CS 839 Probability and Learning in High Dimension Lecture 8 - 02/21/2022

Lecture 8: Convex Relaxation for Community Detection

Lecturer: Yudong Chen *Scribe: Jingqi Duan*

In last few lectures, we focused on the class of spectral methods. Starting from this lecture, we will discuss another class of methods, namely convex relaxation.[1]

1 Convex Relaxation Methods

ML/STAT Model	$\longrightarrow$	(Hard) optimization problem	$\longrightarrow$	convex program
true parameter θ^*		$\min f(\theta)$ s.t. $\theta \in C$		$\min \bar{f}(\theta)$ s.t. $\theta \in \bar{C}$

Suppose that we want to estimate some machine learning or statistical model parameterized by some unknown ground truth parameter θ^*. To estimate this true θ^*, we may construct some optimization problem in which we minimize or maximize an objective function f (such as training loss and likelihood) subject to the constraint $\theta \in C$. Many problems involve the nonconvex optimization, which is in general computationally hard to solve. Convex relaxation methods aim to address this computational challenge by building a convex surrogate of the original optimization problem. Convex optimization problems can often be solved efficiently, and the hope is that the solution obtained is a good estimator of θ^*.

2 Community Detection

We will consider the community detection problem (a.k.a. graph clustering), for which convex relaxation is very powerful. Given a network of n nodes, the high level goal of community detection is to partition nodes into clusters such that there are (1) many connections within clusters and (2) few connections across clusters. That is, we would like to detect community structure in the network.

2.1 Stochastic Block Model (a.k.a. Planted Partition Model)

The Stochastic Block Model (SBM) is a popular probabilistic model for studying community detection. It assumes that the observe graph is generated randomly from some underlying unknown communities.

Specifically, we assume that n nodes are partitioned into k unknown equal-sized clusters, each containing $\frac{n}{k}$ nodes. An edge is placed between node i and j with probability p if i, j are in the same cluster, and with probability q if they are in different clusters. We assume that $p > q$, so on average there are more edges within the cluster than across clusters. The resulting random graph can be represented by an adjacency matrix $A \in \{0,1\}^{n \times n}$, with the distribution

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(p) & \text{if } i, j \text{ in the same cluster} \\ \text{Bernoulli}(q) & \text{if } i, j \text{ in different clusters} \end{cases}$$

independently across all $i \neq j$. We may encode the true clusters by a "cluster matrix" $Y^* \in \{0,1\}^{n \times n}$, where

$$Y^*_{ij} = \begin{cases} 1 & \text{if } i, j \text{ in the same cluster} \\ 0 & \text{if } i, j \text{ in different clusters.} \end{cases}$$

[1] Reading:

- Original paper: [2]
- Improvement: [1]

Note that Y^* is a binary, block-diagonal matrix with k blocks; when $k = 2$, Y^* looks like $\begin{array}{|c|c|}\hline 1 & 0 \\ \hline 0 & 1 \\ \hline \end{array}$ (after appropriately permuting the rows and columns.)

The goal is to estimate true clusters Y^* given the observed graph A.

2.2 SDP Relaxation

Assume that p and q are known (which can be relaxed). We want to find a matrix Y that is maximally correlated with the observed graph A and has the same block-diagional structure as Y^*. Specifically, we try to solve the following optimization problem

$$\begin{aligned} \max_{Y \in \mathbb{R}^{n \times n}} \quad & \left\langle A - \frac{p+q}{2}, Y \right\rangle = \sum_{i,j} \left(A_{ij} - \frac{p+q}{2} \right) Y_{ij} \\ \text{s.t.} \quad & Y_{ij} \in \{0,1\}, \forall i,j \\ & Y_{ii} = 1, \forall i \\ & \sum_j Y_{ij} = n/k, \forall i \\ & \operatorname{rank}(Y) = k \end{aligned} \tag{1}$$

What is the rationale for considering this optimization problem? The expectation of the quantity $A - \frac{p+q}{2}$, which appears in the objective function, is a block matrix of the form $\mathbb{E}\, A - \frac{p+q}{2} = \begin{array}{|c|c|}\hline \frac{p-q}{2} & -\frac{p-q}{2} \\ \hline -\frac{p-q}{2} & \frac{p-q}{2} \\ \hline \end{array}$, which has the property

$$\begin{cases} Y^*_{ij} = 1 & \text{when } \left(\mathbb{E}\, A - \frac{p+q}{2}\right)_{ij} > 0, \\ Y^*_{ij} = 0 & \text{when } \left(\mathbb{E}\, A - \frac{p+q}{2}\right)_{ij} < 0. \end{cases}$$

It is then easy to see that if we were to replace A by $\mathbb{E}A$ in the optimization problem (1), then Y^* would be the unique optimal solution. Therefore, it is reasonable to expect that the optimal solution to the actual optimization problem (1) is a good estimator of Y^*.

However, the optimization problem in (1) is computationally hard to solve due to the nonconvex constraints (the objective function is linear). We may consider a convex relaxation of (1) by replacing nonconvex constraints with convex ones. One such convex relaxations is

$$\begin{aligned} \hat{Y} = \max_{Y} \quad & \left\langle A - \frac{p+q}{2}, Y \right\rangle \\ \text{s.t.} \quad & 0 \le Y_{ij} \le 1, \ \forall i,j \\ & Y_{ii} = 1, \ \forall i \\ & Y \succeq 0. \end{aligned} \tag{2}$$

Note that all feasible solutions to the problem in (1) is also feasible to the convex problem (2); in particular, Y^* is feasible to (2). We remark that the convex problem (2) is a semidefinite program (SDP), which can be solved in polynomial time.

Next, we study whether the convex relaxation solution $\hat{Y}$ is a good estimator of Y^*. We would like to bound the quantity

$$\left\| \hat{Y} - Y^* \right\|_1 \triangleq \sum_{ij} \left| \hat{Y}_{ij} - Y^*_{ij} \right|, \tag{3}$$

which is the element-wise ℓ_1 distance between $\hat{Y}$ and Y^*.

2.3 Explicit Clustering When $k = 2$

In many cases, besides considering whether the matrix $\hat{Y}$ is close to Y^*, we are also interested in estimating the clusters themselves. We would like to extract from $\hat{Y}$ explicit clustering, which are hopefully close to the true clusters. The spectral algorithm is one way to extract clusters from $\hat{Y}$.

Algorithm Here we focus on the special case with $k = 2$ clusters. Recall that $\hat{Y}$ is the optimal solution to the SDP. Compute the top singular vector $\hat{u}$ of the matrix $\hat{Y} - \frac{1}{2}$. For each node i, assign i to cluster 1 if $\hat{u}_i > 0$, and to cluster 2 otherwise.[2]

Our goal is to show these clusters are close to true clusters. In particular, we would to bound the number of nodes that are incorrectly clustered with respect to true clusters.

3 Theoretical Guarantees

We establish the following theorem, which bounds the matrix ℓ_1 error.

Theorem 1 ([2]). *If $p \geq \frac{1}{n}$, then with probability $\geq 1 - 2(\frac{2}{e})^n$, we have*

$$\frac{1}{n^2} \left\| \hat{Y} - Y^* \right\|_1 \lesssim \sqrt{\frac{p}{(p-q)^2 n}}. \tag{4}$$

Recall that n is the number of nodes. p and q are the probability of having an edge between nodes within the cluster and across clusters, respectively. This theorem says that the larger the gap between p and q is, the smaller the error is. The matrix error is normalized by n^2, the number of entries of $\hat{Y}$.

Remark The error bound in (4) is nontrivial when

$$\begin{aligned} &\text{RHS} \lesssim 1 \\ \Rightarrow &\frac{(p-q)^2}{p} \gtrsim \frac{1}{n} \\ \Rightarrow &p \gtrsim \frac{1}{n} \end{aligned}$$

RHS $\lesssim 1$. Hence, this error bound applies even when the edge probabilities are very small, all the way down to $p \asymp \frac{1}{n}$.

- $p \asymp \frac{1}{n}$ is sometimes called the *sparse graph regime*. Community detection is challenging in this regime:
  - $\mathbb{E}$[degree of each node] $= O(1)$.
  - With high probability, the graph is not connected.
- In contrast, $p \gtrsim \frac{\log n}{n}$ is the called dense regime, which is relatively easier:
  - $\mathbb{E}$[degree of each node] $\to \infty$ as $n \to \infty$.
  - With high probability, the graph is connected.

Many previous results on community detection only apply to the dense graph regime, especially those on spectral methods. The spectrum of a graph is quite stable in the dense regime while not well-behaved in the sparse regime with $p \asymp \frac{1}{n}$. A naive spectral method that uses the eigenvectors of the adjacency matrix or the graph Lapacian is known to provably fail in the sparse regime.

It is remarkable that the SDP relaxation approach, without any sophisticated modification, has non-trivial performance guarantees all the way down to the sparse graph regime. This is one example that demonstrates the power of convex relaxation.

[2] To generalize to $k > 2$ clusters, one may take first k singular vectors to form an $n \times k$ matrix, where each row is considered a point in $\mathbb{R}^k$. Then run a clustering algorithm (e.g., the Lloyd's k-means algorithm or Single-Linkage Hierarchical Clustering) on these n points.

3.1 Grothendieck's Inequality

To prove Theorem 1, we make use of the following powerful inequality.

Theorem 2 (Grothendieck's Inequality)**.** *For any $B \in \mathbb{R}^{n\times n}$, it holds that*

$$\max_{\substack{u_i,v_j:\|u_i\|_2=\|v_j\|_2=1\\ \forall i,j=1,\dots,n,}} \left|\sum_{i,j} B_{ij}\langle u_i, v_j\rangle\right| \le K \max_{\substack{x_i,y_j\in\{\pm 1\}\\ \forall i,j=1,\dots,n}} \left|\sum_{i,j} B_{ij}x_i y_j\right|, \tag{5}$$

where $K \le 1.783$.

Remark The constant K is independent of B and n.

Remark LHS of (5) maximizes over infinitely many vectors; RHS maximizes over finitely many numbers. LHS maximizes $\langle B, UV^\top\rangle$ over n-by-n matrices of the form $UV^\top$, which may have an arbitrary rank between 1 and n. RHS maximizes $\langle B, xy^\top\rangle$ over *rank-one* sign matrices of the form $xy^\top$.

Remark RHS is an integer program. If we further assume that $u_i = v_i, \forall i$, then the LHS can be cast as an SDP:

$$\begin{aligned} \max_{Y\in\mathbb{R}^{n\times n}} \quad & \sum_{i,j} B_{ij}Y_{ij} \\ \text{s.t.} \quad & Y \text{ is p.s.d.} \\ & Y_{ii} = 1, \forall i \end{aligned}$$

where $Y_{ij} = \langle u_i, u_j\rangle$ and hence $Y = UU^T$. This SDP is sometimes called the "standard" convex relaxation of the integer program on the RHS. Grothendieck's inequality says that this SDP relaxation is a good approximation.

3.2 Proof of Theorem 1

Proof

Recall that $\hat{Y}$ is an optimal solution to SDP and Y^* is a feasible solution to SDP. Therefore,

$$\begin{aligned} & \left\langle A - \frac{p+q}{2}, \hat{Y}\right\rangle \ge \left\langle A - \frac{p+q}{2}, Y^*\right\rangle \\ \iff \quad & 0 \ge \left\langle Y^* - \hat{Y}, A - \frac{p+q}{2}\right\rangle = \left\langle Y^* - Y, \mathbb{E}\,A - \frac{p+q}{2}\right\rangle + \langle Y^* - Y, A - \mathbb{E}\,A\rangle \end{aligned} \tag{6}$$

Note that if nodes i, j are in the same cluster, $Y^*_{ij} = 1 \ge \hat{Y}_{ij}$; if i, j are in different clusters, $Y^*_{ij} = 0 \le \hat{Y}_{ij}$. It follows that $Y^* - \hat{Y} = \begin{array}{|c|c|}\hline \ge 0 & \le 0 \\ \hline \le 0 & \ge 0 \\ \hline \end{array}$ and $\mathbb{E}\,A - \frac{p+q}{2} = \begin{array}{|c|c|}\hline \frac{p-q}{2} & -\frac{p-q}{2} \\ \hline -\frac{p-q}{2} & \frac{p-q}{2} \\ \hline \end{array}$. In particular, the entries of these two matrices have matching signs. Thus, we obtain that

$$\left\langle Y^* - Y, \mathbb{E}\,A - \frac{p+q}{2}\right\rangle = \frac{p-q}{2}\sum_{i,j}\left|Y^*_{ij} - \hat{Y}_{ij}\right| = \frac{p-q}{2}\left\|\hat{Y} - Y^*\right\|_1 \tag{7}$$

The result in (7) involves the ℓ_1 matrix error that we would like to bound. Combining (7) with (6), we obtain

$$\frac{p-q}{2}\left\|\hat{Y} - Y^*\right\|_1 \le \left\langle \hat{Y} - Y^*, A - \mathbb{E}\,A\right\rangle. \tag{8}$$

To control the RHS in (8), we observe that

$$\begin{aligned}\left\langle \hat{Y} - Y^*, A - \mathbb{E}\, A \right\rangle &\le \left| \left\langle \hat{Y}, A - \mathbb{E}\, A \right\rangle \right| + |\langle Y^*, A - \mathbb{E}\, A\rangle| \qquad \text{(by triangluar inequality)} \\ &\le 2 \max_{\substack{Y_{ii}=1,\forall i \\ Y \succeq 0}} |\langle Y, A - \mathbb{E}\, A\rangle| \\ &\le 2K \max_{\substack{i,j=1,\dots,n \\ x_i, y_j \in \{\pm 1\}}} \left| \sum_{i,j} (A_{ij} - \mathbb{E}\, A_{ij}) x_i y_j \right|. \qquad \text{(by Grothendieck's inequality)} \end{aligned} \tag{9}$$

Note that the last RHS is a maximization over random quantities. Thanks to the Grothendieck's inequality, we only need to bound the maximization over *finitely* many possible x_i's and y_j's. To this end, we can establish a high-probability upper bound for each fixed pair of (x_i, y_j) and then apply a union bound.

Fix an arbitrary pair $(x, y) \in \{\pm 1\}^n \times \{\pm 1\}^n$. Set $Z_{ij} := (A_{ij} - \mathbb{E}\, A_{ij}) x_i y_j$ and $Z := \sum_{i,j} Z_{ij}$. The random variables $\{Z_{ij}\}$ are independent, zero-mean, and bounded by 1 in absolute value. Moreover, for all i, j we have

$$\mathrm{Var}(Z_{ij}) = \mathrm{Var}(A_{ij}) = p(1-p) \text{ or } q(1-q) \le p.$$

Bernstein's inequality ensures that for any $t \ge 0$,

$$\mathbb{P}(|Z| \ge t) \le 2 \exp\left(- \frac{ct^2}{\sum_{i,j} \mathrm{Var}(Z_{ij}) + t} \right) \le 2 \exp\left(- \frac{ct^2}{pn^2 + t} \right).$$

Taking the union bound over all possible x_i and y_j, we obtain

$$\mathbb{P}\left(\max_{(x,y) \in \{\pm 1\}^n \times \{\pm 1\}^n} \left| \sum_{i,j} (A_{ij} - \mathbb{E}\, A_{ij}) x_i y_j \right| \ge t \right) \le 2 \exp\left(- \frac{ct^2}{pn^2 + t} \right) 4^n. \tag{10}$$

We would like to choose t large enough so that RHS in (10) is small. To this send, we set $t = c\left(\sqrt{pn^3} + n\right)$, then RHS $\le (2e)^{-n}$ for a sufficiently large constant $c > 0$. It follows that with probability $\ge 1 - (2e)^{-n}$,

$$\max_{(x,y) \in \{\pm 1\}^n \times \{\pm 1\}^n} \left| \sum_{i,j} (A_{ij} - \mathbb{E}\, A_{ij}) x_i y_j \right| \lesssim \sqrt{pn^3} + n \lesssim \sqrt{pn^3},$$

where the last step holds under the assumption $p \ge \frac{1}{n}$.

Combining the inequality above with (8) and (9), we can conclude that w.h.p.,

$$\frac{p-q}{2} \left\| \hat{Y} - Y^* \right\|_1 \lesssim \sqrt{pn^3} \quad \Longleftrightarrow \quad \frac{1}{n^2} \left\| \hat{Y} - Y^* \right\|_1 \lesssim \sqrt{\frac{p}{(p-q)^2 n}}.$$

□

Remark If we define $\mathrm{SNR} := \frac{(p-q)^2 n}{p}$, which measures the "signal-to-noise ratio" in the stochastic block model, then Theorem 1 says that

$$\frac{1}{n^2} \left\| \hat{Y} - Y^* \right\|_1 \lesssim \sqrt{\frac{1}{\mathrm{SNR}}}.$$

This bound can be improved to

$$\frac{1}{n^2} \left\| \hat{Y} - Y^* \right\|_1 \lesssim e^{-\Omega(\mathrm{SNR})}$$

for the same SDP relaxation [1].

References

[1] Yingjie Fei and Yudong Chen. Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Transactions on Information Theory*, 65(1):551–571, 2019.

[2] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.