CS 839 Probability and Learning in High Dimension Lecture 6 - 02/14/2022

Lecture 6: Spectral Algorithms I

Lecturer: Yudong Chen *Scribe: Nayoung Lee*

In this lecture,[1] we will introduce the singular perturbation theory. Previously, we mainly focused on the operator or Frobenius norm of the difference in the estimated matrix and the true matrix. This lecture will focus on the singular values/vectors of the estimator and the true matrix. We first begin with a motivating example where a small perturbation in the matrix can lead to a big difference in the singular vector of the matrix and show that a gap in the singular values are needed for stability. Then we introduce a new metric, the subspace distance, and two theorems — the Wedin's $\sin\Theta$ theorem and the Weyl Inequality. We apply the two theorem in a simple case of the matrix completion problem.

1 Notation

A quick summary of the notation.

- For a matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, we use $\|\boldsymbol{A}\|_{\mathrm{F}}$ to denote its Frobenius norm and $\|\boldsymbol{A}\|_{\mathrm{op}}$ its operator/spectral norm (i.e., the largest singular value of $\boldsymbol{A}$).
- $\sigma_i(\boldsymbol{A})$ is the i-th largest singular value of matrix $\boldsymbol{A}$.
- $\mathrm{col}(\boldsymbol{A})$ denotes the column space of the given matrix $\boldsymbol{A}$.

2 Recap

In previous lectures, we covered spectral algorithms for two problems: (i) low-rank matrix completion and (ii) non-parametric Bradley-Terry model. We followed the following setup where given a (noisy/partial) observation $\boldsymbol{Y}$ of some unknown matrix $\boldsymbol{Y}^*$, our goal is to find an good estimator $\hat{\boldsymbol{Y}}$ of $\boldsymbol{Y}^*$.

$$\underset{\text{unknown}}{\boldsymbol{Y}^*} \longrightarrow \underset{\text{observation}}{\boldsymbol{Y}} \longrightarrow \underset{\text{estimator}}{\hat{\boldsymbol{Y}}}$$

We derived bounds on the error of estimating $\boldsymbol{Y}^*$, in terms of $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}^*\|_{op}$ and/or $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}^*\|_F$. However, sometimes the row/column spaces and the eigen/singular vectors of $\boldsymbol{Y}^*$ are of interest rather than the matrix $\boldsymbol{Y}^*$ itself.

3 Motivating Examples

Example 1 (Bradley-Terry Model)**.**

An estimate of true ranking can be extracted from eigenvector of $\hat{\boldsymbol{Y}}$ (will be covered in next lecture)

[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.

Example 2 (Spectral Initialization for Non-convex Matrix Completion[2])**.**

Suppose that the spectral estimator and true matrix have rank-r singular value decomposition $\boldsymbol{Y}^* = \boldsymbol{U}^*\boldsymbol{\Sigma}^*\boldsymbol{V}^{*\top}$ and $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top$. Let

$$\begin{aligned} \boldsymbol{F}_0 &= \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}^{1/2}, \quad \boldsymbol{F}^* = \boldsymbol{U}^*\boldsymbol{\Sigma}^{*1/2}, \\ \boldsymbol{G}_0 &= \hat{\boldsymbol{V}}\hat{\boldsymbol{\Sigma}}^{1/2}, \quad \boldsymbol{G}^* = \boldsymbol{V}^*\boldsymbol{\Sigma}^{*1/2}, \end{aligned}$$

where $\hat{\boldsymbol{F}}, \hat{\boldsymbol{G}}, \boldsymbol{F}^*, \boldsymbol{G}^* \in \mathbb{R}^{n\times r}$. Then the estimator and true matrix can be written in the factorized form $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{F}}\hat{\boldsymbol{G}}^\top$ and $\boldsymbol{Y}^* = \boldsymbol{F}^*\boldsymbol{G}^{*\top}$.

The matrices $\hat{\boldsymbol{F}}, \hat{\boldsymbol{G}}$, given by the singular values/vectors of the spectral estimator, are often used as an initial solution for gradient descent method applied to the non-convex formulation:

$$\min_{\boldsymbol{F},\boldsymbol{G}\in\mathbb{R}^{d\times r}} \sum_{(i,j)\text{ observed}} \left(\left(\boldsymbol{F}\boldsymbol{G}^\top\right)_{i,j} - \boldsymbol{Y}^*_{ij}\right)^2.$$

We want to show that $\hat{\boldsymbol{F}}, \hat{\boldsymbol{G}}$ is close to $\boldsymbol{F}^*, \boldsymbol{G}^*$, so they are in a local convex region of the above (globally nonconvex) objective function. In this region, one may use standard arguments from convex optimization to show that converges to the true $\boldsymbol{F}^*, \boldsymbol{G}^*$.

The question is:

If $\hat{\boldsymbol{Y}} \approx \boldsymbol{Y}^*$, then do we have $\hat{\boldsymbol{F}} \approx \boldsymbol{F}^*$ and $\hat{\boldsymbol{G}} \approx \boldsymbol{G}^*$?

We will return to this example at the end of the lecture.

Remark (Comparison with spectral algorithm)
Recall our previous error bound for spectral algorithm:

$$\frac{1}{n^2}\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}^*\|_F^2 \le \frac{r\log n}{pn}.$$

Even when p is close to 1, in which case most entries are observed, the bound above gives a non-zero error. The non-convex formulation can be viewed as a refinement of the spectral estimator; when p is sufficiently large, this formulation can achieve *exact* recovery of Y^* (under appropriate conditions).

4 Singular Perturbation Theory

Let $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ be the *original matrix*, and $\hat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{H}$ be the *perturbed matrix*, where $\boldsymbol{H}$ is the *perturbation* or noise. Assume $n_1 \ge n_2$. The singular value decomposition (SVD) of $\boldsymbol{M}$ and $\hat{\boldsymbol{M}}$ is given by,

$$\begin{aligned} \boldsymbol{M} &= \begin{bmatrix}\boldsymbol{U}_0 & \boldsymbol{U}_1\end{bmatrix} \begin{bmatrix}\boldsymbol{\Sigma}_0 & 0 \\ 0 & \boldsymbol{\Sigma}_1 \\ 0 & 0\end{bmatrix} \begin{bmatrix}\boldsymbol{V}_0^\top \\ \boldsymbol{V}_1^\top\end{bmatrix}, \\ \hat{\boldsymbol{M}} &= \begin{bmatrix}\hat{\boldsymbol{U}}_0 & \hat{\boldsymbol{U}}_1\end{bmatrix} \begin{bmatrix}\hat{\boldsymbol{\Sigma}}_0 & 0 \\ 0 & \hat{\boldsymbol{\Sigma}}_1 \\ 0 & 0\end{bmatrix} \begin{bmatrix}\hat{\boldsymbol{V}}_0^\top \\ \hat{\boldsymbol{V}}_1^\top\end{bmatrix}, \end{aligned} \tag{1}$$

where

[2]

- Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation, Yudong Chen and Yuejie Chi. IEEE Signal Processing Magazine, vol. 35, no. 4, pp. 14-31, 2018. https://arxiv.org/abs/1802.08397
- Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview, Yuejie Chi, Yue M. Lu, Yuxin Chen. IEEE Transactions on Signal Processing, vol. 67, no. 20, pp. 5239 -5269, 2019. https://arxiv.org/abs/1809.09573

- $\boldsymbol{U}_0 \in \mathbb{R}^{n_1 \times r}, \boldsymbol{\Sigma}_0 \in \mathbb{R}^{r \times r}, \boldsymbol{V}_0 \in \mathbb{R}^{n_2 \times r}$ correspond to the top-r singular values/vectors of $\boldsymbol{M}$;
- $\hat{\boldsymbol{U}}_0 \in \mathbb{R}^{n_1 \times r}, \hat{\boldsymbol{\Sigma}}_0 \in \mathbb{R}^{r \times r}, \hat{\boldsymbol{V}}_0 \in \mathbb{R}^{n_2 \times r}$ correspond to the top-r singular values/vectors of $\hat{\boldsymbol{M}}$;
- similarly, $\boldsymbol{U}_1 \in \mathbb{R}^{n_1 \times (n_1 - r)}, \boldsymbol{\Sigma}_1 \in \mathbb{R}^{(n_2 - r) \times (n_2 - r)}, \boldsymbol{V}_1 \in \mathbb{R}^{n_2 \times (n_2 - r)}$ correspond to the bottom singular values/vectors of $\boldsymbol{M}$;
- $\hat{\boldsymbol{U}}_1 \in \mathbb{R}^{n_1 \times (n_1 - r)}, \hat{\boldsymbol{\Sigma}}_1 \in \mathbb{R}^{(n_2 - r) \times (n_2 - r)}, \hat{\boldsymbol{V}}_1 \in \mathbb{R}^{n_2 \times (n_2 - r)}$ correspond to the bottom singular values/vectors of $\hat{\boldsymbol{M}}$.

Example 3. Consider two matrices $\boldsymbol{M}$ and $\hat{\boldsymbol{M}}$, where $\boldsymbol{M} - \hat{\boldsymbol{M}}$ is small:

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1+\epsilon \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$\hat{\boldsymbol{M}} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix} = \left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right) \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{bmatrix} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right).$$

We consider two cases:

- Setting I: $r = 1$
  In this case, $\boldsymbol{U}_0$ and $\hat{\boldsymbol{U}}_0 \in \mathbb{R}^{2 \times 1}$ correspond to the top-1 left singular vectors of $\boldsymbol{M}$ and $\hat{\boldsymbol{M}}$, respectively. Although $\boldsymbol{M} - \hat{\boldsymbol{M}}$ is small, their top left singular vectors are very different:

  $$\boldsymbol{U}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \hat{\boldsymbol{U}}_0 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

  This is an example where a small perturbation can significantly change the singular vectors of a matrix. Reason: $\sigma_1(\boldsymbol{M}) = \sigma_2(\boldsymbol{M})$. For stability, a gap in singular values $\sigma_1(\boldsymbol{M}) - \sigma_2(\boldsymbol{M}) \gtrsim \epsilon$ is needed.

- Setting II: $r = 2$
  In this case, $\boldsymbol{U}_0$ and $\hat{\boldsymbol{U}}_0 \in \mathbb{R}^{2 \times 2}$ correspond to the top-2 singular vectors of $\boldsymbol{M}$ and $\hat{\boldsymbol{M}}$, respectively. We see that $\boldsymbol{U}_0$ and $\hat{\boldsymbol{U}}_0$ are very different:

  $$\boldsymbol{U}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad \hat{\boldsymbol{U}}_0 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

  However, $\mathrm{col}(\boldsymbol{U}_0) = \mathrm{col}(\hat{\boldsymbol{U}}_0)$; that is, the columns of $\boldsymbol{U}_0$ and $\hat{\boldsymbol{U}}_0$ span the the same vector space $\mathbb{R}^2$, although they have different basis representation.

Setting II above shows that $\|\hat{\boldsymbol{U}}_0 - \boldsymbol{U}_0\|_F$ or $\|\hat{\boldsymbol{U}}_0 - \boldsymbol{U}_0\|_{op}$ is not the right choice of metric. Instead, we use the following subspace distance metric.

Definition 1 (Subspace Distance)**.** *The subspace distance between $\boldsymbol{U}_0$ and $\hat{\boldsymbol{U}}_0$ given in (1) is defined as*

$$\begin{aligned} \mathrm{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0) &\triangleq \|\hat{\boldsymbol{U}}_0 \hat{\boldsymbol{U}}_0^\top - \boldsymbol{U}_0 \boldsymbol{U}_0^\top\|_{op} \\ &= \|\hat{\boldsymbol{U}}_0^\top \boldsymbol{U}_1\|_{op} = \|\boldsymbol{U}_0^\top \hat{\boldsymbol{U}}_1\|_{op} \\ &= \max\{|\sin\theta_1|, \cdots, |\sin\theta_r|\}, \end{aligned}$$

where

$$\theta_i \triangleq \textit{i-th principal angle between col}(\boldsymbol{U}_0) \textit{ and col}(\hat{\boldsymbol{U}}_0).$$

Some remarks:

- $\mathrm{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0) \in [0, 1]$.
- $\hat{\boldsymbol{U}}_0 \hat{\boldsymbol{U}}_0^\top$ and $\boldsymbol{U}_0 \boldsymbol{U}_0^\top$ are the projection matrices onto $\mathrm{col}(\hat{\boldsymbol{U}}_0)$ and $\mathrm{col}(\boldsymbol{U}_0)$, respectively.

- When $r = 1$ (so $\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0 \in \mathbb{R}^n$), we have

$$\theta_1 = \arccos(\hat{\boldsymbol{U}}_0^\top \boldsymbol{U}_0) = \angle(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0),$$

$$(\sin\theta_1)^2 = 1 - (\cos\theta_1)^2 = 1 - (\hat{\boldsymbol{U}}_0^\top \boldsymbol{U}_0)^2 = (\hat{\boldsymbol{U}}_0^\top \boldsymbol{U}_1)^2.$$

- For the proof of the equalities in Definition 1, see, e.g., Part 2 of Theorem I.5.5 in G. W. Stewart and Ji guang Sun (1990), Matrix Perturbation Theory, Academic Press, Boston.

- Technicality: one needs to be a bit careful with the relationship between r and $n - r$, since at most $k := \min\{r, n - r\}$ of the principal angles are nonzero (equivalently, one may define only the first k principal angles, as done in Stewart and Sun's book.)

4.1 Wedin's $\sin\Theta$ Theorem and Weyl Inequality

Recall that $\hat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{H}$. For the singular vectors, we have Wedin's $\sin\Theta$ Theorem.

Theorem 4 (Wedin's $\sin\Theta$ Theorem)**.**
Suppose $\sigma_r(\boldsymbol{M}) - \sigma_{r+1}(\hat{\boldsymbol{M}}) \geq \Delta > 0$. Then the following inequality holds:

$$\begin{aligned}\max\{\mathrm{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0), \mathrm{dist}(\hat{\boldsymbol{V}}_0, \boldsymbol{V}_0)\} &\leq \frac{\max\{\|\boldsymbol{H}\boldsymbol{V}_0\|_{op}, \|\boldsymbol{H}^\top \boldsymbol{U}_0\|_{op}\}}{\Delta} \\ &\leq \frac{\|\boldsymbol{H}\|_{op}}{\Delta}.\end{aligned}$$

The second inequality above follows from $\|\boldsymbol{H}\boldsymbol{V}_0\|_{op} \leq \|\boldsymbol{H}\|_{op}\|\boldsymbol{V}_0\|_{op} = \|\boldsymbol{H}\|_{op}$. Note that the above bound is useful only when the right hand side is less than 1.

For the singular values we have Weyl's Inequality.

Theorem 5 (Weyl's Inequality)**.**
We have

$$|\sigma_i(\boldsymbol{M}) - \sigma_i(\hat{\boldsymbol{M}})| \leq \|\boldsymbol{H}\|_{op}, \quad \forall i = 1, 2, \ldots, n.$$

Consequently

$$\sigma_r(\boldsymbol{M}) - \sigma_{r+1}(\hat{\boldsymbol{M}}) \geq \sigma_r(\boldsymbol{M}) - \sigma_{r+1}(\boldsymbol{M}) - \|\boldsymbol{H}\|_{op}.$$

Combining the above two theorems, we have:

Corollary 1.
The following inequality holds, given that the denominator is positive:

$$\max\{\mathrm{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{M}_0), \mathrm{dist}(\hat{\boldsymbol{V}}_0, \boldsymbol{V}_0)\} \leq \frac{\|\boldsymbol{H}\|_{op}}{\sigma_r(\boldsymbol{M}) - \sigma_{r+1}(\hat{\boldsymbol{M}}) - \|\boldsymbol{H}\|_{op}}.$$

Note that the inequality is valid even if $\sigma_1(\boldsymbol{M}) = \sigma_2(\boldsymbol{M}) = \cdots = \sigma_r(\boldsymbol{M})$ — only the gap between the r-th and $(r+1)$-th singular value is required. The perturbation may change the ordering of these r singular values, but it does not matter since we are using a distance metric in the column space.

Remark
Similar bounds for eigenvectors of symmetric matrices are given in the **Davis-Kahan sinΘ Theorem**.

4.2 Application: Matrix Completion

Suppose $\boldsymbol{Y}^* \in \mathbb{R}^{n\times n}$ is rank r with SVD $\boldsymbol{Y}^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_r), \boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$. From Lecture 1, the spectral algorithm gives a rank-r estimator $\hat{\boldsymbol{Y}}$ bound on $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}^*\|_{op}$.
Now, suppose $\hat{\boldsymbol{Y}}$ has SVD $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top$, where $\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}} \in \mathbb{R}^{n\times r}$. Then the following can be derived.

- Apply Theorem 4 to bound $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{op}$.
- Apply Corollary 1 to get bounds on $\mathrm{dist}(\hat{\boldsymbol{U}}, \boldsymbol{U}), \mathrm{dist}(\hat{\boldsymbol{V}}, \boldsymbol{V})$.
- We can then use $\hat{\boldsymbol{F}} \triangleq \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}^{1/2}$ and $\hat{\boldsymbol{G}}_0 \triangleq \hat{\boldsymbol{V}}\hat{\boldsymbol{\Sigma}}^{1/2}$ at the initial solution for gradient descent for the non-convex formulation in Example 2.

A simple case (rank $r = 1$):
Suppose $\boldsymbol{Y} = \boldsymbol{u}\boldsymbol{v}^\top$ where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{u} = \frac{1}{\|\tilde{\boldsymbol{u}}\|_2}\tilde{\boldsymbol{u}}$, $\boldsymbol{v} = \frac{1}{\|\tilde{\boldsymbol{v}}\|_2}\tilde{\boldsymbol{v}}$, $\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \sim \mathcal{N}(0, \boldsymbol{I}_{n\times n})$. Let the rank-1 estimator $\hat{\boldsymbol{Y}} = \hat{\sigma}\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\top$ be obtained by the spectral algorithm.

Lemma 1.
If the sampling probability satisfies $p \geq C\frac{\log^3 n}{n}$ for a sufficiently large constant C, then w.h.p.,

$$\mathrm{dist}(\hat{\boldsymbol{u}}, \boldsymbol{u}) \leq \frac{1}{100}, \quad \mathrm{dist}(\hat{\boldsymbol{v}}, \boldsymbol{v}) \leq \frac{1}{100}, \quad |\hat{\sigma} - 1| \leq \frac{1}{100}.$$

Sketch of Proof

1. Show $\|\tilde{\boldsymbol{u}}\|_2^2 \asymp \|\tilde{\boldsymbol{v}}\|_2^2 \asymp n$ w.h.p. by scalar Bernstein's inequality
2. Show $\max_i |\tilde{u}_i| \lesssim \sqrt{\log n}, \max_j |\tilde{v}_j| \lesssim \sqrt{\log n}$ w.h.p by bounding each $|\tilde{u}_i|$ and applying a union bound.
3. Bound $\max_{i,j} |Y^*_{ij}|$ by combining 1 and 2.
4. Bound $\|\hat{\boldsymbol{Y}} - \boldsymbol{Y}^*\|_{op}$ (similar to Lecture 1, using matrix Bernstein - Hint: Matrix Bernstein require boundedness, ensured by 3.)
5. Bound $\mathrm{dist}(\hat{\boldsymbol{u}}, \boldsymbol{u}), \mathrm{dist}(\hat{\boldsymbol{v}}, \boldsymbol{v})$ by Corollary 1.

□