CS 598: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Sameera Somisetty

Lecture #

1 Dense Graphs

1.1 Introduction

In the previous lecture, we talked about Stochastic Block Models (SBM) where Graph $G = (V, E)$ with $\mid V \mid = n$ is generated by a random process, where for every (i, j) in G, probability of (i, j) belonging to the same subgraph=p and probability of (i, j) belonging to different subgraphs=q. The adjacency matrix of graph G can be written as:

$$A \sim SBM(p, q, n)$$

We introduced the notion of an *ideal matrix* M where $M = E[A]$ such that A can be though of as a *perturbation* of M. In all our discussions we will consider two subgraphs S_1 and S_2 of G. So M can be written as:

$$M = E[A] = \begin{bmatrix} P & Q \\ Q & P \end{bmatrix}$$

where P is a $n/2 \times n/2$ matrix with all values=p, Q is a $n/2 \times n/2$ matrix with all values=q.

$$M = \frac{n}{2} \times (p + q) \times v_1 {v_1}^T + \frac{n}{2} \times (p - q) \times v_2 {v_2}^T$$

where

$$v_1 = (\frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} ... \frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}})$$

$$v_2 = (\frac{1}{\sqrt{n}} ... \frac{1}{\sqrt{n}} \frac{-1}{\sqrt{n}} ... \frac{-1}{\sqrt{n}})$$

So we can think of Matrix A as the sum of Matrix M and a noise matrix R.

$$A = M + R$$

For all R_{ij} in R:

$$E[R_{ij}] = 0$$

We also saw that,

Theorem 1. *Let $R_{n \times n}$ be a random matrix such that $E[R_{ij}] = 0$ and $\mid R_{ij} \mid \leq 1$ and $max(Var(R_{ij})) = \sigma^2$. Then, if $\sigma^2 \geq \frac{c \log^4 n}{n}$ with high probability*

$$||R|| \leq 3\sigma\sqrt{n}$$

In our case:

- When $i, j \in S_1$ OR $i, j \in S_2$

$$Var(R_{ij}) = p(1-p)^2 + (1-p)p^2 = p(1-p)$$

- When ($i \in S_1$ and $j \in S_2$) OR ($i \in S_2$ and $j \in S_1$)

$$Var(R_{ij}) = q(1-q)^2 + (1-q)q^2 = q(1-q)$$

Applying Theorem 1 on the above results, we can say, with high probability

$$||R|| \leq 3 \times \sqrt{p(1-p)}\sqrt{n} \leq 3\sqrt{n}p$$

In other words, the Spectral Norm of the matrix is quite small.

1.2 Effect of Perturbation on Singular Vectors

We have seen above that the spectral norm of the noise matrix is quite small. We will now prove that Singular Values of Matrix M and Matrix A are close to each other.

Theorem 2. ***Davis-Kahan Theorem:***
Let $M_{n\times n}, \tilde{M}_{n\times n}$ be two matrices such that $||M - \tilde{M}||_2 \leq \delta$. Let $v_1.....v_n$ be the singular vectors of M and $w_1.....w_n$ be the singular vectors of $\tilde{M}$.
Then, $\forall i \in [n]$

$$||v_i - (\pm w_i)||_2 \leq \frac{2\delta}{\min_{j\neq i}|\sigma_j - \sigma_i|}$$

In the above theorem $\tilde{M}$ is a perturbation of M. In our case, $\tilde{M} = A$. It basically says that, if the singular values are well separated then the singular vectors are closer to each other. In general, in Machine Learning singular values are well separated.

In our case, we are interested in the effect of perturbation on the second singular vector v_2 of M. Let the second singular vector of A be $\hat{v_2}$. Applying Davis-Kahan Theorem:

$$||v_2 - (\pm\hat{v}_2)||_2 \leq \frac{2.3\sqrt{n}p}{\min(|\sigma_1 - \sigma_2|, |\sigma_2 - \sigma_3|)}$$

$$\implies ||v_2 - (\pm\hat{v}_2)||_2 \leq \frac{2.3\sqrt{n}p}{\min(|nq|, |\frac{n}{2}(p-q)|)}$$

$$\implies ||v_2 - (\pm\hat{v}_2)||_2 \leq \frac{6\sqrt{n}p}{\frac{n}{2}(c\sqrt{\frac{p\log n}{n}})}$$

$$\implies ||v_2 - (\pm\hat{v}_2)||_2 \leq O(\frac{1}{\sqrt{\log n}})$$

If $\hat{v_2}$ makes k mistakes, we know that, $||v_2 - (\pm\hat{v}_2)||_2 \geq \frac{k}{n}$. Using the above inequality, we can show that $k \leq O(\frac{n}{\sqrt{\log n}})$

2 Sparse Graphs

2.1 Basic Definition:

$$A \sim SBM(p, q, n)$$

where $p = a/n$ and $q = b/n$ where a,b are constants.

2.2 Characteristics of Sparse Graphs:

- Subgraphs S_1 and S_2 are not well defined, they can have disconnected components. Hence we can only hope to recover S_1 and S_2 upto some error.
- Average degree of a vertex=$(a + b)/2$ is a constant
- Sparse graphs contain hubs, defined as, nodes of degree $> \frac{logn}{loglogn}$. For example connections on twitter is a sparse graph where some celebrity profiles having a high number of followers can be thought of as hubs.

2.3 Spectral Clustering

If we naively apply spectral clustering on a sparse graph containing hubs, then the clustering algortithm wil tend to cluster all the hubs in one subgraph and other points in the other subgraph. A simple solution to this problem is to restrict the maximum degree of the graph, or in other words remove the hubs. In practice this makes sense, for example, if we want to find the political affiliation of people on Facebook, then the follow pattern of celebrities may not be very important and can be ignored.

2.3.1 Regularized Spectral Clustering

Algorithm 1:

1. Let A be the adjacency matrix of G
2. For any vertex i of degree $> 4(a + b)$ set $A_i = 0$,$A^(i) = 0$. That is, delete the vertex.
3. Run Spectral Clustering on $\hat{A}$, the regularized matrix.

Theorem 3. *If $(a - b) > \frac{c}{\epsilon^2}\sqrt{a}$ (a,b are well separated) then with probabilty $= 1 - \frac{1}{n^3}$ regularized spectral clustering outputs $S_1{}'$ and S_2' such that,*

$$|S_1 \Delta S_1{}'| + |S_2 \Delta S_2{}'| \leq \epsilon n$$

For the analysis of Algorithm 1, it is key to argue that spectral norm of the regularised matrix is small. Note that in the regularization process, we are taking a set of points generated by a random process and creating dependencies in them by deleting points, hence the analysis of spectral norm of regularized sparse matrices is not trivially similar to that of dense matrices.

Theorem 4. *For M=E[A], $\hat{A}$ = regularised A and spectral norm of noise matrix $||R|| = ||M - \hat{A}||$, with high probability*

$$||R|| \leq O(\sqrt{a})$$

To prove Theorem 4, we will use the following Lemma:

Lemma 1. ***Decomposition Lemma*** *Let $A_{n\times n}$ be a random matrix with entries in $\{0,1\}$ such that $E[A_{ij}] = P_{ij}$, $d = n\max_{i,j} P_{ij}$ where d is the expected degree of a vertex. Choose any set of $\frac{10n}{d}$ vertices and reduce the weights of corresponding rows and columns arbitrarily to create $\hat{A}$. Then with probability $\geq 1 - \frac{1}{n^3}$*

$$||\hat{A} - E[A]|| \leq O(\sqrt{d} + \sqrt{d'})$$

where $\hat{A}$ is the new matrix, d' is the l_1-norm of any row/column.