Scribed by Elizabeth Yang

Lecture 14

In which we finish the proof of exact reconstruction in the stochastic block model, and introduce a new semirandom model with an adversarial component.

1 Overview

We first continue the proof of exact reconstruction in the stochastic block model (SBM), by proving that the optimum to the minimum bisection SDP we have studied so far is also the *unique* optimum. We then introduce **semirandom models**, where both random and adversarial choices are allowed. We will demonstrate some analogous results to the non-adversarial SBM: how to achieve ε-approximate recovery and exact recovery after introducing a special set of adversarial choices.

2 Exact Reconstruction, Continued

Recall our setup for the stochastic block model: We construct a random graph $G = (V, E)$, with $|V| = n$. We partition V into S_1 and S_2, with $|S_1| = |S_2| = \frac{n}{2}$. We place edges between vertices of the same S_i with probability p, and edges between our partition sets with probability $q < p$.

We write $a = \frac{np}{2}$ and $b = \frac{nq}{2}$. Last lecture, we started proving the following theorem:

Theorem 1 *Given $G = (V, E)$ drawn from our SBM distribution, if $a - b > c \cdot \sqrt{\log n} \cdot \sqrt{a+b}$, then we can exactly reconstruct our bisection $V = (S_1, S_2)$ with high probability.*

Recap: We began our proof in Lecture 12 by writing our SDP for minimum bisection.

$$\max \sum_{u,v} A_{u,v} \langle \mathbf{x}_u, \mathbf{x}_v \rangle$$
$$\text{subject to } \|\mathbf{x}_v\|^2 = 1 \text{ for all } v \in V$$
$$\|\sum_v \mathbf{x}_v\|^2 = 0$$

We define our intended solution $\{\mathbf{x}_v\}_{v\in V}$ below, and the matrix X by $X_{uv} = \langle \mathbf{x}_u, \mathbf{x}_v\rangle$:

$$\mathbf{x}_v = \frac{1}{\sqrt{n}}[1, 1, \dots, 1] \text{ if } v \in V_1 \text{ and } \mathbf{x}_v = \frac{1}{\sqrt{n}}[-1, -1, \dots, -1] \text{ if } v \in V_2$$
$$\text{so } X_{u,v} = 1 \text{ if } u, v \text{ are on the same side}$$
$$X_{u,v} = -1 \text{ if } u, v \text{ are on different sides}$$

Recall that $X = \chi\chi^T$, where χ is an indicator vector for the cut:

$$\chi \in \mathbb{R}^n : \chi_v = 1 \text{ if } v \in S_1, \chi_v = -1 \text{ if } v \in S_2$$

Let a_v be the number of neighbors of v on the same side of the bisection, and let b_v be the number of neighbors of v on the opposite. Then:

$$\text{Cost}(\{\mathbf{x}_v\}_{v\in V}) = \sum_v (a_v - b_v)$$

We proved that $\{\mathbf{x}_v\}_{v\in V}$ was an optimal solution to our SDP by showing that it was feasible, applying SDP duality and obtaining a dual solution of the same cost. Below is the dual:

$$\min \sum_{v=1}^{n} y_v$$
$$\text{subject to } \text{diag}(y_1, \dots, y_n) + y_0 \cdot J \succeq A$$

We saw that with high probability, the solution $y_0 = \frac{a+b}{2}$ and $y_v = (a_v - b_v)$ is feasible for the dual. Then, $\text{Cost}(\mathbf{y}) = \text{Cost}(\{y_v\}_{v=0}^{n}) = \sum_{v=1}^{n} y_v = \sum_v (a_v - b_v)$ as well.

Claim 2 *The solution $\{\mathbf{x}_v\}_{v\in v}$ is the **unique** optimum for the minimum bisection SDP.*

Proof: Fix graph G. We previously showed that $y_0, \dots, y_n$ defined above is feasible for the dual, with high probability. This implies:

$$\begin{aligned}
\sum_v (a_v - b_v) &= \text{Cost}(\{\mathbf{x}_v\}_{v\in V}) = A \cdot X \\
&\leq [\text{diag}(a_1 - b_1, \dots, a_n - b_n) + \frac{a+b}{n} \cdot J] \cdot X \\
&\leq \text{diag}(a_1 - b_1, \dots, a_n - b_n) \cdot X + \frac{a+b}{n}[J \cdot X] \\
&= \sum_v y_v \cdot X_{v,v} + \frac{a+b}{n} \cdot \sum_{u,v} \langle \mathbf{x}_u, \mathbf{x}_v\rangle \\
&= \sum_v y_v + \frac{a+b}{n} \|\sum_v \mathbf{x}_v\|^2 \\
&= \text{Cost}(\mathbf{y}) + 0 = \sum_v (a_v - b_v)
\end{aligned}$$

Thus, all of the above inequalities are actually equalities, and we have:

$$A \cdot X = [\operatorname{diag}(a_1 - b_1, \ldots, a_n - b_n) + \frac{a+b}{n} \cdot J] \cdot X$$

$$\text{which implies } [\operatorname{diag}(a_1 - b_1, \ldots, a_n - b_n) + \frac{a+b}{n} \cdot J - A] \cdot X = 0$$

Let $M = [\operatorname{diag}(a_1 - b_1, \ldots, a_n - b_n) + \frac{a+b}{n} \cdot J - A]$. To show uniqueness of our solution X, it suffices to show that the X is the only solution that satisfies $M \cdot X = 0$.

Recall in Lecture 12, that we showed $M\chi = \vec{0}$. We also showed for all $\mathbf{x} \perp \chi$:

$$\mathbf{x}^T M \mathbf{x} \geq (a - b) - O(\sqrt{\log n}\sqrt{a+b}) > 0$$

We can also write SDPs as positive combinations of certain rank 1 matrices $\mathbf{z}_i \mathbf{z}_i^T$, so:

$$X = \sum_i \lambda_i \mathbf{z}_i \mathbf{z}_i^T \text{ where } \lambda_i > 0$$

$$M \cdot X = M \cdot (\sum_i \lambda_i \mathbf{z}_i \mathbf{z}_i^T) M = \sum_i \lambda_i (\mathbf{z}_i^T M \mathbf{z}_i)$$

The quantity $\mathbf{z}_i^T M \mathbf{z}_i$ will always be positive, unless $\mathbf{z}_i = \chi$. Therefore, if $M \cdot X = 0$, we must have $X = \chi\chi^T$, which proves uniqueness of our solution. We conclude that this solution is also optimal for the minimum bisection problem. □

3 Semirandom Models for Cut Problems

While there are many different semirandom models that are analogous to various random graph and planted solution models, we focus this lecture on the semirandom stochastic block model. Here, we have a random selection of a graph, but we also have an adversary that is permitted to perform certain moves on the graph. Construct our graph as follows:

1. Start with the SBM, with edge probabilities p, q and bisection $V = (S_1, S_2)$.
2. The adversary is now allowed to do any of the following steps as much as it wants:
    - Add non-crossing edges, effectively "increasing p."
    - Remove crossing edges, effectively "decreasing q."

It may appear that the adversary only helps us recover the bisection, as the quantity $a - b$ is larger now. Intuitively, the adversary is making the clusters from the bisection more apparent. However, it is not obvious that the SDP techniques we developed in the previous lectures still apply here. The adversary can change the graph so that $\|A - \frac{d}{n}J\|$ no longer tells us about the hidden cut.

We see now that our SDP techniques can still apply to the semirandom stochastic block model, with some relatively simple changes to our proofs.

3.1 Exact reconstruction

The aim of this section is to prove the following analogue in the semirandom model:

Theorem 3 *Given $G = (V, E)$ from the semirandom SBM, if $a - b \geq c \cdot \sqrt{\log n} \cdot \sqrt{a+b}$, then we are able to exactly reconstruct our bisection $V = (S_1, S_2)$ with high probability.*

The proof requires the following lemma:

Lemma 4 *If M is symmetric, and for all i, we have $M_{i,i} \geq \sum_{j \neq i} |M_{i,j}|$, then $M \succeq 0$. (We call M "symmetric and diagonally-dominated," or **SDD**.)*

Proof:[Proof of Lemma 4] Fix an $\mathbf{x}$. Then:

$$\begin{aligned}
\mathbf{x}^T M \mathbf{x} = \sum_{i,j} M_{ij}\mathbf{x}_i\mathbf{x}_j &= \sum_i M_{ii}\mathbf{x}_i^2 + \sum_{i \neq j} M_{ij}\mathbf{x}_i\mathbf{x}_j \\
&\geq \sum_i M_{ii}\mathbf{x}_i^2 - \sum_{i \neq j} |M_{ij}||\mathbf{x}_i| \cdot |\mathbf{x}_j| \\
&\geq \sum_i M_{ii}\mathbf{x}_i^2 - \sum_{i \neq j} |M_{ij}| \cdot (\frac{\mathbf{x}_i^2 + \mathbf{x}_j^2}{2}) \\
&=\geq \sum_i M_{ii}\mathbf{x}_i^2 - \sum_i (\frac{\mathbf{x}_i^2}{2} \cdot \sum_{j \neq i} |M_{ij}|) - \sum_j (\frac{\mathbf{x}_j^2}{2} \cdot \sum_{i \neq j} |M_{ij}|) \\
&= \sum_i \mathbf{x}_i^2 \cdot (M_{ii} - \sum_{j \neq i} |M_{ij}|) \geq 0
\end{aligned}$$

We obtain the third line by applying Cauchy-Schwarz, and the final line by using the fact that M is symmetric. □

Proof:[Proof of Theorem 3] We use the same SDP for minimum bisection:

$$\begin{aligned}
&\max \sum_{u,v} A_{u,v} \langle \mathbf{x}_u, \mathbf{x}_v \rangle \\
&\text{subject to } \|\mathbf{x}_v\|^2 = 1 \text{ for all } v \in V \\
&\|\sum_v \mathbf{x}_v\|^2 = 0
\end{aligned}$$

However, our adjacency matrix A now behaves differently. We can think of $A = A_R + A_N$, where A_R is the adjacency matrix for the fully random contribution to G's construction, and A_N is the adjacency matrix for the adversary's contribution.

Note that A_R is the adjacency matrix for an instance of the non-adversarial SBM, while A_N has entries in $\{-1, 0, 1\}$. 0 corresponds to an edge untouched, 1 to an edge added, and -1 to an edge removes.

Despite these changes from our usual A_R, we can still show that the indicator for the hidden bisection (S_1, S_2) is the unique optimal solution to our SDP. Since our constraints did not change, it is certainly feasible. To analyze the cost, define:

$$\begin{aligned}
a_v &= \text{ the number of randomly chosen neighbors on the same side} \\
b_v &= \text{ the number of randomly chosen neighbors on the opposite side} \\
\hat{a}_v &= \text{ the number of non-random neighbors on the same side} \\
\hat{b}_v &= \text{ the number of deleted neighbors on the opposite side}
\end{aligned}$$

Using the same indicator solution $\{\mathbf{x}_v\}_{v \in V}$ (more specifically, $\mathbf{x}_v = \frac{1}{\sqrt{n}}[1, 1, \ldots, 1]$ if $v \in V_1$ and $\mathbf{x}_v = \frac{1}{\sqrt{n}}[-1, -1, \ldots, -1]$ if $v \in V_2$), we then have:

$$\text{Cost}(\{\mathbf{x}_v\}_{v \in V}) = \sum_v (a_v - b_v) + \sum_v (\hat{a}_v + \hat{b}_v)$$

As before, we want to show that there is still a feasible dual solution with the same cost, to prove optimality of the hidden cut in the primal, For reference, here is the dual:

$$\min \sum_{v=1}^{n} y_v$$
$$\text{subject to } \text{diag}(y_1, \ldots, y_n) + y_0 \cdot J \succeq A$$

A natural candidate for a solution here is $y_0 = \frac{a+b}{n}$, and $y_v = (a_v - b_v) + (\hat{a}_v + \hat{b}_v)$.

$$\text{Cost}(\mathbf{y}) = \sum_v (a_v - b_v) + (\hat{a}_v + \hat{b}_v) = \text{Cost}(\{\mathbf{x}_v\}_{v \in V})$$

This solution has the desired cost. We are now left to see if:

$$\text{diag}((a_1 - b_1) + (\hat{a}_v + \hat{b}_v), \ldots, (a_n - b_n) + (\hat{a}_v + \hat{b}_v)) + \frac{a+b}{n} \cdot J \succeq A_R + A_N$$

We already know that $\text{diag}(a_1 - b_1, \ldots, a_n - b_n) + \frac{a+b}{n} \cdot J \succeq A_R$, so it suffices to prove:

$$\text{diag}(\hat{a}_v + \hat{b}_v, \ldots, \hat{a}_v + \hat{b}_v) - A_N \succeq 0$$

We can now apply Lemma 4 to $\text{diag}(\hat{a}_v + \hat{b}_v, \ldots, \hat{a}_v + \hat{b}_v) - A_N$, and conclude that it is PSD. We have now established that our current $\mathbf{y} = \{y_v\}_{v \in V}$ is feasible.

As before, to prove uniqueness, we define the matrix M as follows:

$$\begin{aligned}
M &= \text{diag}((a_1 - b_1) + (\hat{a}_v + \hat{b}_v), \ldots, (a_n - b_n) + \frac{a+b}{n} \cdot J - (A_R + A_N) \\
&= [\text{diag}(a_1 - b_1, \ldots, a_n - b_n) + \frac{a+b}{n} \cdot J - A_R] + [\text{diag}(\hat{a}_1 + \hat{b_1}, \ldots, \hat{a_n} + \hat{b_n}) - A_N]
\end{aligned}$$

Let M_1 be the first "random" part that is the same as the matrix non-adversarial case, and M_2 be the "adversarial" part we introduced in this proof. Again, we show that our X with

$X_{u,v} = \langle \mathbf{x}_u, \mathbf{x}_v \rangle$ is the unique solution to $M \cdot X = 0$. And again, let χ be the indicator vector for our cut. Consider any $\mathbf{x} \perp \chi$:

$$\mathbf{x}^T M \mathbf{x} = \mathbf{x}^T (M_1 + M_2)\mathbf{x} = \mathbf{x}^T M_1 \mathbf{x} + \mathbf{x}^T M_2 \mathbf{x} \geq \mathbf{x}^T M_1 \mathbf{x} > 0$$

Here, we see that $\mathbf{x}^T M_2 \mathbf{x} \geq 0$ since we proved that M_2 is PSD earlier. We also know that $\mathbf{x}^T M_1 \mathbf{x} > 0$, with strict inequality, from the uniqueness proof in the non-adversarial setting. This tells us that X is the unique optimum as well. □

We now ask: where did we use the assumption that the adversary only removed crossing edges, and only added non-crossing edges? That assumption guarantees we **add** $\hat{a}_v$ and $\hat{b}_v$.

3.2 ε-approximate reconstruction

We can also get an analogue of ε-approxmate reconstruction in the semirandom model:

Theorem 5 *Given $G = (V, E)$ drawn from the semirandom SBM and $\varepsilon > 0$, if $a - b \geq c_\varepsilon \cdot \sqrt{a+b}$, then we can exactly reconstruct our bisection $V = (S_1, S_2)$ with fewer than εn misclassified vertices.*

PROOF: We will use the same SDP for minimum bisection, and the same solution $\{\mathbf{x}_v\}_{v \in V}$ as in the previous two sections. Then, applying a fact from ε-approximate recovery in the non-adversarial SBM:

$$\text{Cost}(\{\mathbf{x}_v\}_{v\in V}) = \sum_v (a_v - b_v) + \sum_v (\hat{a}_v + \hat{b}_v) \geq_{\text{w.h.p.}} [n(a-b) - O(n)] + \sum_v (\hat{a}_v + \hat{b}_v)$$

For the non-adversarial case (with adjacency matrix A_R as defined in the previous section), we used Grothendieck's inequality and Chernoff bounds to show that w.h.p., for all $\{\mathbf{x}_v\}$ and $\{\mathbf{y}_v\}$ such that $\|\mathbf{x}_v\|^2 = \|\mathbf{y}_v\|^2 = 1$:

$$\sum_{u,v} [(A_R)_{uv} - (\mathbb{E}(A_R))_{uv}] \cdot \langle \mathbf{x}_u, \mathbf{y}_v \rangle \leq O(n\sqrt{a+b})$$

We compile the following equations, where A_N is also defined as in the previous section:

$$(A_R - \mathbb{E}(A_R)) \cdot X \leq O(n\sqrt{a+b}) \tag{1}$$

$$A_N \cdot X \leq \sum_{u,v} (A_N)_{uv} = \sum_v (\hat{a}_v + \hat{b}_v) \text{ (using Fröbenius norm)} \tag{2}$$

$$(A_R + A_N) \cdot X \geq [n(a-b) - O(n)] + \sum_v (\hat{a}_v + \hat{b}_v) \tag{3}$$

If we compute (1) - (2) - (3), we then have:

$$\mathbb{E}(A_R) \cdot X \geq n(a-b) - O(n) - O(n\sqrt{a+b})$$

Recall from Lecture 10 (non-adversarial setting) that we could write $\mathbb{E}(A_R) = \frac{a+b}{n} \cdot J + \frac{a-b}{n} \cdot C$.

$$\begin{aligned}\frac{a-b}{n} \cdot C \cdot X &\geq n(a-b) - O(n\sqrt{a+b}) \\ C \cdot X &\geq n^2(a-b) - O(n^2\sqrt{a+b}) \\ &\geq n^2(1 - O(\frac{\sqrt{a+b}}{a-b}))\end{aligned}$$

We use the above to bound the Fröbenius norm of $C - X$, to see how "close" C is to X:

$$\begin{aligned}\|C - X\|_F^2 &= C \cdot C + X \cdot X - 2C \cdot X \\ &= 2n^2 - 2n^2(1 - O(\frac{\sqrt{a+b}}{a-b})) \leq \frac{2}{c}n^2\end{aligned}$$

We run the same argument at the very end of Lecture 10 to conclude that the unit eigenvector corresponding to the maximum eigenvalue of X is a good approximation to $\frac{1}{\sqrt{n}}\chi$, where χ is the indicator vector for the cut. □

3.3 Other semirandom models

We can also consider the usual stochastic block model with planted sets S_1 and S_2, but with a different kind of adversary. The adversary is allowed to both add and remove edges within the S_i, but is only allowed to delete edges between S_1 and S_2. While we can't hope to recover the original position, we do have the following result:

Theorem 6 *If $q \geq c \cdot \frac{\sqrt{\log n}}{n}$, we can find a balanced cut with $O(\frac{qn^2}{4})$ edges.*