EE 381V: Large Scale Learning	Spring 2013
Lecture 7 — February 5	
Lecturer: Caramanis & Sanghavi	*Scribe: Dohyung Park*

7.1 Topics covereed

- The planted model for spectral clustering

7.2 Perturbation approach

In the previous lecture, we proved the $\sin\theta$ theorem. Applying the theorem, we can find a performance guarantee for the spectral clustering.

Let us first recap the $\sin\theta$ theorem. The distance $d_p(E_0, F_0)$ between two subspaces spanned by the columns of E_0 and F_0, respectively, is defined as

$$d_p(E_0, F_0) \triangleq \|E_0E_0^* - F_0F_0^*\|_2 = \|\sin\Theta\|_2 \tag{7.1}$$

where Θ is a diagonal matrix with principal angles.

Theorem 7.1 (The $\sin\theta$ theorem). *Consider matrices $M, \Delta \in \mathbb{S}_n$ where*

$$M = [E_0|E_1]\begin{bmatrix} \mathrm{diag}(M_0) & 0 \\ 0 & \mathrm{diag}(M_1) \end{bmatrix}[E_0|E_1]^*,$$
$$M + \Delta = [F_0|F_1]\begin{bmatrix} \mathrm{diag}(\hat{M}_0) & 0 \\ 0 & \mathrm{diag}(\hat{M}_1) \end{bmatrix}[F_0|F_1]^*$$

are the eigenvalue decompositions of the matrices. If $M_0 \subseteq [a, b]$, $\hat{M}_1 \subseteq (-\infty, a-\delta) \cup (b+\delta, \infty)$, then

$$d_p(E_0, F_0) \le \frac{1}{\delta}\|\Delta\|_2. \tag{7.2}$$

The $\sin\theta$ theorem bounds the distance between the column spaces of E_0 and F_0. In spectral clustering, once we take the k eigenvectors with the k smallest eigenvalues, we cluster n rows of the matrix whose columns are the k eigenvectors. Therefore, the performance of the spectral clustering must be measured as the gap between the n rows obtained from the perturbed Laplacian, $\hat{L}_n$, and the n rows from the unperturbed Laplacian, L_n, up to rotation. In other words, let Y and $\hat{Y}$ denote the matrices with the first k eigenvectors of L and $\hat{L}$, respectively. They are described as

$$Y = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix} = \begin{bmatrix} - & y_1 & - \\ & \vdots & \\ - & y_n & - \end{bmatrix}, \ \hat{Y} = \begin{bmatrix} | & & | \\ \hat{u}_1 & \cdots & \hat{u}_k \\ | & & | \end{bmatrix} Q = \begin{bmatrix} - & \hat{y}_1 & - \\ & \vdots & \\ - & \hat{y}_n & - \end{bmatrix} Q, \tag{7.3}$$

where $u_1, \dots, u_k$ are the first k eigenvectors of L, and $\hat{u}_1, \dots, \hat{u}_k$ are the first k eigenvectors of $\hat{L}$. Q is a $k \times k$ unitary matrix. The performance of spectral clustering gets better as $\hat{y}_1, \dots, \hat{y}_n$ are closer to $y_1, \dots, y_n$, respectively. Hence we need to measure $\frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2$. Since we have

$$\frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 \le \frac{1}{n}\|Y - \hat{Y}\|_F^2 \le \frac{k}{n}\|Y - \hat{Y}\|_2^2, \tag{7.4}$$

we need to bound $\|Y - \hat{Y}\|_2^2$. To do so, we define another measure of distance between two subspaces.

Definition 7.2.

$$\begin{aligned} d_c(E_0, F_0) &\triangleq \min_{Q,R \in O(k)} \|E_0 Q - F_0 R\|_2 \\ &= \min_{R \in O(k)} \|E_0 - F_0 R\|_2 \end{aligned} \tag{7.5}$$

Before we consider the main theorem, we check two useful lemmas.

Lemma 7.3.

$$d_p(E_0, F_0) \le d_c(E_0, F_0) \le \sqrt{2} d_p(E_0, F_0) \tag{7.6}$$

Proof: (Proof) □

Lemma 7.4. *Let λ_{k+1} and $\hat{\lambda}_{k+1}$ be the $(k+1)$-th smallest eigenvalue of matrices L and $\hat{L}$, respectively. Then*

$$\hat{\lambda}_{k+1} \ge \lambda_{k+1} - \|\hat{L} - L\|_2. \tag{7.7}$$

Proof: Let $u_1, \dots, u_k$ be the first k eigenvectors (with the k smallest eigenvalues) of L. Then it follows that

$$\begin{aligned} \hat{\lambda}_{k+1} &= \max_{V:\dim(V)=k} \min_{x:x\in V^{\perp}, \|x\|=1} \langle x, \hat{L}x \rangle \\ &\ge \min_{x:x\in \mathrm{span}\{u_1,\dots,u_k\}^{\perp}, \|x\|=1} \langle x, \hat{L}x \rangle \\ &= \min_{x:x\in \mathrm{span}\{u_1,\dots,u_k\}^{\perp}, \|x\|=1} \left\{ \langle x, Lx \rangle - \langle x, (\hat{L} - L)x \rangle \right\} \\ &\ge \min_{x:x\in \mathrm{span}\{u_1,\dots,u_k\}^{\perp}, \|x\|=1} \langle x, Lx \rangle - \max_{\|x\|=1} \langle x, (\hat{L} - L)x \rangle \\ &\ge \min_{x:x\in \mathrm{span}\{u_1,\dots,u_k\}^{\perp}, \|x\|=1} \langle x, Lx \rangle - \max_{\|x\|=\|y\|=1} \langle y, (\hat{L} - L)x \rangle \\ &= \lambda_{k+1} - \|\hat{L} - L\|_2. \end{aligned}$$

□

An interpretation of Lemma 7.4 is the following: If we add $\hat{L} - L$ to L, λ_{k+1} will change. It is maximally reduced when the kth eigenvector of L is perfectly aligned to the eigenvector with the smallest (negative) eigenvalue of $\hat{L} - L$. Since the value is greater than $-\|\hat{L} - L\|_2$, λ_{k+1} cannot be reduced more than $\|\hat{L} - L\|_2$. There is a chance that another eigenvalue of L smaller than λ_{k+1} will become the $(k+1)$-th smallest eigenvalue of $\hat{L}$, but it doesn't matter because the value will be greater than $\lambda_{k+1} - \|\hat{L} - L\|_2$.

Using the above lemmas, we obtain the main theorem.

Theorem 7.5. *There exists a unitary matrix $Q \in O(k)$ such that if*

$$Y = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix}, \ \hat{Y} = \begin{bmatrix} | & & | \\ \hat{u}_1 & \cdots & \hat{u}_k \\ | & & | \end{bmatrix} Q, \tag{7.8}$$

then $d_c(E_0, F_0) = \|Y - \hat{Y}\|_2$, and

$$\frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 \le \frac{2k}{n(\lambda_{k+1} - \lambda_k - \|\hat{L}_n - L_n\|)^2}\|\hat{L}_n - L_n\|_2^2. \tag{7.9}$$

Proof:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 &= \frac{1}{n}\|Y - \hat{Y}\|_F^2 \\
&\le \frac{k}{n}\|Y - \hat{Y}\|_2^2 \\
&= \frac{k}{n} d_c(Y, \hat{Y})^2 \\
&\le \frac{2k}{n} d_p(Y, \hat{Y})^2 \\
&\le \frac{2k}{n(\hat{\lambda}_{k+1} - \lambda_k)^2}\|L_n - \hat{L}_n\|_2^2 \\
&\le \frac{2k}{n(\lambda_{k+1} - \lambda_k - \|L_n - \hat{L}_n\|)^2}\|L_n - \hat{L}_n\|_2^2
\end{aligned}$$

- The first equality holds by the definition of Frobenius norm.
- The second inequality holds because $\|X\|_F \le \sqrt{k}\|X\|_2$ for any matrix X with rank k.
- The third equality holds by the definition of the distance measure d_c.
- The fourth inequality holds by Lemma 7.3.
- The fifth inequality follows from Theorem 7.1.
- The last inequality follows from Lemma 7.4.

□

In the next section, we apply this theorem to the planted model.

7.3 The planted model

Consider a graph with k clusters. There is an edge with probability p between a pair of vertices in the same cluster, while vertices in different cluster are connected with probability q. It is natural that we should have $p > q$ to correctly split the vertices into k clusters. The main question is that: *How big must the gap $p - q$ be?*

Let us build a mathematical model before we consider the problem. The matrices P^{un} and P are defined as

$$P^{\text{un}}_{ij} = \begin{cases} p & \text{if vertices } i \text{ and } j \text{ are in the same cluster,} \\ 0 & \text{if vertices } i \text{ and } j \text{ are in different clusters,} \end{cases}$$

$$P_{ij} = \begin{cases} p & \text{if vertices } i \text{ and } j \text{ are in the same cluster,} \\ q & \text{if vertices } i \text{ and } j \text{ are in different clusters.} \end{cases}$$

Then an adjacency matrix A based on P is generated as

$$A_{ij} = \begin{cases} 1 & \text{with probability } P_{ij} & \text{if } i \le j, \\ 0 & \text{with probability } 1 - P_{ij} & \text{if } i \le j, \\ A_{ji} & \text{with probability } P_{ij} & \text{if } i > j. \end{cases}$$

Once we have an adjacency matrix A, we do the spectral clustering.

Note that P and A can be thought of as perturbations of P^{un} and P, respectively. Therefore, we apply Theorem 7.5 for the following two cases.

- A deterministic model : $L_n = I - D^{-\frac{1}{2}} P^{\text{un}} D^{-\frac{1}{2}}, \hat{L}_n = I - D^{-\frac{1}{2}} P D^{-\frac{1}{2}}$
- The planted model : $L_n = I - D^{-\frac{1}{2}} P D^{-\frac{1}{2}}, \hat{L}_n = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$

In the following sections, we will find a lower bound on $p - q$ for exact partitioning by spectral clustering. The planted model is what we are interested in, but we first consider the deterministic model.

7.3.1 Spectral clustering for the deterministic model

Let $P = U \Lambda U^{-1}$ be the eigenvalue decomposition of P. Since $D = (qn + (p-q)n/k)I$ is a multiple of identity, we have that

$$\begin{aligned} \hat{L}_n &= I - D^{-\frac{1}{2}} P D^{-\frac{1}{2}} \\ &= UU^{-1} - \frac{1}{\sqrt{\gamma}} I \cdot U \Lambda U^{-1} \cdot \frac{1}{\sqrt{\gamma}} I \\ &= U \left(I - \frac{1}{\gamma} \Lambda \right) U^{-1} \end{aligned} \tag{7.10}$$

where $\gamma = qn + (p-q)n/k$. Note that $I - \frac{1}{\gamma}\Lambda$ is diagonal. This means that (7.10) is the eigenvalue decomposition of $\hat{L}_n$. Therefore, the eigenvectors corresponding to the k smallest eigenvalues of $\hat{L}_n$ are equal to the eigenvectors with the k largest eigenvalues of P. This leads to the following fact.

Proposition 7.6. *Clustering according to the bottom k eigenvectors of $\hat{L}_n$ is equivalent to clustering by the top k eigenvectors of P.*

Can we also consider the bound in Theorem 7.5 in terms of P and P^{un}? The following proposition is the key property.

Proposition 7.7. *The bound* (7.9) *is invariant to scaling and translation by addition of a multiple of identity.*

Proof: (Proof) □

As D is a multiple of identity, P is obtained by scaling $\hat{L}_n$ and adding a multiple of identity to it. It follows that the gap between two eigenvalues of $\hat{L}_n$ are scaled by the absolute number of the scaling factor. Therefore, the bound (7.9) is written as

$$\frac{1}{n}\sum_{i=1}^{n}\|y_i-\hat{y}_i\|_2^2 \le \frac{2k}{n(\lambda_{n-k+1}(P)-\lambda_{n-k}(P)-\|P-P^{\text{un}}\|_2)^2}\|P-P^{\text{un}}\|_2^2. \tag{7.11}$$

Since the eigenvalues of P are given by

$$\begin{aligned}\lambda_n(P) &= qn+(p-q)\frac{n}{k},\\ \lambda_{n-1}(P)=\cdots=\lambda_{n-k+1}(P) &= (p-q)\frac{n}{k},\\ \lambda_{n-k}(P)=\cdots=\lambda_1(P) &= 0,\end{aligned}$$

the bound is written as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\hat{y}_i-e_{\text{cluster}(i)}\|^2 &\le \frac{2k}{n(qn+(p-q)n/k-\|P-P^{\text{un}}\|_2)^2}\|P-P^{\text{un}}\|_2^2\\ &= \frac{2k}{n(qn+(p-q)n/k-qn)^2}(qn)^2\\ &= \frac{2k^3q^2}{n(p-q)^2}\end{aligned} \tag{7.12}$$

which means that the spectral clustering works while

$$q \le p\left\{1-\frac{ck^{3/2}}{n^{1/2}}\right\}. \tag{7.13}$$

7.3.2 Spectral clustering for the planted model

As mentioned in the previous case, we can consider the k largest eigenvalues and their corresponding eigenvectors of A instead of the k smallest eigenvalues and eigenvectors of L_n, when D for A is a multiple of identity. Here, we also look at A because each diagonal entry of D is given by the sum of independent random variables, $d_i = \sum_{j=1}^n A_{ij}$, and for sufficiently large n, it is close to a constant $\sum_{j=1}^n P_{ij} = \frac{n}{k}(p-q) + nq$.

The planted model can be rewritten as

$$A = P + X,$$

where

$$X_{ij} = \begin{cases} 1 - P_{ij} & \text{with probability } P_{ij}, \\ -P_{ij} & \text{otherwise.} \end{cases}$$

Note that $E[X] = 0$. The bound (7.9) is then given by

$$\frac{1}{n}\sum_{i=1}^n \|y_i - \hat{y}_i\|_2^2 \leq \frac{2k}{n(\lambda_{n-k+1}(P) - \lambda_{n-k}(P) - \|X\|_2)^2}\|X\|_2^2 \tag{7.14}$$

Since the eigenvalues of P are given by

$$\begin{aligned} \lambda_n(P) &= qn + (p-q)\frac{n}{k}, \\ \lambda_{n-1}(P) &= \cdots = \lambda_{n-k+1}(P) = (p-q)\frac{n}{k}, \\ \lambda_{n-k}(P) &= \cdots = \lambda_1(P) = 0, \end{aligned}$$

the bound is written as

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \|y_i - \hat{y}_i\|_2^2 &\leq \frac{2k}{n((p-q)(n/k) - \|X\|_2)^2}\|X\|_2^2 \\ &= \frac{2k}{n(n\epsilon/k - \|X\|_2)^2}\|X\|_2^2 \end{aligned} \tag{7.15}$$

where $\epsilon = p - q$. Since $\|X\|_2$ is still a random variable, we need an expression to bound it with high probability.

One reasonable try can be to use Chebyshev's inequality: For any random variable Z with mean μ and variance σ^2,

$$P(|Z - \mu| \geq \alpha\sigma) \leq \frac{1}{\alpha^2} \tag{7.16}$$

for any number $\alpha > 0$. Since we have that

$$\sigma^2 = E[\|X\|_2^2] \leq E[\|X\|_F^2] = \sum_{i,j} E[X_{ij}^2] \leq n^2, \tag{7.17}$$

it follows that

$$P(\|X\|_2 \geq 10n) \leq P(\|X\|_2 \geq 10\sigma^2) \leq \frac{1}{100}. \tag{7.18}$$

Putting the bound $\|X\|_2 \leq 10n$ to Theorem 7.5, we get

$$\frac{1}{n}\sum_{i=1}^{n}\|y_i-\hat{y}_i\|_2^2 \leq \frac{2k}{n(n\epsilon/k-\|X\|_2)^2}\|X\|_2^2 \leq \frac{2k}{n(n(\epsilon/k-10))^2}(10n)^2 \leq \frac{200k}{n(\epsilon/k-10)^2}.$$

This bound doesn't give any useful result for ϵ. In the above derivation, (7.17) and (7.18) hold even if the entries of X are correlated. To obtain a useful bound, *we must exploit independence of i and j.*

Another try is to use the Matrix Bernstein inequality. [1]

Theorem 7.8 (Matrix Bernstein inequality). *Let $Z_1,\ldots,Z_m \in \mathbb{S}^n$ be independent random matrices where $E[Z_i]=0$, $\|Z_i\|_2 \leq R$, and $\|\sum_{i=1}^m E[Z_i^2]\| \leq \sigma^2$, and let $X=\sum_{i=1}^m Z_i$. Then we have that*

$$P\left(\|X\|_2 > t\right) \leq n\exp\left(-\frac{t^2}{6(Rt+\sigma^2)}\right). \tag{7.19}$$

We can apply Theorem 7.8 to $X = P - A$ by defining

$$Z_{(ij)} \triangleq X_{ij}(e_ie_j^* + e_je_i^*)$$

where the superscript * denotes the transpose. Note that all the entries of $Z_{(ij)}$ are zero except for the entries at (i,j) and (j,i) that are equal to $X_{ij} = X_{ji}$. Then X can be described as the sum of n^2 matrices

$$X = \sum_{i,j} Z_{(ij)}.$$

The random matrices $Z_{(ij)}$ have the following properties.

- $E[Z_{(ij)}] = 0, \|Z_{(ij)}\|_2 \leq 2.$
- $Z_{(ij)}^2 = X_{ij}(e_ie_j^*+e_je_i^*)^* \cdot X_{ij}(e_ie_j^*+e_je_i^*) = X_{ij}^2(e_ie_i^*+e_je_j^*).$
- $$\sum_{i,j:i\leq j} E[Z_{(ij)}^2] = \begin{pmatrix} \sum_{j=1}^n X_{1j}^2 & & \\ & \ddots & \\ & & \sum_{j=1}^n X_{nj}^2 \end{pmatrix}, \quad \sigma^2 = \left\|\sum_{i\leq j} E[Z_{(ij)}^2]\right\|_2 \leq n.$$

Using Theorem 7.8 and the above properties, we obtain that

$$\begin{aligned} P\left(\|X\|_2 > t\right) &\leq n \exp\left(-\frac{t^2}{6(Rt + \sigma^2)}\right) \\ &\leq n \exp\left(-\frac{t^2}{6(2t + n)}\right). \end{aligned} \tag{7.20}$$

Consider $t = 10\sqrt{n \log n}$. Then we have that

$$\begin{aligned} P\left(\|X\|_2 > 10\sqrt{n \log n}\right) &\leq n \exp\left(-\frac{100n \log n}{6(n + 20\sqrt{n \log n})}\right) \\ &\leq n \exp\left(-\frac{100n \log n}{10n}\right) \\ &\leq n \exp\left(-10 \log n\right) = n^{-9}. \end{aligned} \tag{7.21}$$

This implies that for sufficiently large n, we have that

$$\|X\|_2 \leq 10\sqrt{n \log n} \tag{7.22}$$

with high probability. Let us drop the $\log n$ factor just to make bounds look clean. Then we obtain that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 &\leq \frac{2k}{n(n\epsilon/k - \|X\|_2)^2}\|X\|_2^2 \\ &\leq \frac{200k}{(n\epsilon/k - 10\sqrt{n})^2} \\ &\leq \frac{200k}{\gamma^2 n} \end{aligned}$$

where $\epsilon \geq \frac{10k}{\sqrt{n}} + \gamma$. This concludes that we need

$$p - q = \epsilon \geq \frac{10k}{\sqrt{n}} \tag{7.23}$$

for the planted model to be correctly partitioned using spectral clustering for sufficiently large n.

Remark 7.9. *If $k = O(1)$ as $n \to \infty$, then we need $\epsilon = O(1/\sqrt{n})$. This implies that if $n >> k$, the spectral clustering can correctly partition the planted model even with a very small gap $\epsilon = p - q$.*

Bibliography

[1] Tropp, J. (2010). User-friendly tail bounds for sums of random matrices. Foundations of Computational Mathematics, 12(4), 389–434.