CS 839 Probability and Learning in High Dimension Lecture 11 - 03/02/2022

Lecture 11: Convex Relaxation for Community Detection: Exact Recovery

Lecturer: Yudong Chen *Scribe: Zhenmei Shi*

In this lecture,[1] we will present exact recovery guarantees based on convex relaxation for community detection problem. We will also discuss the semirandom stochastic block model and show that convex relaxation is robust under this model.

Notation: Matrix inner product $\langle X, Y \rangle := \sum_{ij} X_{ij} Y_{ij}$. Spectral norm: $\|X\|_{op}$:= largest singular value of X. Nuclear norm $\|X\|_*$:= sum of singular values of X; if X is p.s.d. then $\|X\|_* = \mathrm{tr}(X)$. Entry-wise ℓ_1 norm: $\|X\|_1 := \sum_{i,j} |X_{ij}|$. Entry-wise ℓ_∞ norm: $\|X\|_\infty := \max_{i,j} |X_{ij}|$. Denote by J the $n \times n$ all-one matrix.

1 Stochastic Block Model (SBM)

n nodes partitioned into k unknown equal-sized clusters.

Encode true clusters as a 0-1 matrix $Y^* \in \{0,1\}^{n \times n}$:

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

Note that Y^* is a binary, rank-k, block-diagonal, semidefinite matrix.

Observe: a random graph with adjacency matrix $A \in \{0,1\}^{n \times n}$, such that for two numbers $1 \geq p > q \geq 0$, and independently across all $i \leq j$,

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(p), & \text{if } Y^*_{ij} = 1, \\ \text{Bernoulli}(q), & \text{if } Y^*_{ij} = 0. \end{cases}$$

and $Y_{ij} = Y_{ji}$ for all $i > j$.

The goal is to recover Y^* given A.

2 SDP Relaxation

We consider the same convex relaxation as in previous lectures:

$$\begin{aligned} \widehat{Y} = \arg \max_{Y \in \mathbb{R}^{n \times n}} \quad & \left\langle Y, A - \frac{p+q}{2} J \right\rangle \\ \text{s.t.} \quad & Y \succcurlyeq 0, \\ & 0 \leq Y_{ij} \leq 1, \forall i \in [n], j \in [n] \\ & Y_{ii} = 1, \forall i \in [n]. \end{aligned} \tag{1}$$

Here J is the $n \times n$ all-one matrix.

We will prove the following:

[1]Reading:

- (Survey paper) Xiaodong Li, Yudong Chen and Jiaming Xu. Convex Relaxation Methods for Community Detection. Statistical Science, 2019. https://arxiv.org/abs/1810.00315 (Section 5) [Li et al., 2021]

Theorem 1. *Under the stochastic block model, if*

$$\underbrace{\frac{(p-q)^2 n}{pk^2}}_{\text{SNR}} \gtrsim \log n, \tag{2}$$

then with probability at least $1-3n^{-3}$, $\widehat{Y} = Y^*$ *is the unique optimal solution to the SDP (1).*

Remark To satisfy the condition (2) , we must at least have $p \gtrsim \frac{\log n}{n}$, which is in the dense graph regime. Compare this from the previous lectures, where we proved the approximate recovery result

$$\text{cluster error} \lesssim \frac{1}{n^2}\|\hat{Y} - Y^*\|_1 \lesssim \sqrt{\frac{1}{\text{SNR}}} = \sqrt{\frac{p}{(p-q)^2 n}},$$

which applies to even the sparse regime $p \asymp \frac{1}{n}$. Theorem 1 can be viewed as a more refined guarantee for the dense regime.

3 Proof of Theorem 1

Because $\widehat{Y}$ is optimal to the SDP and Y^* is feasible, we have

$$\begin{aligned}
0 &\leq \left\langle \widehat{Y} - Y^*, A - \frac{p+q}{2}J \right\rangle \\
&= \left\langle \widehat{Y} - Y^*, \mathbb{E}A - \frac{p+q}{2}J \right\rangle + \left\langle \widehat{Y} - Y^*, A - \mathbb{E}A \right\rangle \\
&\Downarrow
\end{aligned}$$

$$\left\langle Y^* - \widehat{Y}, \mathbb{E}A - \frac{p+q}{2}J \right\rangle \leq \left\langle \widehat{Y} - Y^*, A - \mathbb{E}A \right\rangle. \tag{3}$$

We lower bound the LHS and upper bound the RHS in the following two lemmas. The first lemma is proved in Section 3.1.

Lemma 2. *Any* Y *feasible to (1) satisfies*

$$\left\langle Y^* - Y, \mathbb{E}A - \frac{p+q}{2}J \right\rangle = \frac{p-q}{2}\left\|Y - Y^*\right\|_1.$$

The second lemma is proved in Section 3.2.

Lemma 3. *With probability at least* $1-3n^{-3}$, *we have*

$$|\langle Y - Y^*, A - \mathbb{E}A\rangle| \leq C\sqrt{\frac{pk^2 \log n}{n}}\left\|Y - Y^*\right\|_1 \qquad \text{simultaneously for all } Y \text{ feasible to (1)},$$

where $C > 0$ *is a universal constant.*

Applying the two lemmas to the two sides of Eq.(3), we obtain that with probability at least $1-3n^{-3}$,

$$\frac{p-q}{2}\left\|\widehat{Y} - Y^*\right\|_1 \leq C\sqrt{\frac{pk^2 \log n}{n}}\left\|\widehat{Y} - Y^*\right\|_1.$$

When the SNR condition (2) holds, we have $C\sqrt{\frac{pk^2 \log n}{n}} \leq \frac{p-q}{4}$, hence $\frac{p-q}{2}\|\widehat{Y} - Y^*\|_1 \leq \frac{p-q}{4}\|\widehat{Y} - Y^*\|_1$, in which case we must have $\widehat{Y} = Y^*$. This completes the proof of Theorem 1.

3.1 Proof of Lemma 2

Proof We write
$$\left\langle Y^* - \widehat{Y}, \mathbb{E}A - \frac{p+q}{2}J \right\rangle = \sum_{i,j} \left(Y^*_{ij} - Y_{ij}\right)\left(\mathbb{E}A_{ij} - \frac{p+q}{2}\right).$$

Now, for each $(i,j) \in [n] \times [n]$, observe that

- If $Y^* = 1$, then $Y^*_{ij} - Y_{ij} \geq 0$ since $Y_{ij} \leq 1$. Moreover, $\mathbb{E}A_{ij} - \frac{p+q}{2} = p - \frac{p+q}{2} = \frac{p-q}{2}$. It follows that
$$\left(Y^*_{ij} - Y_{ij}\right)\left(\mathbb{E}A_{ij} - \frac{p+q}{2}\right) = \frac{p-q}{2}\left|Y^*_{ij} - Y_{ij}\right|.$$

- If $Y^*_{ij} = 0$, then $Y^*_{ij} - Y_{ij} \leq 0$ since $Y_{ij} \geq 0$. Moreover, $\mathbb{E}A_{ij} - \frac{p+q}{2} = q - \frac{p+q}{2} = -\frac{p-q}{2}$. It follows that
$$\left(Y^*_{ij} - Y_{ij}\right)\left(\mathbb{E}A_{ij} - \frac{p+q}{2}\right) = \frac{p-q}{2}\left|Y^*_{ij} - Y_{ij}\right|.$$

Combining, we see that
$$\sum_{i,j}\left(Y^*_{ij} - Y_{ij}\right)\left(\mathbb{E}A_{ij} - \frac{p+q}{2}\right) = \frac{p-q}{2}\sum_{i,j}\left|Y^*_{ij} - Y_{ij}\right| = \frac{p-q}{2}\left\|Y - Y^*\right\|_1.$$

This completes the proof of Lemma 2. □

3.2 Proof of Lemma 3

Proof Define the shorthand $W := A - \mathbb{E}A$ for the noise matrix. Note that the entries of W are centered Bernoulli random variables, whose variances are either $p(1-p)$ or $q(1-q)$, both of which are less than p.

Let $U \in \mathbb{R}^{n\times k}$ be the matrix whose columns are the top-k singular vectors of Y^*. Explicitly, by equal-sized clusters assumption, we have
$$U_{i\ell} = \begin{cases} \sqrt{\frac{k}{n}} & \text{if node } i \text{ is in cluster } \ell, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, U can be interpreted as the ground-truth *cluster membership* matrix. Note that $UU^\top = \frac{k}{n}Y^*$ and $U^\top U = I_{k\times k}$. We define the projection $P_T : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ and its complement $P_{T^\perp} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ by
$$P_T(A) := UU^\top A + AUU^\top - UU^\top AUU^\top,$$
$$P_{T^\perp}(A) := A - P_T(A) = (I - UU^\top)A(I - UU^\top).$$

Note that P_T is the orthogonal projection onto the set
$$T := \{Z : Z = \underbrace{UB^\top}_{\text{column space}} + \underbrace{EU^\top}_{\text{row space}}, B \in \mathbb{R}^{n\times k}, E \in \mathbb{R}^{n\times k}\},$$

which is the linear subspace spanned by matrices with the same column *or* row space as Y^*. Correspondingly, $P_{T^\perp}$ is the orthogonal projection onto
$$T^\perp := \{Z : U^\top Z = ZU = 0\},$$

the linear subspace of matrices whose column *and* row spaces are orthogonal to Y^*.

Fix an arbitrary Y that is feasible to the SDP. Let $D := Y - Y^*$. Observe that

$$\begin{aligned} |\langle Y - Y^*, W\rangle| &= |\langle P_T D, W\rangle + \langle P_{T^\perp} D, W\rangle| \\ &\le \|D\|_1 \|P_T W\|_\infty + \|P_{T^\perp} D\|_* \|W\|_{op}, \end{aligned} \tag{4}$$

where the last step follows from the generalized Holder's inequality since $\|\cdot\|_1$ ($\|\cdot\|_*$, resp.) is the dual norm of $\|\cdot\|_\infty$ ($\|\cdot\|_{op}$, resp.)

We first bound $\|P_T W\|_\infty$ in Eq. (4). For each (i,j), the quantity

$$(UU^T W)_{ij} = \frac{k}{n} \sum_{\ell : Y^*_{i\ell}=1} W_{\ell j}$$

is the average of $\frac{n}{k}$ independent[2] centered Bernoulli RV's with variance at most p. By the standard Bernstein's inequality and the union bound ,we have

$$\left\|UU^T W\right\|_\infty \le \frac{k}{n} \cdot C \left(\sqrt{\frac{pn}{k} \log n} + \log n \right) \le 2C \sqrt{\frac{pk \log n}{n}} \tag{5}$$

with probability at least $1 - n^{-3}$; the second inequality above holds thanks to the condition (2). The same bound holds for $\left\|WUU^T\right\|_\infty$. Moreover, we have $\left\|UU^T W UU^T\right\|_\infty \le \left\|UU^\top W\right\|_\infty$. It follows from the triangle inequality and the union bound that

$$\|P_T W\|_\infty \le \left\|UU^T W\right\|_\infty + \left\|WUU^T\right\|_\infty + \left\|UU^T W UU^T\right\|_\infty \le 6C \sqrt{\frac{pk \log n}{n}}$$

with probability at least $1 - 2n^{-3}$.

For the second term in Eq. (4), we know that with probability at least $1 - n^{-3}$, there holds

$$\|W\|_{op} \le C\sqrt{pn \log n}; \tag{6}$$

this can be proved using matrix Bernstein inequality. On the other hand, we have the following chain of inequalities:

$$\begin{aligned} \|P_{T^\perp} D\|_* &= \operatorname{tr}(P_{T^\perp} D) && P_{T^\perp}(Y - Y^*) = P_{T^\perp}(Y) = (I - UU^\top) Y (I - UU^\top) \text{ is psd} \\ &= \operatorname{tr}\left((I - UU^\top) D (I - UU^\top)\right) \\ &= \operatorname{tr}\left((I - UU^\top) D\right) && \text{cyclic property of trace; } (I - UU^\top)^2 = I - UU^\top \\ &= \operatorname{tr}(D) - \operatorname{tr}(UU^\top D) \\ &= 0 - \left\langle UU^\top, D \right\rangle && \operatorname{tr}(Y) - \operatorname{tr}(Y^*) = n - n = 0 \\ &\le \left\|UU^\top\right\|_\infty \|D\|_1 \\ &\le \frac{k}{n} \|D\|_1 . \end{aligned}$$

Combining pieces, we conclude that with probability at least $1 - 3n^{-3}$,

$$|\langle Y - Y^*, W\rangle| \le \left(6C \sqrt{\frac{pk \log n}{n}} + C\sqrt{pn \log n} \cdot \frac{k}{n} \right) \|D\|_1 \le 7C \sqrt{\frac{pk^2 \log n}{n}} \|D\|_1 .$$

This completes the proof of Lemma 3. $\square$

[2]They are actually not independent since W is a symmetric matrix. One solution is to write W as the sum of its upper and lower triangular parts and bound each of them separately.

4 Refined Guarantee with Optimal Constant

In this section we focus on the special case with $k = 2$ equal-sized clusters. Let us write the in-cluster and cross-cluster edge probabilities as $p = \frac{a \log n}{n}$ and $q = \frac{b \log n}{n}$, where $a > b > 0$ are two fixed constants.

Theorem 1 shows that SDP achieves exact recovery w.h.p. provided that

$$\frac{(a-b)^2}{a} \geq C$$

for a sufficiently large constant $C > 0$. This condition is implied by

$$\sqrt{a} - \sqrt{b} \geq \sqrt{C}.$$

We can establish a sharper result with an explicit, optimal constant.

Theorem 4. *Under the stochastic block model with $k = 2$ equal sized clusters, if*

$$\sqrt{a} - \sqrt{b} > \sqrt{2}, \tag{7}$$

then with probability at least $1 - n^{-\Omega(1)}$, Y^* is the unique optimal solution to the SDP.

Moreover, it is known that if $\sqrt{a} - \sqrt{b} < \sqrt{2}$, then *all* methods fail to exactly recover Y^* w.h.p.[3] In other words, the condition (7) is *sufficient and necessary* for exact recovery[4]—it is hence called the *sharp threshold*. It is remarkable that this sharp threshold can be achieved by a polynomial-time algorithm (in particular, by SDP relaxation).

For references for this line of work, see Section 3 in [Li et al., 2021] survey paper, which also contains a proof of Theorem 4. Below we discuss the high-level proof ideas in a broader context.

4.1 Two Approaches for Analyzing Convex Relaxation

Consider an abstract statistical setting where Y^* is the unknown ground-truth, and we compute an estimator $\widehat{Y}$ by solving a convex program of the form

$$\widehat{Y} := \arg\min_{Y \in \mathcal{C}} f(Y),$$

where f is a convex function and $\mathcal{C}$ is a convex set satisfying $Y^* \in \mathcal{C}$. Our goal is to show that $\widehat{Y}$ is close to Y^* in a desired sense.[5] There are two general approaches to this goal.

1. **Basic inequality:** We make use of the fact that

   $$f(\widehat{Y}) - f(Y^*) \leq 0, \tag{8}$$

   which holds since $\widehat{Y}$ is optimal and Y^* is feasible. Note that equation (8), sometimes called a *basic inequality*, is a necessary (but not sufficient) condition for the optimality of $\widehat{Y}$. This approach is also referred to as the *primal* approach (as opposed to the dual approach below). This approach is

[3] See

- Abbe, E., Bandeira, A. S. and Hall, G. (2016). Exact Recovery in the Stochastic Block Model. IEEE Transactions on Information Theory 62 471–487. arXiv 1405.3267. [Abbe et al., 2015]
- Mossel, E., Neeman, J. and Sly, A. (2015a). Consistency Thresholds for the Planted Bisection Model. In Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing. STOC '15 69–75. ACM, New York, NY, USA. [Mossel et al., 2015]

[4] Here we ignore the critical point $\sqrt{a} - \sqrt{b} = \sqrt{2}$.

[5] We ignore the issue of uniqueness of $\widehat{Y}$ in this informal discussion.

particularly useful when f is strongly convex or quadratic-like, in which case one can further lower bound the LHS of (8) as

$$0 \geq f(\widehat{Y}) - f(Y^*) \geq \left\langle \nabla f(Y^*), \widehat{Y} - Y^* \right\rangle + \frac{\mu}{2} \left\| \widehat{Y} - Y^* \right\|^2,$$

where μ is the strong convexity parameter (related to Fisher Information). Rearranging terms gives the ℓ_2 error bound

$$\left\| \widehat{Y} - Y^* \right\|^2 \lesssim \mu^{-1} \left\langle \nabla f(Y^*), \widehat{Y} - Y^* \right\rangle.$$

This primal approach is used for establishing Theorem 1 and most of the results covered before this lecture.

2. **Dual certificate:** Here we make use of a *sufficient and necessary* condition for the optimality of $\widehat{Y}$. For unconstrained convex optimization (i.e., $\mathcal{C} = \mathbb{R}^{n \times n}$), $\widehat{Y}$ is optimal if and only if $\nabla f(\widehat{Y}) = 0$, from which one can deduce the properties of $\widehat{Y}$. For example, in sparse linear regression one may use this approach to study the support recovery property of the Lasso estimator. As an important special case of this approach, one may establish *exact* recovery by showing $\nabla f(Y^*) = 0$. In the more general, constrained optimization setting, a sufficient and necessary condition for optimality is the Karush-Kuhn-Tucker (KKT) condition, which involves certain Lagrangian multipliers/dual variables. These multipliers are sometimes called a dual certificate for the optimality of $\widehat{Y}$. The dual certificate approach is more general/precise than the basic inequality approach, but it is also often more complicated, as it involves proving existence of the certificate. Theorem 4 and many other exact recovery results are established using this approach.

We remark that there are other approaches, e.g., primal-dual approach, which combines the power of the above two.

5 Semi-random Robustness of Convex Relaxation

In this section, we will introduce semirandom models.[6] where data points are generated from a random model and then an adversary changes all data points in a monotone but otherwise arbitrary way.

5.1 Semi-random Stochastic Block Model

Unknown partition of n nodes into k clusters, represented by the groundtruth cluster matrix $Y^* \in \{0,1\}^{n \times n}$.

Generate random graph $A \in \{0,1\}^{n \times n}$, such that for two numbers $1 \geq p > q \geq 0$,

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

Then, an adversary observes A and is allowed to

[6] The following papers motivate the study of semirandom models. In particular, they argue that the semirandom models are not easier than the original models; in contrast, they foil many existing algorithms.

- Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection? In Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, 2016. [Moitra et al., 2016]
- M. Krivelevich, D. Vilenchik. Semirandom models as benchmarks for coloring algorithms. In: ANALCO, pp. 211–221 (2006) [Krivelevich and Vilenchik, 2006]
- Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. Journal of Computer and System Sciences, 63(4):639–671, 2001. [Feige and Kilian, 2001]

The result in this section is folklore. One version of this result was proved in Lemma 1 of the following paper:

- Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved Graph Clustering. IEEE Transactions on Information Theory, vol. 60, no. 10, pp. 6440–6455, 2014. https://arxiv.org/abs/1210.3335 [Chen et al., 2014]

- remove some edges between clusters,
- add some edges inside clusters.

The two actions could be correlated and depend on A (and the algorithm used for community detection). Let A' be the modified graph produced by the adversary.

The goal is to recover the true clusters Y^* given the modified graph A'.

Remark The adversary does not necessarily make the problem easier:

- Random models are very "rigid" and we have many tools such as sharp concentration for degrees, eigenvalue/vectors of A, triangle count and so on. Under the semirandom model, the adversary may destroy these rigid structures.
- Many standard algorithms over-exploit properties of random models and provably fail under semi-random models, e.g., spectral algorithms and algorithms based on counting degrees/triangles in the graph.

5.2 Robustness of SDP Relaxation

Optimization-based methods (e.g., SDP relaxation) are often more robust under the semirandom model. Sometimes SDP achieves such robustness "automatically".

Let SDP(A) denote the following SDP with the original graph A as the input:

$$\begin{aligned}\max_{Y} \ & f_A(Y) \triangleq \langle Y, A - \frac{p+q}{2} \rangle \\ \text{s.t. } & Y \succcurlyeq 0, \\ & Y_{ii} = 1, \forall i, \\ & 0 \le Y_{ii} \le 1, \forall i, j.\end{aligned}$$

Correspondingly, SDP(A') is the SDP with the modified graph A' as the input.

The following theorem shows that if SDP achieves exact recovery under the original graph A, then it achieves the same under the modified graph A'.

Theorem 5. *If Y^* is the unique optimum of SDP(A), then Y^* is also the unique optimum of SDP(A').*

Proof Since Y^* is the unique optimum for SDP(A), we have

$$f_A(Y^*) > f_A(Y)$$

for all feasible Y. On the other hand, we have

$$\begin{aligned} f_{A'}(Y^*) - f_A(Y^*) =& \sum_{i,j} (A'_{ij} - A_{ij}) Y^*_{ij} \\ \overset{(*)}{\ge}& \sum_{i,j} (A'_{ij} - A_{ij}) Y_{ij} \\ =& f_{A'}(Y) - f_A(Y). \end{aligned}$$

Adding up the last two display equations, we obtain $f_{A'}(Y^*) > f_{A'}(Y)$, for all feasible Y. We hence conclude that Y^* is the unique optimum of SDP(A').

Proof of $(*)$:

$$\begin{aligned}
\text{if } i,j \text{ in the same cluster } &\implies A'_{ij} - A_{ij} \geq 0, Y^*_{ij} = 1 \geq Y_{ij}, \\
\text{if } i,j \text{ in different clusters } &\implies A'_{ij} - A_{ij} \leq 0, Y^*_{ij} = 0 \leq Y_{ij}.
\end{aligned}$$

$\square$

5.2.1 Consequence

Recall Theorem 1: If $(p-q)^2 \gtrsim \frac{pk^2 \log n}{n}$, then with high probability, we have Y^* is the unique optimal solution to SDP(A). Combining with Theorem 5, we have the following corollary for exact recovery under the semirandom model.

Corollary 1. *Let A' be generated from the semirandom SBM with parameters n, k, p, q. If $(p-q)^2 \gtrsim \frac{pk^2 \log n}{n}$, then with high probability, Y^* is the unique optimal solution for SDP(A').*

References

[Abbe et al., 2015] Abbe, E., Bandeira, A. S., and Hall, G. (2015). Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487.

[Chen et al., 2014] Chen, Y., Sanghavi, S., and Xu, H. (2014). Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455.

[Feige and Kilian, 2001] Feige, U. and Kilian, J. (2001). Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671.

[Krivelevich and Vilenchik, 2006] Krivelevich, M. and Vilenchik, D. (2006). Semirandom models as benchmarks for coloring algorithms. In *2006 Proceedings of the Third Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 211–221. SIAM.

[Li et al., 2021] Li, X., Chen, Y., and Xu, J. (2021). Convex relaxation methods for community detection. *Statistical Science*, 36(1):2–15.

[Moitra et al., 2016] Moitra, A., Perry, W., and Wein, A. S. (2016). How robust are reconstruction thresholds for community detection? In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 828–841.

[Mossel et al., 2015] Mossel, E., Neeman, J., and Sly, A. (2015). Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 69–75.