LECTURE NOTES ON APPLICATIONS OF GROTHENDIECK'S INEQUALITY

COMMUNITY DETECTION

JOP BRIËT

ABSTRACT. In this lecture we cover an efficient approximation algorithm for community detection in the sparse stochastic block model due to Guédon and Vershynin [GV15].

1. Community detection in the SBM

The *stochastic block model* is a simple model for inhomogeneous networks. Here, an n-element vertex set is partitioned into two $(n/2)$-sided vertex subsets $S, T \subseteq \{1, \ldots, n\}$, the communities. For some real numbers $1 \geq p > q \geq 0$, each pair of distinct vertices independently forms an edge with probability p if they belong to the same community and with probability q if they belong to different communities. For the purpose of this lecture, we shall consider the slightly non-standard model in which each loop is also included independently with probability p. Denote the resulting probability distribution over graphs by $\bar{G}(n, p, q)$. The *community detection problem* asks to recover the communities S and T when we are given a single sample from $\bar{G}(n, p, q)$.

There are many results for the dense regime, where the average degree $(p+q)n$ is of order $\Omega(\log n)$. Less is known about the sparse case, where the average degree is *constant*, $p = \Theta(1/n)$. In this lecture we will see an efficient approximation algorithm from [GV15] for the latter case.

Theorem 1.1 (Guédon and Vershynin)**.** *Let $\varepsilon \in (0, 1)$ and $n \geq 10^4/\varepsilon^2$. Furthermore assume that for $p = a/n$ and $q = b/n$, we have*

$$\max\{(p(1-p), q(1-q)\} \geq \frac{20}{n} \quad \text{and} \quad (a-b)^2 \geq \frac{10^4(a+b)}{\varepsilon^2}.$$

Then, there exists a polynomial-time algorithm for community detection that, with probability at least $1 - e^3/5^n$, misclassifies at most εn vertices.

2. The ideal situation

To gain intuition for the algorithm, we first consider a highly idealized situation. Let A be the adjacency matrix of a random graph with distribution $\bar{G}(n,p,q)$. Let $\bar{A} = \mathbb{E}[A]$ be its expectation and observe that if $S = \{1,\ldots,n/2\}$ and $T = \{n/2+1,\ldots,n\}$, then

$$\bar{A} = \begin{bmatrix} pJ_{n/2} & qJ_{n/2} \\ qJ_{n/2} & pJ_{n/2} \end{bmatrix},$$

where J_d denotes the $d \times d$ all-ones matrix.

Define the random matrix $B = A - \frac{p+q}{2}J_n$ and its expectation $\bar{B} = \mathbb{E}[B]$. Note that if S and T are again arranged consecutively, then

$$\bar{B} = \bar{A} - \frac{p+q}{2}J_n = \frac{p-q}{2}\begin{bmatrix} J_{n/2} & -J_{n/2} \\ -J_{n/2} & J_{n/2} \end{bmatrix}.$$

The idealized situation we will consider is that we posses the matrix $\bar{A}$, which implies that we also posses $\bar{B}$, and that we can efficiently find an optimizer of the following integer quadratic optimization problem:

$$\begin{aligned} \mathrm{OPT}(M) \;=\; & \text{maximize} \quad \langle M, xx^{\mathsf{T}} \rangle \\ & \text{subject to} \quad x \in \{-1,1\}^n. \end{aligned}$$

The unique optimizer to $\mathrm{OPT}(\bar{B})$ is easily seen to be $\bar{x} = 1_S - 1_T$ (up-to a sign). Hence, this optimizer correctly labels the vertices with $+1$ if they belong to S and with -1 if they belong to T.

Unfortunately, this situation is unrealistic for two reasons. First, we do not have $\bar{A}$, we only have single sample, A. Second, it is NP-hard to solve the above optimzation problem in general (and so an efficient way to find $\bar{x}$ is unlikely to exist).

3. The algorithm

To deal with the real situation in which we posses the random matrix A, and therefore B, and are content with an approximate solution, we consider instead the following alternative optimization problem.

$$\begin{aligned} &\text{maximize} && \langle B, Z\rangle \\ &\text{subject to} && Z \succeq 0 \\ &&& \operatorname{diag}(Z) = \mathbf{1}, \end{aligned} \tag{1}$$

where $\operatorname{diag}(Z)$ denotes the vector forming the diagonal of A and $\mathbf{1}$ denotes the all-ones vector. An optimal solution for this semidefinite program can be found in polynomial time.[1] The following theorem and lemma suffice to prove Theorem 1.1 (see the exercise section below).

Theorem 3.1 (Guédon–Vershynin)**.** *There is an absolute constant $C \in (0, \infty)$ such that the following holds. Let n, p, q, ε be as in Theorem 1.1. Let $\widetilde{Z}$ be an optimal solution for the semidefinite program* (1)*. Then, with probability at least $1 - 10^4/5^n$, we have*

$$\|\widetilde{Z} - \bar{x}\bar{x}^{\mathsf{T}}\|_F^2 \le C\varepsilon n,$$

where $\|X\|_F^2 = \langle X, X\rangle$ is the squared Frobenius norm and $\bar{x} = 1_S - 1_T$.

Lemma 3.2. *Let $\tilde{x}$ be an eigenvector of $\widetilde{Z}$ corresponding to its largest eigenvalue with $\|\tilde{x}\|_2^2 = n$. Then,*

$$\min_{\alpha \in \{-1,1\}} \|\alpha\tilde{x} - \bar{x}\|_2^2 \le \varepsilon n.$$

In particular, either $(\operatorname{sign}(\tilde{x}_i))_{i=1}^n$ or its negation agrees with $\bar{x}$ on at least a $(1-\varepsilon)$-fraction of the coordinates.

4. Analysis of the algorithm

Grothendieck's inequality plays a crucial part in the analysis of the algorithm. Let us recall it once more for our convenience.

Theorem 4.1 (Grothendieck's inequality)**.** *There exists an absolute constant $K_G \in (1, 2)$ such that the following holds. For any positive integer n and matrix $A \in \mathbb{R}^{n\times n}$, we have*

$$\|A\|_G \le K_G \|A\|_{\infty\to 1}. \tag{2}$$

We will work out the parts of the analysis that use the inequality, in particular in the proof of Theorem 3.1. The other parts are not difficult and can be found in [GV15]. (There, the proof of Lemma 3.2 uses a

[1] More precisely, for any $\delta > 0$, there is a $\operatorname{poly}(n, \log(1/\delta))$-time algorithm giving a feasible matrix Z such that $\langle B, Z\rangle$ is within δ of the optimum.

simple application of a theorem of Davis and Khan.) For $C \in \mathbb{R}^{n\times n}$, denote by $\mathrm{SDP}(C)$ the semidefinite program (1) with B replaced by C.

Proposition 4.2. *The semidefinite program* $\mathrm{SDP}(\bar{B})$ *has a unique optimizer given by* $\bar{Z} = \bar{x}\bar{x}^{\mathsf{T}}$.

Proposition 4.3. *There exists an absolute constant* $C > 0$ *such that with probability at least* $1 - 10^4/5^n$, *we have*

$$\|B - \bar{B}\|_{\infty\to 1} \leq C\sqrt{p+q}\, n^{3/2}.$$

The proof of Proposition 4.3 follows from a standard but careful application of the Hoeffding bound.

Now we apply Grothendieck's inequality.

Proposition 4.4. *Let* $\delta = K_G\|B - \bar{B}\|_{\infty\to 1}$. *Let* $\widetilde{Z}$ *be an optimal solution to the (random) semidefinite program* $\mathrm{SDP}(B)$. *Then,*

$$\langle \bar{B}, \widetilde{Z}\rangle \geq \langle \bar{B}, \bar{Z}\rangle - 2\delta.$$

Proof: By Grothendieck's inequality

$$\langle B, \widetilde{Z}\rangle - \langle \bar{B}, \widetilde{Z}\rangle = \langle B - \bar{B}, \widetilde{Z}\rangle \leq \|B - \bar{B}\|_G \leq \delta \tag{3}$$

$$\langle \bar{B}, \bar{Z}\rangle - \langle B, \bar{Z}\rangle = \langle \bar{B} - B, \bar{Z}\rangle \leq \|\bar{B} - B\|_G \leq \delta. \tag{4}$$

Since $\widetilde{Z}$ is optimal for $\mathrm{SDP}(B)$, we get

$$\langle \bar{B}, \widetilde{Z}\rangle \overset{(3)}{\geq} \langle B, \widetilde{Z}\rangle - \delta \geq \langle B, \bar{Z}\rangle - \delta \overset{(4)}{\geq} \langle \bar{B}, \bar{Z}\rangle - 2\delta.$$

□

Proof of Theorem 3.1: Proposition 4.2 asserts that $\bar{Z} = \bar{x}\bar{x}^{\mathsf{T}}$. Since both $\bar{Z}$ and $\widetilde{Z}$ belong to $[-1,1]^{n\times n}$,

$$\|\widetilde{Z} - \bar{Z}\|_F^2 \;=\; \|\widetilde{Z}\|_F^2 + \|\bar{Z}\|_F^2 - 2\langle \widetilde{Z}, \bar{Z}\rangle \;\leq\; 2(n^2 - \langle \widetilde{Z}, \bar{Z}\rangle).$$

Observe that $\bar{B} = \frac{p-q}{2}\bar{Z}$. Then, by Proposition 4.4,

$$\begin{aligned}\frac{p-q}{2}\langle \widetilde{Z}, \bar{Z}\rangle &= \langle \bar{B}, \widetilde{Z}\rangle \\ &\geq \langle \bar{B}, \bar{Z}\rangle - 2\delta \\ &= \frac{p-q}{2}n^2 - 2\delta.\end{aligned}$$

Rearranging gives $\|\widetilde{Z} - \bar{Z}\|_F^2 \leq 2\big(n^2 - n^2 + 2\delta\big) = 4\delta/(p-q)$. By Proposition 4.3 and our assumptions on p and q, it follows δ is sufficiently small with the desired probability. The result now follows from. □

5. Exercises

Exercise 5.1. Let $\tilde{x}$ and $\alpha \in \{-1, 1\}^n$ be as in Lemma 3.2. Show that the sign vector $y = \text{sign}(\alpha\tilde{x})$ differs from $\bar{x} = 1_S - 1_T$ in at most εn coordinates. In particular, conclude with Theorem 1.1.

Exercise 5.2. In the proof of Theorem 3.1 we claimed that both $\bar{Z}$ and $\widetilde{Z}$ belong to $[-1, 1]^{n\times n}$. Why is this true?

Exercise 5.3. Prove Proposition 4.2.

References

[GV15] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendiecks inequality. *Probability Theory and Related Fields*, pages 1–25, 2015.