EE290 Mathematics of Data Science Lecture 7 - 09/24/2019

Lecture 8: Exact recovery of stochastic block model: Converse

Lecturer: Jiantao Jiao *Scribe: Feynman Liang*

1 Exact recovery of stochastic block model

Definition 1 (Symmetric stochastic block model)**.** *The* **symmetric stochastic block model***, denoted by* $SSBM(n, 2, p_{in} = \frac{a \log n}{n}, p_{out} = \frac{b \log n}{n} \mid \sigma)$*, is a probability distribution over graphs* (V, E) *on* n *vertices where:*

- *Each vertex* $v \in V$ *belongs to one of* 2 *communities, denoted by* $\sigma_v \in \{1, 2\}$
- ***Symmetric****: exactly* $\frac{n}{2}$ *vertices in each community*
- *The probability of an edge between two vertices in the same community is* $p_{in} = \frac{a \log n}{n}$
- *The edge probability between different communities is* p_{out}*.*

Notice that we have chosen to parameterize $p_{in} = \frac{a \log n}{n}$ and $p_{out} = \frac{b \log n}{n}$. Some intuition for the log is to recall that $G(n, c \log n/n)$ is connected whp iff $c > 1$. For SSBM, we have a similar threshold where G is connected whp iff the average of the edge probability coefficients $\frac{a+b}{2} > 1$.

We are interested in ***exact recovery in SSBM***: let $G = (V, E) \sim SSBM(n, 2, p_{in}, p_{out} \mid \sigma^*)$, can we construct an estimator $\hat{\sigma}(G)$ such that as $n \to \infty$

$$\Pr[\sigma^* \neq \hat{\sigma}] \to 0 \tag{1}$$

The goal over the next lectures will be to establish the following phase transition regarding the hardness of exact recovery in SSBM:

Theorem 2. *Exact recovery in* $SSBM(n, 2, \frac{a \log n}{n}, \frac{b \log n}{n})$ *is efficiently solvable if* $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ *and unsolvable if* $|\sqrt{a} - \sqrt{b}| < \sqrt{2}$*.*

Remark We can rewrite $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ as $\frac{a+b}{2} > 1 + \sqrt{ab}$ and compare against the $\frac{a+b}{2} > 1$ connectivity threshold for SSBM. As expected, exact recovery implies connectivity. Furthermore, exact recovery requires a $\sqrt{ab}$ over-sampling factor.

Remark For $|\sqrt{a} - \sqrt{b}| = \sqrt{2}$, exact recovery is efficiently solvable if $a, b > 0$.

1.1 Genie-aided hypothesis test

Today's lecture focuses on converse proof. Our proof will relate exact recovery to the following "genie-aided" hypothesis test:

Example 3 (Genie-aided hypothesis test)**.** Consider the one dimensional problem of genie-aided hypothesis testing problem where the genie reveals the true communities σ_v of all vertices except for one, say σ_0, and we test $\mathcal{H}_0 = \{\sigma_0 = 1\}$ against $\mathcal{H}_a = \{\sigma_0 = 2\}$. Suppose that except for node 0, there are $\frac{n}{2}$ nodes with label 1, and $\frac{n}{2}$ nodes with label 2.

The probability of error is minimized by the MAP estimator, which picks $\sigma_0 = u$ maximizing the posterior probability

$$\Pr[\sigma_0 = u \mid G = g, \sigma_{\setminus 0} = \sigma_{\setminus 0}] \tag{2}$$

Since $P(\sigma_0 = u) = 1/2$ for $u \in \{1, 2\}$, the posterior probability is

$$\Pr[\sigma_0 = u \mid G = g, \sigma_{\setminus 0} = \sigma_{\setminus 0}] = \frac{\overbrace{\Pr[\sigma_0 = u]}^{=1/2} \Pr[G = g, \sigma_{\setminus 0} = \sigma_{\setminus 0} \mid \sigma_0 = u]}{\Pr[G = g, X_{\setminus 0} = x_{\setminus 0}]} \tag{3}$$

$$\propto \Pr[G = g, \sigma_{\setminus 0} = \sigma_{\setminus 0} \mid \sigma_0 = u] \tag{4}$$

which depends only on the number of edges between vertex 0 and the two communities.

Let $T = \#\{v \in V \setminus \{0\} : \sigma_v = 1 \text{ and } (0, v) \in E\}$ count the number of edges between vertex 0 and all the vertices in community 1 (provided by the genie through $\sigma_{\setminus 0}$). Notice $T \mid \sigma_0 = 1 \sim B(\frac{n}{2}, p_{in})$ and $T \mid \sigma_0 = 2 \sim B(\frac{n}{2}, q_{out})$, so the error probability for a hypothesis test comparing the number of edges between 0 and community 1 and the community 2 is upper bounded as

$$p_e \leq P(B(\frac{n}{2}, p_{in}) \leq B(\frac{n}{2}, p_{out})) \tag{5}$$

$$= n^{-\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 + o(1)}, \tag{6}$$

where $B(\frac{n}{2}, p_{in})$ and $B(\frac{n}{2}, p_{out})$ are independent. In fact, the $\leq$ can be replaced with $=$ here.

We will spend the remainder of this lecture showing that exact recovery is not solvable if $np_e \to \infty$.

Important intuition: Let $X = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} P$ or Q, $\mathcal{H}_0$ be the hypothesis that the samples are from P, and $\mathcal{H}_1$ that they are from Q. The minimum probability of error (under an equally probable prior) is

$$\frac{1}{2}\left(1 - \mathrm{TV}(p^{\otimes n}, q^{\otimes n})\right) \tag{7}$$

We will be concerned with bounds involving a different discrepancy metric.

Definition 4 (Squared Hellinger distance)**.** *The* **squared Hellinger distance**

$$H^2(P, Q) = \mathbb{E}_Q\left[\left(1 - \sqrt{\frac{P}{Q}}\right)^2\right] \geq 0 \tag{8}$$

$$= \mathbb{E}_Q\left[1 + \frac{P}{Q} - 2\sqrt{\frac{P}{Q}}\right] \tag{9}$$

$$= 1 + 1 - 2\int \sqrt{PQ} = 2\left(1 - \int \sqrt{PQ}\right) \tag{10}$$

It sandwiches total variation distance in the following sense:

$$0 \leq \frac{1}{2}H^2(P, Q) \leq \mathrm{TV}(P, Q) \leq H(P, Q)\sqrt{1 - \frac{H^2}{4}} \leq 1 \tag{11}$$

Lemma 5. *For any sequence $\{p_n\}$, $\{q_n\}$, as $n \to \infty$*

$$\mathrm{TV}(p_n^{\otimes n}, q_n^{\otimes n}) \to 0 \iff H^2(p_n, q_n) = o(1/n) \tag{12}$$

$$\mathrm{TV}(p_n^{\otimes n}, q_n^{\otimes n}) \to 1 \iff H^2(p_n, q_n) = \omega(1/n). \tag{13}$$

Clearly, to show that $\mathrm{TV}(p_n^{\otimes n}, q_n^{\otimes n}) \to 0$, it suffices to show $\mathrm{TV}(p_n, q_n) = o(1/n)$.

1.2 Proof for converse

Without loss of generality, let $C_1 = [1 : \frac{n}{2}] = \{v : (\sigma_0)_v = 1\}$ and $C_2 = [\frac{n}{2} + 1 : n] = \{v : (\sigma_0)_v = 2\}$ where σ_0 are the true labels. Let $G \sim P_{G|\sigma}(\cdot \mid \sigma_0)$ be the SSBM graph generated from this community assignment.

Definition 6 (Bad pairs)**.** *For a community assignment $\sigma \in \{0,1\}^n$, let $\sigma[u \leftrightarrow v]$ denote σ except with the community assignments for u and v swapped.*

The **bad pairs** *of vertices are*

$$\mathcal{B}(G) = \{(u,v) : u \in C_1, v \in C_2, \Pr_{G|\sigma}[G \mid \sigma_0] \leq \Pr_{G|\sigma}[G \mid \sigma_0[u \leftrightarrow v]]\} \tag{14}$$

The reason why these pairs are bad is because if $(u,v) \in \mathcal{B}(G)$ then the MAP estimator would assign greater probability to the incorrectly swapped $\sigma_0[u \leftrightarrow v]$ labels than the true σ_0 labels, therefore:

Corollary 7. *If $\mathcal{B}(G)$ is non-empty with non-vanishing probability, then exact recovery is not possible.*

To characterize the individual bad vertices involved in bad pairs, notice that swapping vertices u and v flips the edge probabilities $p_{out} \leftrightarrow p_{in}$ for all the edges containing u and v ***except*** for the (u,v) edge (if it exists). When $p_{in} > p_{out}$, we have

$$\Pr_{G|\sigma}[G \mid \sigma_0] \leq \Pr_{G|\sigma}[G \mid \sigma_0[u \leftrightarrow v]] \iff d_+(u) + d_+(v) \leq d_-(u \setminus v) + d_-(v \setminus u) \tag{15}$$

This motivates the following definition:

Definition 8 (Bad vertices for each community)**.** *For $i \in \{1,2\}$, the* **bad vertices within community** *i are*

$$\mathcal{B}_i(G) = \{u \in C_i : d_+(u) \leq d_-(u) - 1\} \tag{16}$$

where $d_+(u) = \#\{$edges u has in its own community$\}$ and $d_-(u)$ similarly but with the other community.

Notice if $u \in \mathcal{B}_1(G)$ and $v \in \mathcal{B}_2(G)$, then

$$d_+(u) + d_+(v) \leq d_-(u) + d_-(v) - 2 \leq d_-(u \setminus v) + d_-(v \setminus u) \tag{17}$$

and therefore $(u,v) \in \mathcal{B}(G)$ and exact recovery fails.

For each fixed node u, the probability that $u \in \mathcal{B}_i(G)$ is equal to the probability that a vertex u has strictly less edges within its own community $d_+(u)$ than with the other community $d_-(u)$, which is precisely the error probability

$$p_e = n^{-\left(\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}}\right)^2 + o(1)} \tag{18}$$

Continuing the proof, the following lemma shows that with high probability there some vertex is bad, completing the proof for when exact recovery fails.

Lemma 9. $\sqrt{a} - \sqrt{b} < \sqrt{2} \Rightarrow \Pr[\exists u \in \mathcal{B}_1(G)] = 1 - o(1)$

Proof Recall the Paley-Zygmund inequality:

Theorem 10 (Paley-Zygmund Inequality)**.** *Let $X \geq 0$, $0 < \mathbb{E}X^2 < \infty$. For any $c \in [0,1]$*

$$\Pr[X > c\mathbb{E}[X]] \geq (1-c)^2 \frac{(\mathbb{E}X)^2}{\mathbb{E}X^2} \tag{19}$$

Let $B_u = \mathbb{1}(d_+(u) \leq d_-(u) - 1)$. Applying Paley-Zygmund with $c = 0$ gives

$$\Pr[\exists u \in \mathcal{B}_1(G)] = \Pr\left[\sum_{u=1}^{\frac{n}{2}} B_u > 0\right] \tag{20}$$

$$\geq \frac{(\mathbb{E}\sum_{u=1}^{\frac{n}{2}} B_u)^2}{\mathbb{E}(\sum_{u=1}^{\frac{n}{2}} B_u)^2} \tag{21}$$

$$= \frac{(\frac{n}{2}\Pr[B_1 = 1])^2}{\frac{n}{2}\Pr[B_1 = 1] + \frac{n}{2}(\frac{n}{2} - 1)\Pr[B_1 = 1, B_2 = 1]} \tag{22}$$

$$= \frac{\frac{n}{2}\Pr[B_1 = 1]}{1 + (\frac{n}{2} - 1)\Pr[B_2 = 1 | B_2 = 1]} \tag{23}$$

Sufficient conditions for the RHS to tend to 1 as $n \to \infty$ are for

- $\frac{n}{2}\Pr[B_1 = 1] = \omega(1)$, which follows from eg:genie-aided-test (eg:genie-aided-test) and the assumption $\sqrt{a} - \sqrt{b} < \sqrt{2}$.
- Approximate independence, i.e. $\frac{\Pr[B_2=1|B_1=1]}{\Pr[B_1=1]} = 1 + o(1)$

The proof for approximate independence is given on page 49 of [1]. □

References

[1] Emmanuel Abbe. Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, 2018.