MATH888: High-dimensional probability and statistics Fall 2021

Lecture 41— December 13, 2021

Sebastien Roch, UW-Madison *Scribe: Govind Gopakumar, Sebastien Roch*

1 Overview

In the last lecture, we began talking about the Stochastic Block Model (SBM) and conditions for community recovery in this model. In this lecture, we establish some results regarding almost exact recovery in the SBM. Initially we recap what was covered in the previous lecture, setup the problem, and then move on to the key results.

2 Setup

We are concerned with recovering cluster/community assignments for n vertices in a graph. The vertices are partitioned into two clusters of equal sizes $(n/2)$. Pairs of vertices within the same cluster/community have an edge between them with probability $p = q_{in}$, whereas the corresponding probability for edges between nodes that belong to different clusters is $p = q_{out}$. We observe an instance of this random graph, where nodes have edges between them with the probabilities mentioned above.

Denote by $(X, G) \sim SBM(n, q_{in}, q_{out})$ the obtained output, where X is the cluster assignment for each node, and G is the resulting graph.

3 Community recovery by spectral clustering

3.1 Exact recovery

We state below a theorem from [ABH16], that characterizes the conditions for exact recovery to be possible.

Theorem 1. *Exact recovery in the SBM model is possible if we have the following,* $q_{in} \approx \frac{a \log n}{n}$, $q_{out} \approx \frac{b \log n}{n}$ *with the relationship* $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$. *It is not possible if we have* $|\sqrt{a} - \sqrt{b}| < \sqrt{2}$.

3.2 Spectral clustering: a special case

Let us look at the spectral clustering algorithm, the following background will be necessary,

1. Add self loops to the graph, each being present with probability q_{in}
2. Let A' be the adjacency matrix that has been so modified

3. Denote by A the matrix $A = A' - \frac{q_{in}+q_{out}}{2} \begin{bmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{bmatrix} \begin{bmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{bmatrix}^T$

4. Note that $\mathbb{E}A = \frac{q_{in}-q_{out}}{2} \begin{bmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{bmatrix} \begin{bmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{bmatrix}^T$

Here we have assumed that the first $n/2$ vertices belong to the first community and the next $n/2$ belong to the second community. This is without loss of generality.

We now compute $\hat{\phi}$, the leading eigenvector of A. Note that the corresponding eigenvector of $\mathbb{E}A$ is $\bar{\hat{\phi}} = \frac{1}{\sqrt{n}} \begin{bmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{bmatrix}$. We can then estimate the community assignments by using $\hat{\phi}$, in the following manner,

$$sign(\hat{\phi}) = \begin{cases} +1 & \phi_i \geq 0 \\ -1 & \phi_i < 0 \end{cases}$$

Observation 2. *In the Assignment, note that we did not have access to the values of q_{in} and q_{out}.*

Observation 3. *This estimator does not use the fact that the communities are balanced anywhere.*

4 Almost exact recovery

We first obtain "almost exact" recovery, as established by the following theorem (Theorem 3.8 from [CCFM21]).

Theorem 4. *Under the conditions $q_{in} \gtrsim \frac{logn}{n}$ and $\sqrt{\frac{q_{in}}{n}} = o(q_{in} - q_{out})$, w.p. exceeding $1 - n^{-8}$ the spectral method achieves almost exact recovery, i.e. $\sum_{i=1}^{n} \mathbb{1}_{\{x_i = x_i^*\}} = n - o(n)$.*

This means that the number of nodes that are mis-clustered as a fraction of n tend to vanish.

We show this by an application of Davis-Kahan, as well as a result from [BH16] about the matrix norm of a matrix with bounded and centered entries.

Theorem 5. *Consider a symmetric matrix $\mathbf{X} = [X_{i,j}] \in \mathbb{R}^{n \times n}$ whose entries are independent and obey, $\mathbb{E}X_{i,j} = 0$ and $X_{i,j} \leq B$, $\forall 1 \leq i, j \leq n$, $\mathbb{E}X_{i,j}^2 \leq \sigma^2$ then w.p. we have $||\mathbf{X}|| \lesssim \sigma\sqrt{n} + B\sqrt{\log n}$.*

Note that if we consider the matrix $A - \mathbb{E}A$, it satisfies these properties, since,

- The entries are centered
- $|A_{i,j} - \mathbb{E}A_{i,j}| \leq 1$
- $\mathbb{E}[A_{i,j} - \mathbb{E}A_{i,j}]^2 \leq q_{in} \vee q_{out}$ since it is essentially a centered Bernoulli random variable.

Applying this theorem, we obtain,

$$||A - \mathbb{E}A|| \leq \sqrt{q_{in}n} + \sqrt{logn}$$
$$\approx \sqrt{q_{in}n}$$

Now, applying Davis-Kahan to show almost exact recovery,

$$\min_{s\in\{-1,+1\}} ||s\phi - \bar{\phi}||_2 \leq \frac{||A - \mathbb{E}A||}{\bar{\lambda}},$$
$$\bar{\lambda} = \frac{n(q_{in} - q_{out})}{2}$$
$$\min_{s\in\{-1,+1\}} ||s\phi - \bar{\phi}||_2 \leq \frac{2\sqrt{q_{in}n}}{n(q_{in} - q_{out})}$$
$$= o(1)$$

Where in the last step, we apply the second property stated in the theorem.

Now, we make a claim that can be obtained from this result,

Claim 6. *Let $N = \{|\phi_i - \bar{\phi}_i| > 1/\sqrt{n}$. Note that if we have a misclassification, that is, $sign(\phi_i) \neq sign(\bar{\phi}_i)$ then $i \in N$. We can see from the result that we have obtained above that $|N| = \frac{||\phi-\bar{\phi}||^2}{1/n} = o(n)$.*

5 Exact recovery: a key lemma

We will sketch the proof of the exact recovery result next time. To obtain the result, we use a key lemma regarding sums of Bernoulli random variables.

Lemma 7 (Lemma 8 in [AFWZ20])**.** *Let $\{W_i\}_{i=1}^{n/2} \sim Bern(q_{in})$ i.i.d. and similarly, $\{Z_i\}_{i=1}^{n/2} \sim Bern(q_{out})$ i.i.d. For any $\epsilon > 0$ we have,*

$$P(\sum W_i - \sum Z_i \leq \epsilon \log n) \leq n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{2} + \epsilon \log \sqrt{\frac{a}{b}}}$$

Observation 8. *Note that here the result we require is in terms of the difference of these bernoulli sums. We can think of this result as characterizing the event that any particular node has more edges to nodes that are not part of its community, compared to edges to nodes that are part of the same community. In case a particular node does have more edges to non-cluster nodes, it becomes hard for any algorithm to distinguish between both assignments of clusters to this node.*

References

[ABH16] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact Recovery in the Stochastic Block Model. *IEEE Transactions on Information Theory*, 62(1):471–487, January 2016. Conference Name: IEEE Transactions on Information Theory.

[AFWZ20] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3):1452–1474, June 2020. Publisher: Institute of Mathematical Statistics.

[BH16] Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of Probability*, 44(4):2479–2506, July 2016. Publisher: Institute of Mathematical Statistics.

[CCFM21] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral Methods for Data Science: A Statistical Perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, October 2021. Publisher: Now Publishers, Inc.