\hat{P} - P$ has independent entries bounded by $\frac{1}{n}$. Apply the matrix Bernstein's Inequality (similar to lecture 1), we have with high probability:

$$\|\hat{P} - P\|_\pi \asymp \|\hat{P} - P\|_{op} \lesssim \frac{1}{n}\sqrt{n \log n} \tag{5}$$

Plugging the bounds (4) and (5) into Theorem 1, and using the inequality $\|\pi(\hat{P}-P)\|_\pi \leq \|\pi\|_\pi \|(\hat{P}-P)\|_\pi$, we have (for sufficiently large n and with high probability)

$$\begin{aligned}\|\hat{\pi} - \pi\|_\pi &\lesssim \frac{\|\pi\|_\pi \frac{1}{n}\sqrt{n \log n}}{c - O(\frac{1}{n}\sqrt{n \log n})} \\ &\lesssim \sqrt{\frac{\log n}{n}} \|\pi\|_\pi.\end{aligned}$$

Consequently, the relative error satisfies

$$\frac{\|\hat{\pi} - \pi\|_2}{\|\pi\|_2} \asymp \frac{\|\hat{\pi} - \pi\|_\pi}{\|\pi\|_\pi} \lesssim \sqrt{\frac{\log n}{n}} \longrightarrow 0$$

as $n \longrightarrow \infty$.

Exercise: Generalize the above results to the partial observation setting, where each entry of Y is observed with probability $p \in (0, 1)$.

[5]For details see Lemma 4 in the Negahban-Oh-Shah 2016 paper.