EE290 Mathematics in Data Science Lecture 14 - 10/17/2019

Lecture 14: Grothendieck Inequality and its Application in SBM

Lecturer: Jiantao Jiao *Scribe: Nived Rajaraman*

We continue the discussion on correlated recovery in the symmetric stochastic block model. Recall that the last lecture concluded with a mention of the spectral clustering based algorithm of Vershynin et al. [LLV15], and in this lecture we focus on characterizing its performance. We also introduce an SDP based algorithm by Guedon et al. [GV14] and show that it achieves similar guarantees. Along the way, we introduce the Grothendieck inequality which provides a powerful SDP representation (upto a constant factor) for the $\|\cdot\|_{\infty\to 1}$ norm.

1 Notation and recap

A quick summary of the notation from previous lectures:

1. The symmetric stochastic block model $\mathsf{SSBM}(n, 2, p, q)$ considers a graph with n vertices and 2 communities specified by a labelling vector $\sigma \in \{\pm 1\}^n$, such that within-community edges occur with probability p, while edges across communities occur with probability q. In the following discussion we assume for simplicity that $p > q$. In particular, the objective is to output a labelling $\hat{\sigma} \in \{\pm 1\}^n$ such that
$$\lim \sup_{n\to\infty} \frac{\mathbb{E}\left[\min_{s\in\{\pm 1\}} \|\sigma + s\hat{\sigma}\|_1\right]}{n} < 1.$$
2. Since we are interested in the regime of correlated recovery, we restrict ourselves to the case of sparse graphs (with constant average degree) by setting $p = \frac{a}{n}$ and $q = \frac{b}{n}$.
3. A result of [MNS12] gives an information theoretic lower bound to correlated recovery in the symmetric stochastic block model. In particular, if $\tau \triangleq \frac{(a-b)^2}{2(a+b)} < 1$, then correlated recovery is impossible.
4. On the other hand, the non-backtracking spectral method proposed in [BLM15] shows that this lower bound is indeed tight, by providing an efficient algorithm that achieves correlated recovery as long as $\tau > 1$.

2 Spectral clustering

Consider a d-regular graph G having adjacency matrix A. Assume that clusters are formed in the graph.

Let us assume the existence of a matrix M such that for any vector x, $x^T M x = \sum_{(u,v)\in E}(x_u - x_v)^2$. Such a matrix must be PSD since the quadratic form is always positive. On the other hand, it is also apparent that $\mathbf{1}$ is an eigenvector of M having eigenvalue 0. In fact, it is not hard to show that M indeed must be equal to the graph Laplacian $\mathcal{L}(G)$. For d-regular graphs $L(G)$ is defined as $I - \frac{1}{d}A$. The Laplacian of a graph has several nice properties[1], including the fact that $\mathsf{Spec}(A) \in [0, 2]$.

A notion which is related to the stochastic block model is that of spectral clustering, where the goal is to find a partition of the vertices of the graph with minimum normalized cut size. That is, find a set of vertices U such that the quantity $\frac{\mathbb{E}(U,U^c)}{|U||U^c|}$ is minimized. It turns out that 2^{nd} smallest eigenvector, v_2, of $\mathcal{L}(G)$ is closely tied to the objective of finding the best spectral partition. This leads to Fiedler's algorithm which simply returns the vertices having the same sign in v_2. Cheeger's inequality precisely gives a guarantee on the cut provided by Fiedler's algorithm.

[1] Luca Trevisan's course is a good resource for spectral graph theory

Concentration of adjacency matrix of random graphs Consider a random graph having entries distributed independently as $A_{ij} \sim \mathsf{Ber}(p_{ij})$. It is known from several results in literature that $\|A - \mathbb{E}[A]\|_{\text{op}} = \mathcal{O}(\sqrt{d})$ if $d \gtrsim \sqrt{\log n}$ where $d \triangleq \max_{i,j} np_{ij}$. However, if $d \lesssim 1\text{m}$ then $\|A - \mathbb{E}[A]\| \gg \|\mathbb{E}[A]\|$. In particular this rules out the possibility of using results such as the Davis-Kahan theorem to bound the deviation of the eigenvectors of A from those corresponding to $\mathbb{E}[A]$. This is indeed a major roadblock because of the dearth of techniques to bound the distance between eigenvectors of A from those of B, when $\|A - B\|_{\text{op}}$ is large.

Feige and Ofek [FO05] informally show that dropping the vertices of large degree allows the deviation between the new matrix A' and its expectation to bounded by a constant even when $d \lesssim \log n$. In particular,

Lemma 1 ([FO05])**.** *Consider a random graph G with adjacency matrix A sampled as $A_{ij} = \mathsf{Ber}(p_{ij})$. Define $d \triangleq \max_{i,j} np_{ij}$. Let A' be the new adjacency matrix generated by deleting vertices from A having degree greater than $2d$ (that is, the rows and columns corresponding to these vertices are made $\mathbf{0}$). Then,*

$$\|A' - \mathbb{E}[A']\|_{\text{op}} \lesssim \sqrt{d}.$$

While this result claims that indeed deleting the vertices of G having high degree generates a graph which is close to its expectation in $\|\cdot\|_{\text{op}}$. However, a counter-point to this lemma is that much of the structure, including possibly the community structure, in the original graph is lost.

The scope of this result was significantly extended by [LLV15], who show that to ensure that $\|A - \mathbb{E}\|_{\text{op}} \lesssim \sqrt{d}$, it suffices to reduce the weights on the edges (unweighted edges are assumed to have weight 1) of the graph such that all weighted degrees of vertices no longer exceed $2d$. In particular,

Lemma 2 ([LLV15])**.** *Consider a random graph G with adjacency matrix A sampled as $A_{ij} = \mathsf{Ber}(p_{ij})$. Define $d \triangleq \max_{i,j} np_{ij}$. Let A'' be the new adjacency matrix generated by arbitrarily re-weighting edges of G so that each vertex i has weighted degree, $\mathsf{deg}(i) \triangleq \sum_j A_{ij}$ at most equal to $2d$. Then,*

$$\|A'' - \mathbb{E}[A]\|_{\text{op}} \lesssim \sqrt{d}.$$

Remark Lemma 1 corresponds to the particular reweighting scheme where columns and rows of the adjacency matrix are made to 0. However, note that Lemma 1 claims that A' is close to $\mathbb{E}[A']$ in operator norm, while Lemma 2 shows that A'' is close to $\mathbb{E}[A]$, the expected adjacency matrix of the original graph, and therefore Lemma 1 is not necessarily a special case of Lemma 2.

We now move to the analysis of the spectral clustering algorithm proposed by Vershynin et al. [LLV15]. Recall that the proposed algorithm is as follows,

Algorithm 1 Spectral clustering based algorithm [LLV15]

procedure LLV(G)
  Define $\tau \triangleq \frac{1}{n}\sum_{i=1}^{d} \mathsf{deg}(i)$ and $A_\tau \triangleq A + \frac{\tau}{n}\mathbf{1}\mathbf{1}^T$ ▷ regularize the adjacency matrix
  v_2 is the eigenvector of $\mathcal{L}(A_2)$ with 2^{nd} smallest eigenvalue ▷ Spectral clustering of $\mathcal{L}(A_\tau)$
  Define $\hat{\sigma} = \mathsf{sign}(v_2)$ ▷ sign function acts co-ordinate wise
  Return $\hat{\sigma}$
end procedure

Theorem 3. *Algorithm 1 outputs a labelling $\hat{\sigma}$ with misclassified proportion $\leq \epsilon$ (that is, $\frac{1}{2}\min_{s\in\{\pm 1\}} \|\hat{\sigma} + s\sigma\|_1 \leq \epsilon n$) as long as $\frac{(a-b)^2}{2(a+b)} \gtrsim \frac{1}{\epsilon^2}$.*

In order to prove this result, it suffices to show that under the conditions of the theorem, with high probability,

$$\min_{s\in\{\pm 1\}} \|v_2(\mathcal{L}(A_\tau)) + s v_2(\mathcal{L}(\mathbb{E}[A_\tau]))\|_1 \lesssim \epsilon.$$

Recall that A_τ is defined as $A + \frac{\tau}{n}\mathbf{1}\mathbf{1}^T$ in Algorithm 1. As a corollary of Theorem 2, we have that

Corollary 4. *The regularized Laplacian of A_τ with $\tau = \frac{1}{n}\sum_{i=1}^{d} \mathsf{deg}(i)$ satisfies the property that,*

$$\|\mathcal{L}(A_\tau) - \mathcal{L}(\mathbb{E}[A_\tau])\| \lesssim \frac{1}{\sqrt{d}}. \tag{1}$$

Note that this is the correct scaling to expect, since $\mathcal{L}(A_\tau) = I - D_\tau^{-\frac{1}{2}} A_\tau D_\tau^{-\frac{1}{2}}$ which scales the operator norm by a factor of $\Theta(d)$ for $\tau \approx d$.

Having observed the difference between $\mathbb{E}[A_\tau]$ as

$$\mathbb{E}[A_\tau] = \left(\frac{a+b}{2}\phi_1\phi_1^T + \frac{a-b}{2}\phi_2\phi_2^T - \frac{a}{n}I\right) + \tau\phi_1\phi_1^T.$$

where $\phi_1 = \frac{1}{\sqrt{n}}\mathbf{1}$ while $\phi_2 = \frac{1}{\sqrt{n}}(1, \cdots, 1, -1, \cdots, -1)^T$. Thus, the 2^{nd} eigenvector of $\mathbb{E}[A_\tau]$ exactly encodes the true community labelling vector upto a scale factor. It is also easy to see that $\mathbb{E}[A_\tau]$ is the sum of symmetric matrices each having constant row sums, and therefore $\mathbb{E}[A_\tau]$ corresponds to the adjacency matrix of a regular graph. In fact, the average degree of this regular graph is $\approx \frac{a+b}{2} + \tau$ (ignoring the $\frac{a}{n}I$ term which has a very small contribution). Observe that $\frac{a+b}{2}$ is equal to the expected degree of a vertex in the graph, while τ is chosen to be $\frac{1}{n}\sum_{i=1}^{d} \mathsf{deg}(i)$, the average degree of the realization of the graph which sharply concentrates around $\frac{a+b}{2}$. One would therefore expect that degree of the regular graph corresponding to $\mathbb{E}[A_\tau]$ is $\approx a + b$.

As a consequence of $\tau \approx \frac{a+b}{2}$, we also have that,

$$\frac{\mathbb{E}[A_\tau]}{a+b} \approx \phi_1\phi_1^T + \frac{a-b}{2(a+b)}\phi_2\phi_2^T + 0I.$$

So in order to be able to detect the 2^{nd} eigenvalue, a necessary condition is that its corresponding eigenvalue should be away from other eigenvalues, namely 0 and 1. In addition, $\frac{a-b}{2(a+b)}$ is closer to 0 than to 1. Note also, that for any adjacency matrix A, $\mathcal{L}(A)$ has an eigenvalue as 0 (this maps to the top eigenvalue of A). As a consequence, by Davis-Kahan theorem using (1),

$$\sin\theta \leq \frac{1}{\sqrt{\frac{a+b}{2}}}\frac{a-b}{2} = \frac{a-b}{\sqrt{2(a+b)}} \lesssim \epsilon,$$

where the last inequality follows by the condition provided in the theorem.

We next introduce Grothendieck's inequality as an aide to discuss an SDP based algorithm proposed by Guedon et al. [GV14] that achieves similar guarantees as Algorithm 1.

3 Grothendieck's inequality

Grothendieck's inequality is a powerful characterization (upto a constant factor) of the $\|\cdot\|_{\infty\to 1}$ norm using a tractable SDP formulation. Recall that the $\|\cdot\|_{\infty\to 1}$ norm of a matrix is defined as,

$$\|A\|_{\infty\to 1} = \sup_{x:\|x\|_\infty \leq 1} \|Ax\|_1. \tag{2}$$

The particular SDP under consideration is the following,

$$\mathsf{SDP}_1(A) \triangleq \sup_{\substack{u_i, v_j \in \mathbb{R}^r \\ \|u_i\|_2 = \|v_j\|_2 = 1}} \sum_{i=1}^{n}\sum_{j=1}^{m} \langle u_i, v_j\rangle A_{ij}. \tag{3}$$

which is derived by plugging in the variational representation of the $\|\cdot\|_1$ norm, $\|x\|_1 = \sup_{\|y\|_\infty \le 1} x^T y$, into (2). Note that (3) is indeed an SDP formulation, since we may write it in terms of the $r \times r$ (for $r \ge n+m$) matrix X,

$$\mathsf{SDP}_1(A) = \sup_{\substack{X \succcurlyeq 0 \\ \mathsf{diag}(X) \le I}} \left\langle \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}, X \right\rangle. \tag{4}$$

Indeed in the case $r = 1$, (3) exactly returns $\|A\|_{\infty\to 1}$,

$$\|A\|_{\infty\to 1} = \sup_{\|x\|_\infty \le 1, \|y\|_\infty \le 1} \sum_{i,j} A_{ij} x_i y_j.$$

On the other hand, Grothendieck's inequality states that even as $r \to \infty$, the solution to $\mathsf{SDP}_1(A)$ is only a constant factor away from $\|A\|_{\infty\to 1}$. That is,

$$\|A\|_{\infty\to 1} \le \mathsf{SDP}_1(A) \le K_G \|A\|_{\infty\to 1},$$

where K_G is a constant ≤ 1.78.

An application of Grothendieck's inequality for Erdös Rényi random graphs

Consider $A \sim G(n,p)$ where $p = \frac{a}{n}$ for some large constant a. Observe that it is easy to compute the $\|\cdot\|_{\infty\to 1}$ norm of $\mathbb{E}[A]$ as being equal to $n(n-1)p = (n-1)a$, since $\mathbb{E}[A]$ has all non-diagonal entries positive and equal to p.

Lemma 5 ([GV14])**.** *Consider some symmetric random matrix A with deterministic entries along the diagonal, and with the off-diagonal entries being independently and arbitrarily distributed with the only constraint being that the support of the distributions are a subset of $[0,1]$. Define the average variance $\bar{P}$ as $\frac{\sum_{i<j} \mathsf{Var}(A_{ij})}{\binom{n}{2}}$. Then, if $\bar{P} \ge \frac{9}{n}$,*

$$\|A - \mathbb{E}[A]\|_{\infty\to 1} \le 3\sqrt{p} n^{\frac{3}{2}}, \quad \textit{with probability } \ge 1 - e^3 5^{-n}.$$

Observe in the Erdös Rényi setting, $\mathsf{Var}(A_{ij}) = p(1-p) \le p = \frac{a}{n} = \bar{P}$. Therefore,

$$\|A - \mathbb{E}[A]\|_{\infty\to 1} \le 3a^{\frac{1}{2}} n.$$

For a sufficiently large constant a, this implies that

$$\underbrace{\|A - \mathbb{E}[A]\|_{\infty\to 1}}_{=3a^{\frac{1}{2}} n} \ll \underbrace{\|\mathbb{E}[A]\|_{\infty\to 1}}_{=a(n-1)}.$$

That is, while A and $\mathbb{E}[A]$ may be far separated in $\|\cdot\|_{\mathrm{op}}$ in the regime of constant average degree in Erdös Rényi graphs, the $\|\cdot\|_{\infty\to 1}$ norm serves as a witness for the closeness of A and $\mathbb{E}[A]$.

4 SDP based algorithm of [GV14]

We now describe the SDP based algorithm proposed in [GV14].

Algorithm 2 SDP based algorithm [GV14]

procedure GV(G)
  Let A denote the adjacency matrix of G.
  Let $\hat{X}$ be the optimal solution to

$$\mathsf{SDP}_2(A) = \sup_{\substack{X \succcurlyeq 0 \\ \mathsf{diag}(X)=I \\ \langle X, \mathbf{1}\mathbf{1}^T \rangle = 0}} \langle A, X \rangle$$

  Define $\hat{\sigma} = \mathsf{sign}(v_1(\hat{X}))$ ▷ v_1 denotes the top eigenvector and sign acts co-ordinate wise
  Return $\hat{\sigma}$
end procedure

Theorem 6. *Algorithm 2 outputs a labelling $\hat{\sigma}$ with misclassified proportion $\leq \epsilon$ (that is, $\frac{1}{2}\min_{s\in\{\pm 1\}} \|\hat{\sigma} + s\sigma\|_1 \leq \epsilon n$) as long as $\frac{(a-b)^2}{2(a+b)} \gtrsim \frac{1}{\epsilon^2}$.*

Equipped with Grothendieck's inequality and the concentration of the $\|\cdot\|_{\infty\to 1}$ of random adjacency matrices with entries $\in [0,1]$, we now try to analyze Algorithm 2 and in particular the following SDP,

$$\mathsf{SDP}_2(A) = \sup_{\substack{X \succcurlyeq 0 \\ \mathsf{diag}(X)=I \\ \langle X, \mathbf{1}\mathbf{1}^T \rangle = 0}} \langle A, X \rangle, \tag{5}$$

the optimal solution to which is denoted $\hat{X}$. Notice how this SDP formulation bears striking resemblance to that in (4) (with an additional balance constraint, $\langle X, \mathbf{1}\mathbf{1}^T \rangle = 0$). We use $\|\cdot\|_{\infty\to 1}$ as a proxy to show that $\hat{X}$ is not far from $X^* \triangleq \sigma\sigma^T$ where σ is the true labelling vector. In particular, observe that,

$$\|\hat{X} - X^*\|_F^2 = \|\hat{X}\|_F^2 + \|X^*\|_F^2 - 2\langle \hat{X}, X^* \rangle.$$

Observe that $\|X^*\|_F^2$ is equal to n^2, while $\|\hat{X}\|_F^2$ is upper bounded by n^2, since any PSD matrix with diagonal upper bounded by I must have all of its entries upper bounded by 1. Therefore,

$$\|\hat{X} - X^*\|_F^2 \leq 2(n^2 - \langle \hat{X}, X^* \rangle). \tag{6}$$

Observe that we can also show the following relation,

$$\langle \mathbb{E}[A], X^* \rangle - \langle \mathbb{E}[A], \hat{X} \rangle = \frac{p-q}{2}(n^2 - \langle \hat{X}, X^* \rangle).$$

Therefore, in order to upper bound the LHS of (6), it suffices to upper bound $\langle \mathbb{E}[A], X^* - \hat{X} \rangle$. However, this is easy by an application of Grothendieck's inequality.

$$\begin{aligned}
\langle \mathbb{E}[A], X^* \rangle - \langle \mathbb{E}[A], \hat{X} \rangle &\leq \langle \mathbb{E}[A], X^* \rangle - \Big(\langle A, \hat{X} \rangle - \langle \mathbb{E}[A] - A, \hat{X} \rangle \Big) \\
&\overset{(i)}{\leq} \langle \mathbb{E}[A], X^* \rangle - \Big(\langle A, \hat{X} \rangle - K_G \|A - \mathbb{E}[A]\|_{\infty\to 1} \Big)
\end{aligned}$$

where (i) uses the fact that since $\hat{X}$ is a feasible solution to (5) and hence SDP_1 in (3) its value cannot exceed the optimizer of SDP_1 which is at most $K_G\|A - \mathbb{E}[A]\|_{\infty\to 1}$. Repeating the same argument for the first term gives,

$$\begin{aligned}
\langle \mathbb{E}[A], X^* \rangle - \langle \mathbb{E}[A], \hat{X} \rangle &\leq \langle A, X^* \rangle - \langle A, \hat{X} \rangle + 2K_G \|A - \mathbb{E}[A]\|_{\infty\to 1} \\
&\leq 2K_G \|A - \mathbb{E}[A]\|_{\infty\to 1}.
\end{aligned}$$

which follows from the fact that $\hat{X}$ is the optimizer to SDP_2 in (5), to which X^* is a feasible solution. As a consequence, from (6),

$$\|\hat{X} - X^*\|_F^2 \leq \frac{8n}{a-b} K_G \|A - \mathbb{E}[A]\|_{\infty \to 1} \overset{(i)}{=} \frac{8K_G n^2}{a-b} \sqrt{\frac{a+b}{2}}, \tag{7}$$

where (i) follows from Lemma 5 (we use the upper bounds $p(1-p) \leq p$ an.

Having established this relation, we can conclude that the top eigenvector of $\hat{X}$ is close to σ. This is because the Frobenius norm upper bounds the operator norm, and we can use Davis-Kahan theorem (after scaling unit-norm eigenvectors by a factor of $\sqrt{n}$),

$$\min_s \|v_1(\hat{X}) + s\sigma\| \leq \sqrt{n} \frac{\sqrt{\frac{8K_G n^2}{a-b}\sqrt{\frac{a+b}{2}}}}{n} \lesssim \sqrt{\epsilon},$$

where the last inequality follows from the condition in the theorem.

References

[BLM15] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs, 2015.

[FO05] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

[GV14] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality, 2014.

[LLV15] Can M. Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs, 2015.

[MNS12] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction, 2012.