Lecture Notes on High Dimensional Data Analysis

Jiaming Xu [1]

Aug 15, 2018

[1]Disclaimer: This is a processed (in progress) version of the individually scribed lectures most of which have not been cleaned up. Apologies for many mistakes.

Contents

Chapter 1

Introduction

Outline

- Introduction of the topics covered
- Curse of the high-dimension
- Blessings of the high-dimension

1.1 Topics will be covered

1.1.1 Data Clustering

Definition 1.1 (Clustering)**.** Given data points $x_1, \ldots, x_n \in \mathcal{X}$, partition them into k groups/clusters. See Fig. 1.1 for an illustration.

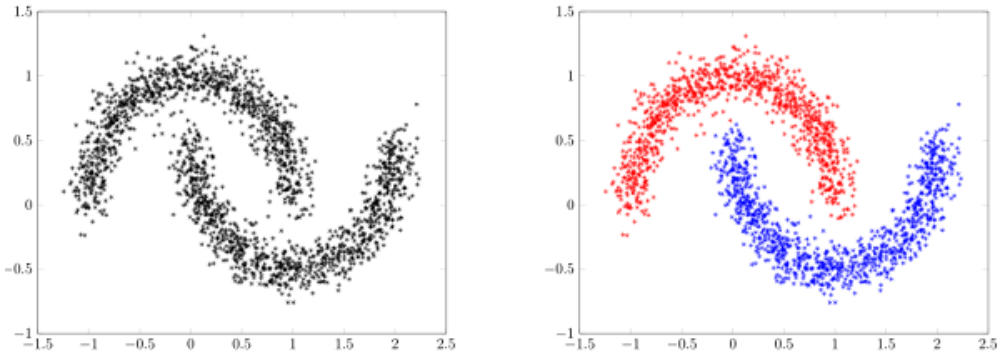


Figure 1.1: An illustration of clustering.

Note:

- $\mathcal{X}$: Input data space, e.g., Euclidean space $\mathbb{R}^d$.
- k: the number of clusters could be known a priori or need to be properly chosen.

Example 1.1 (Cluster species)**.** Cluster n species into two clusters: plants or animals.

- Features of species, e.g., color (green or brown), movement (static or could move);

- d: the total number of features available;
- $x_i \in \mathbb{R}^d$: the feature vector of specie i.
- $k = 2$.

Example 1.2 (Cluster customers)**.** Cluster n customers into k clusters based on the their preferences.

- d: the total number of products;
- $x_i \in \mathbb{R}^d$ denote the preference of customer i, where $x_{ij} = 1$ if i likes product j; $x_{ij} = -1$ if i dislikes product j; $x_{ij} = 0$ if i has not purchased product j yet.
- k is unknown a priori.

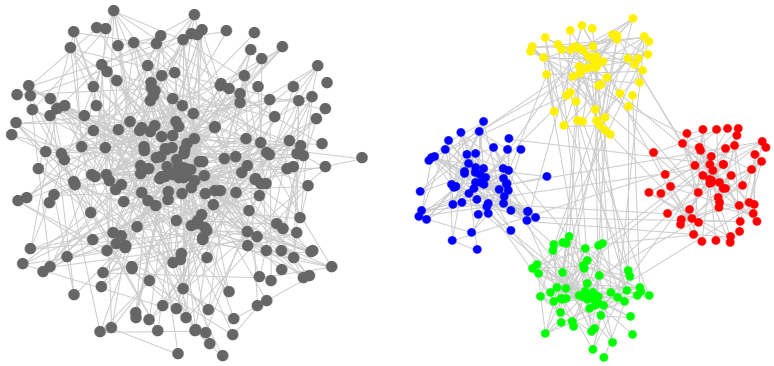
Note: In Example 1.1, n could be large but d is relatively small, while in Example 1.2, both n and d could be extremely large.

Question: How shall we cluster high-dimensional data?

1.1.2 Community Detection

Definition 1.2 (Community Detection/Graph Clustering)**.** Given a (weighted) graph $G = (V, E)$, partition n vertices into k communties/clusters.

Note: Assortative community structure: communities are more densely connected internally than externally (See Fig. 1.2); Disassortative community structure: the other way around.



(a) (b)

Figure 1.2: A synthetic network. Panel (a): Nodes are placed arbitrarily with community structure hidden. Panel (b): Nodes are grouped with community structure revealed.

Example 1.3 (Political blog network)**.** The network formed by web blogs, where two blogs are connected if there exist hyperlinks between them. See Fig. 1.3. for an illustration.

Example 1.4 (Facebook ego network)**.** The friendship network formed by a person's friends. All the friends may form communities like family members, colleagues, college friends, and so on. One would like to find the community partition of her/his friends. See Fig. 1.4 for an illustration.

Example 1.5 (Amazon product co-purchasing network)**.** The network formed by Amazon products, and two products are connected by an edge if they have been bought together for at least once.

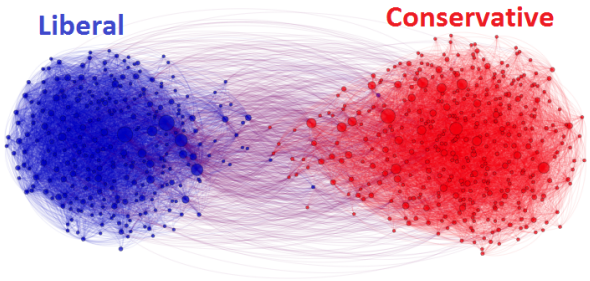


Figure 1.3: An illustration of a friendship ego network

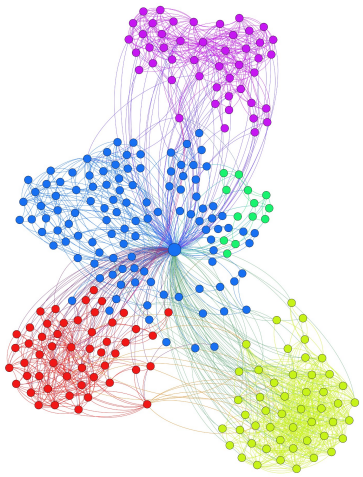
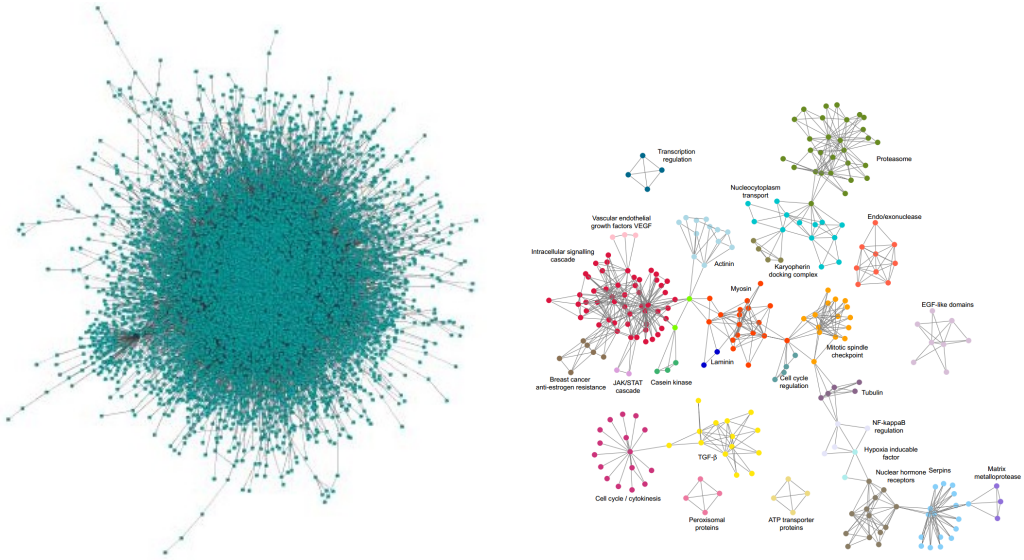


Figure 1.4: An illustration of a friendship ego network

Example 1.6 (Protein-protein interaction network)**.** The network formed by proteins, and two proteins are connected by an edge if they have been interacted with each other in a certain biological process. See Fig. 1.5.



(a) (b)

Figure 1.5: A synthetic network. Panel (a): Protein interaction network. Panel (b): Protein interaction network with functional community structure.

Question: How to detect communities in a large-scale network?

1.1.3 Submatrix localization

Definition 1.3 (Submatrix Localization)**.** Given a large random matrix, find a submatrix with atypical entries.

Example 1.7 (DNA microarray analysis)**.** The DNA microarray data can be represented as a matrix, with rows corresponding to different genes, columns corresponding to different samples, and (i, j)-th entry recording the expression level of gene i in sample j. One is interested in finding a group of genes that are differentially expressed in a set of patients. See Fig. 1.6.

Example 1.8 (Terrorist network detection)**.** Given a communication network, one is interested in identity the terrorist network with frequent communication. See Fig. 1.7 for illustration.

1.1.4 Covariance matrix estimation and Sparse PCA

Covariance matrix estimation plays a central role in statistics. It deals with the problem of how to estimate the true covariance matrix based on observations of samples drawn from a multivariate distribution. Estimation of covariance is useful for us to understand the correlation between different variables, and is an important subroutine in many statistical analysis including regression and classification.

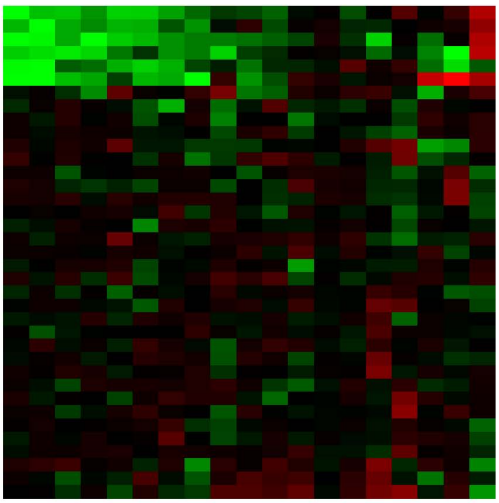
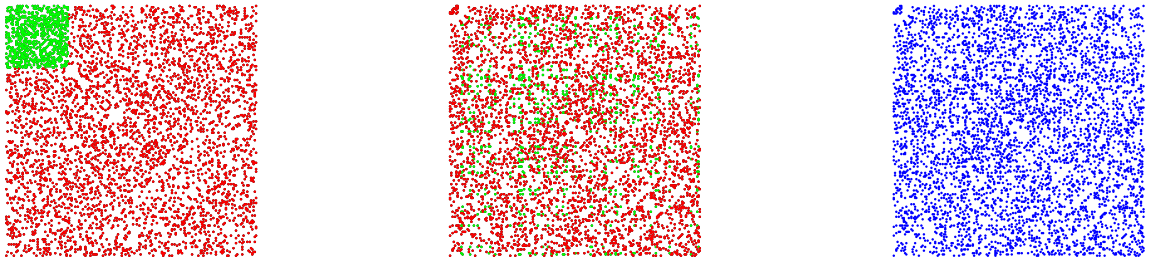


Figure 1.6: Gene expression matrix



(a) (b) (c)

Figure 1.7: The adjacency matrix of a network. Panel (a): The green submatrix corresponds to a group of nodes which are more densely connected than the other part of the network. Panel (b): The nodes are randomly permutated, so the entries of the green submatrix are dispersed over the whole matrix. Panel (c): The green color is erased, and the goal is to find the underlying green submatrix from the observation of the network.

Recall that for a random vector $X \in \mathbb{R}^d$, its mean is defined as $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$ and its covariance matrix $\Sigma \in \mathcal{S}_+^{d\times d}$ is defined as

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])].$$

We can succinctly write $\Sigma = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\right]$.

Definition 1.4 (Covariance matrix estimation)**.** Given n independent and identically distributed samples $x_1, \ldots, x_n$ from a distribution in $\mathbb{R}^d$ with zero mean and unknown covariance matrix $\Sigma = \mathbb{E}\left[x_1 x_1^\top\right] \in \mathcal{S}_+^{d\times d}$, estimate the covariance matrix Σ.

Note: If $\mathbb{E}[x_1] = 0$, a standard estimator of Σ is the *sample covariance matrix*:

$$\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top.$$

Since $\mathbb{E}\left[x_i x_i^\top\right] = \Sigma$, the random matrix $\widehat{\Sigma}$ is an unbiased estimator of Σ, but how close the sample covariance matrix $\widehat{\Sigma}$ is to the true covariance matrix Σ ?

The true covariance matrix admits an eigenvalue decomposition as $\Sigma = \sum_{i=1}^{d} \lambda_i v_i v_i^\top$ where $\lambda_1 \geq \lambda_2 \geq \cdots$. In many scenarios, we are not only interested in Σ itself, but also interested in its leading eigenvectors corresponding to the few largest eigenvalues. These leading eigenvectors of Σ are widely known as *principal components*. Note that

$$v_1 = \arg\max_{\|v\|_2=1} v^\top \Sigma v = \arg\max_{\|v\|_2=1} v^\top \mathbb{E}\left[XX^\top\right] v = \arg\max_{\|v\|_2=1} \mathbb{E}\left[(v^\top X)^2\right].$$

If $\mathbb{E}[X] = 0$, then $\mathbb{E}\left[(v^\top X)^2\right] = \mathsf{var}(v^\top X)$. Hence, the first principal component v_1 maximizes the variance of projection X to v among all directions. See Fig. 1.8

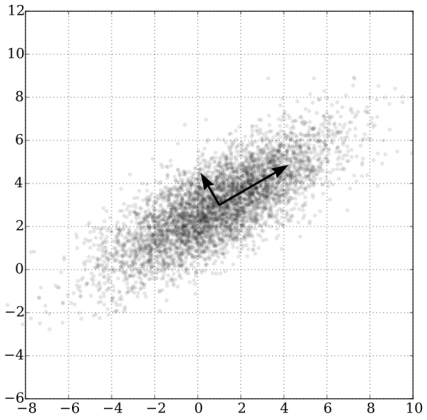


Figure 1.8: PCA of a multivariate Gaussian distribution centered at $(1, 3)$ with a standard deviation of 3 in roughly the $(0.866, 0.5)$ direction and of standard deviation 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue. Refer to https://en.wikipedia.org/wiki/Principal_component_analysis.

A standard estimator of v_1 is the first leading eigenvector of the sample covariance matrix $\widehat{\Sigma}$, i.e.,

$$\widehat{v}_1 = \arg\max_{\|v\|_2=1} v^\top \widehat{\Sigma} v.$$

Question: How close is $\widehat{v}_1$ to v_1?
In some applications, the principal components admit additional structures, e.g., sparsity.

Definition 1.5 (Sparse PCA)**.** Assume the true covariance matrix $\Sigma = \beta v v^\top + I$ with $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$ and $\|v\|_0 \le k$, estimate the sparse principal component v from n i.i.d. samples $x_1, \ldots, x_n$?

Note: v is the first principal component of Σ corresponding to eigenvalue $\beta + 1$. where β can be interpreted as the signal strength of v. If we neglect the sparsity constraint on v, i.e., $\|v\|_0 \le k$, then one can just use the first leading eigenvector $\widehat{v}_1$ of the sample covariance matrix $\widehat{\Sigma}$ as an estimator v. Is it good enough? Can we do better by exploiting the sparsity constraint?

1.2 Overview of tools and techniques

Two central questions will be addressed in this course:

- How shall we characterize the limit above which the task of extracting information is fundamentally possible and below which it is fundamentally impossible?
- How shall we develop computationally efficient algorithms that attain the fundamental limit, or understand the lack thereof.

Techniques

- Linear algebra: eigenvalue (singular value) decomposition. **Note**: In many scenarios, data can be represented as matrix. In the example of clustering customers based on purchase history, one can form data matrix with rows corresponding to customers, and columns corresponding to products; In the example of community detection, graph can be represented using adjacency matrix, where (i, j)-th entry equals one if vertices i and j are connected, and zero otherwise.
- Concentration inequalities. **Note**: Data contains noise which is often randomly distributed.
- Information and statistical theory. Used to characterize the fundamental limit of extracting information from data.
- Algorithms design. For example, convex programming and belief propagation algorithms.

Note: The following standard asymptotic notations will be used in this course. Fro two sequences of numbers a_n and b_n, where $b_n > 0$ for all sufficiently large n. Then

- $a_n = O(b_n)$ as $n \to \infty$ if there exist constants C and n_0 such that $|a_n| \le C b_n$ for all $n \ge n_0$.
- $a_n = \Omega(b_n)$ as $n \to \infty$ if there exist constants $c > 0$ and n_0 such that $a_n \ge c b_n$ for all $n \ge n_0$.
- $a_n = \Theta(b_n)$ as $n \to \infty$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$.
- $a_n \asymp b_n$ if $a_n = \Theta(b_n)$.
- $a_n \sim b_n$ if $a_n / b_n \to 1$.

- $a_n = o(b_n)$ as $n \to \infty$ if $a_n/b_n \to 0$.
- $a_n = \omega(b_n)$ as $n \to \infty$ if $a_n/b_n \to \infty$.
- $a_n \ll b_n$ if $a_n \geq 0$ and $a_n = o(b_n)$.

Also, we say that a sequence of events $\mathcal{E}_n$ holds with high probability (w.h.p.), if $\mathbb{P}[\mathcal{E}_n] \to 1$ as $n \to \infty$.

Note: The following standard norms for vectors will be used. For a vector $x \in \mathbb{R}^d$:

- L_2 norm $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$
- L_0 norm $\|x\|_0 = |\{i : x_i \neq 0\}|$.
- L_1 norm $\|x\|_1 = \sum_{i=1}^n |x_i|$.
- L_∞ norm $\|x\|_\infty = \max_{i \in [n]} |x_i|$.

1.3 What can go wrong in high dimensions?

Curse in dimensionality

- **Curse in computational efficiency.**

  In clustering or community detection problem, there are k^n different partitions of n data points, so enumerating all possible cluster partitions takes at least exponential time and thus is computationally intractable.

  In the submatrix localization problem, there are $\binom{n}{s}$ different supports for $s \times s$ principal matrices. Hence, if s scales with n, enumerating all possible supports is computational intractable. Similarly, in the sparse PCA problem, there are $\binom{n}{k}$ different supports for a k-sparse principal component.

- **Curse in statistical efficiency.**

  Substantially more samples are needed for estimation as dimension grows.

  In the example of clustering customers based on their preferences, consider dot product $\langle x_i, x_j \rangle$ as a measure of the similarity between customer i and customer j. Suppose each customer purchase m products out of d products uniformly at random. Then

  $$\mathbb{E}[\#\text{ of products co-purchased by customers } i \text{ and } j] = \frac{m^2}{d}.$$

  Hence, m has to scale as $\sqrt{d}$ so that $\langle x_i, x_j \rangle$ is not vanishing on expectation. However, in practice, m may remain fixed as d increases.

  In the sparse PCA problem, assume $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \beta v v^\top + I)$ and let $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \in \mathcal{S}_+^{d \times d}$. Let $\gamma = \frac{d}{n}$.

  - If $\beta \leq \sqrt{\gamma}$, then $\lambda_1(\widehat{\Sigma}) \to (1 + \sqrt{\gamma})^2$ and $\langle v_1(\widehat{\Sigma}), v \rangle \to 0$ almost surely.

– If $\beta > \sqrt{\gamma}$, then $\lambda_1(\widehat{\Sigma}) \to (1+\beta)(1+\gamma/\beta)$ and $\langle v_1(\widehat{\Sigma}), v\rangle \geq \epsilon$ for some fixed constant $\epsilon > 0$ almost surely.

Note: The sharp change at threshold $\beta = \sqrt{\gamma}$ is often referred as *BBP phase transition*. It implies that

– Low dimension ($d/n \to 0$): For any fixed constant $\beta > 0$, $\lambda_1(\widehat{\Sigma})$ is a consistent estimator of v.

– High dimension ($d/n \to \infty$): For any fixed constant $\beta > 0$, $v_1(\widehat{\Sigma})$ is asymptotically orthogonal to v.

1.4 What can help us in high dimensions

Structure hidden in the problems

- Low-rank structure: In the submatrix localization problem, if we assume that the entries inside the hidden submatrix are independently and identically distributed as $\mathcal{N}(\mu, 1)$, and the entries outside of the hidden submatrix are i.i.d. as $\mathcal{N}(0,1)$, then the expected value of the whole matrix A is given by $\mathbb{E}[A] = \mu \mathbf{1}_S \mathbf{1}_T$, where S (T) is the row (column) support of the hidden matrix and $\mathbf{1}_S$ is the indicator vector of set S. We can see that $\mathbb{E}[A]$ has rank 1.

- Sparsity structure: In the submatrix localization problem, the principal eigenvector of $\mathbb{E}[A]$ is proportional to the indicator vector of the support $\mathbf{1}_S$ and hence is $|S|$-sparse.

  In the sparse PCA problem, the principal eigenvector of the true covariance matrix Σ is v which is k-sparse.

Blessings of Dimensionality

- **Concentration of measure.**

  Theorem 1.1 (Gaussian concentration inequality for Lipschitz functions)**.** *Let $X \sim \mathcal{N}(0, I_d)$, and $F : \mathbb{R}^d \to \mathbb{R}$ be a 1-Lipchitz function (i.e., $|F(x) - F(y)| \leq \|x - y\|_2$ for all $x, y \in \mathbb{R}^d$). Then for any t,*

$$\mathbb{P}\left[|F(X) - \mathbb{E}[F(X)]| \geq t\right] \leq C\exp(-ct^2)$$

  for some absolute constants $C, c > 0$.

  In plain English, any 1-Lipchitz function of a standard Gaussian random vector, regardless of the dimension, exhibits concentration like a scalar standard Gaussian random variable.

  Example 1.9 (χ distribution)**.** Take $F(x) = \|x\|_2$ and let $X \sim \mathcal{N}(0, I_d)$. Then

$$\mathbb{P}\left[\|X\|_2 - \mathbb{E}[\|X\|_2] \geq t\right] \leq C\exp(-ct^2).$$

  Moreover, $\mathbb{E}[\|X\|_2] \leq \sqrt{\mathbb{E}\left[\|X\|_2^2\right]} = \sqrt{d}$. Hence, for any sequence $\rho_d \to \infty$ as $d \to \infty$, $\mathbb{P}\left[\|X\|_2 \leq \sqrt{d} + \rho_d\right] \to 1$. In fact, one can show $\mathbb{P}\left[\sqrt{d} - \rho_d \leq \|X\|_2 \leq \sqrt{d} + \rho_d\right] \to 1$. In other words, for high-dimensional standard Gaussian distribution, almost all probability mass is concentrated around the shell at distance $\sqrt{d}$ away from the origin.

Example 1.10 (maximum of d i.i.d. $\mathcal{N}(0,1)$ random variables)**.** Take $F(x) = \max_{1\le i\le d} x_i$ and let $X \sim \mathcal{N}(0, I_d)$. Then F is 1-Lipchitz (check) and thus

$$\mathbb{P}\left[\max_{1\le i\le d} X_i - \mathbb{E}\left[\max_{1\le i\le n} X_i\right] \ge t\right] \le C\exp(-ct^2).$$

Moreover, one can show $\mathbb{E}\left[\max_{1\le i\le d} X_i\right] \le \sqrt{2\log d}$ (HW). Hence, for any sequence $\rho_d \to \infty$ as $d \to \infty$, $\mathbb{P}\left[\max_{1\le i\le d} X_i \le \sqrt{2\log d} + \rho_d\right] \to 1$.

Aside: Another heuristic way to derive $\sqrt{2\log d}$. For any i, $\mathbb{P}\left[X_i \ge x\right] \approx e^{-x^2/2}$. Since there are d i.i.d. random variables $X_1, \ldots, X_n$, $\mathbb{P}\left[\max_{1\le i\le d} X_i \ge x\right] \approx de^{-x^2/2}$. Setting $de^{-x^2/2} = 1$, we get that $x = \sqrt{2\log d}$.

Using the fact that $-X_i$ has the same distribution as X_i, one can get that

$$\mathbb{P}\left[\min_{1\le i\le d} X_i \ge -\sqrt{2\log d} - \rho_d\right] \to 1.$$

Example 1.11 (Gaussian submatrix detection)**.** Let $A \in \mathbb{R}^{n\times n}$ denote a random matrix and $S \times T \subset [n] \times [n]$ denote the support of an $s \times s$ submatrix of A. Assume that $A_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ for $(i,j) \in S \times T$ and $A_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $(i,j) \notin S \times T$. where $\mu > 0$. The goal is to estimate the supports S and T from observation of A.

Method 1: Ignore the matrix structure and treat A as an n^2-dimensional random vector. Pick s^2 entries in A with the s^2 largest values, and declare those entries constitute the hidden submatrix. This method works only if the minimum value of $\{A_{ij} : (i,j) \in S \times T\}$ is strictly larger than the maximum value of $\{A_{ij} : (i,j) \notin S \times T\}$. Since with probability converging to 1,

$$\min_{(i,j)\in S\times T} A_{ij} \ge \mu - \sqrt{2\log s^2} - \rho_d,$$

and

$$\max_{(i,j)\in S\times T} A_{ij} \le \sqrt{2\log(n^2 - s^2)} + \rho_d,$$

we need $\mu > 2\rho_d + 2\sqrt{\log s} + \sqrt{2\log(n^2 - s^2)}$.

Method 2: For row i, its row sum $r_i := \sum_j A_{ij} \sim \mathcal{N}(s\mu, n)$ if $i \in S$; otherwise, $r_i \sim \mathcal{N}(0, n)$. Also, row sums are independent across different rows. Hence, one can estimate S by picking s rows with the s largest row sums. Similarly, one can estimate T by picking s columns with the s largest column sums. This method works only if the $\min_{i\in S} r_i > \max_{i\notin S} r_i$. Since with probability converging to 1,

$$\min_{i\in S} r_i \ge s\mu - \sqrt{2n\log s} - \rho_d,$$

and

$$\max_{i\notin S} r_i \le \sqrt{2n\log(n-s)} + \rho_d,$$

we need $s\mu > 2\rho_d + \sqrt{2n\log s} + \sqrt{2n\log(n-s)}$.

We see that method 1 requires smaller μ than method 2 to succeed if $s = o(\sqrt{n})$, and vice versa if $s = \omega(\sqrt{n})$.

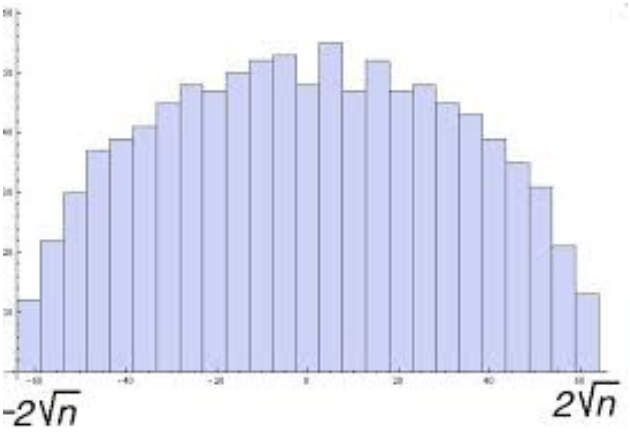
- **Asymptotics**. By letting either $d \to \infty$ or $n \to \infty$, we often get sharp asymptotic limits.

  Example 1.12 (Erdős-Rényi random graph)**.** Let $\mathcal{G}(n,p)$ denote the Erdős-Rényi random graph with n vertices and each pair of two vertices are connected by an edge with probability p independently at random. Let $p = \frac{a \log n}{n}$ for a fixed constant a. Then as $n \to \infty$,

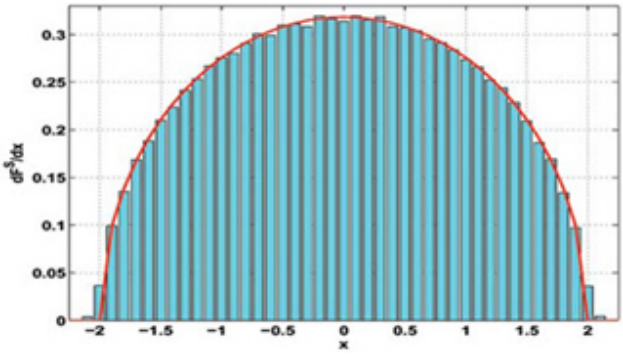
$$\mathbb{P}\left[G \text{ is connected}\right] \to \begin{cases} 1 & \text{if } a > 1, \\ 0 & \text{if } a < 1. \end{cases} \tag{1.1}$$

Example 1.13 (The semi-circular law)**.** Let A be a real, symmetric $n \times n$ matrix such that A_{ij} are independent for all $i \le j$, and $A_{ij} \sim \mathcal{N}(0,1)$ for all $i < j$ and $A_{ii} \sim \mathcal{N}(0,2)$ for all i. Then the eigenvalue histogram of $\frac{1}{\sqrt{n}} A$ (more precisely, the empirical eigenvalue distribution) converges to the semi-circular law $\mu(x)$ supported in $[-2,2]$:

$$\mu(x) = \frac{1}{2\pi} \sqrt{4 - x^2}.$$



(a)



(b)

Figure 1.9: Panel (a): Histogram of the eigenvalues of a 1000×1000 symmetric matrix with independent $\mathcal{N}(0,1)$ entries. Refer to Benedek Valko's course on random matrices http://www.math.wisc.edu/~valko/courses/833/833.html. Panel (b): Semi-circular law distribution. Image by Alan Edelman, MIT open courseware 18.996 / 16.399 Random Matrix Theory and Its Applications

Chapter 2

k-means Clustering

Outline

- Optimization formulation of k-means
- Convergence of k-means
- Failure cases of k-means

2.1 Optimization formulation of k-means

Recall in the data clustering problem, we are given n data points $x_1, x_2, \ldots, x_n \in \mathcal{X}$, and interested in partitioning them into k clusters.

- A psuedo-distance d is a mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}_+$, i.e.., $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.
- k-partition of $[n]$: $S_1 \cup S_2 \cup \cdots \cup S_k = [n]$ such that $S_i \cap S_j = \emptyset$ for all $i \neq j$.
- Center of a group $S \subset [n]$:
$$c(S) \in \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i \in S} d(x_i, z) \right\}.$$
- Cost of a k-partition $\underline{S} = (S_1, \ldots, S_k)$:
$$c(\underline{S}) := \sum_{a=1}^{k} \sum_{i \in S_a} d\left(x_i, c(S_a)\right).$$
- Seek a k-partition $\underline{S}$
$$\min_{\underline{S}} c(\underline{S}). \tag{2.1}$$

Note: : It is important to constrain the number of clusters k in the minimization problem (2.1). If instead the number of clusters is unconstrained, and one minimizes $c(\underline{S})$ over all possible partitions of $[n]$, then the minimizer is trivially given by treating each data point as an individual cluster. Determining a good choice of k from data is a non-trivial task in general.

Algorithm 1 k-means clustering

1: Input: Data $\{x_i\}_{i=1}^n$ and initial partition $\underline{S}$.
2: Output: New partition S'.
3: (Update step): let $c_a = c(S_a)$ for $1 \le a \le k$.
4: (Assignment step) for $1 \le a \le k$,

$$S'_a = \left\{ i \in [n] : a = \arg\min_{b \in [k]} d(x_i, c_b) \right\}.$$

5: Iterate steps 1–4 until $c(\underline{S}') \ge c(\underline{S}) - \epsilon$.

Note: In the update step of Algorithm 1, we need subroutine:

$$c(S) \in \arg\min_{z \in \mathcal{X}} \left\{ \sum_{i \in S} d(x_i, z) \right\}.$$

- Quadratic distance. If $X \equiv \mathbb{R}^n$ and $d(x, y) = \|x - y\|_2^2$, then

$$c(S) = \frac{1}{|S|} \sum_{i \in S} x_i \, .$$

  See Fig. 20.4 in Mackay's book [Mac03] for illustration of k-means clustering with quadratic distance.

- Spherical distance. If $X \equiv \mathcal{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ and $d(x, y) = 1 - \langle x, y \rangle$, then

$$c(S) = \frac{\sum_{i \in S} x_i}{\| \sum_{i \in S} x_i \|_2}.$$

- Kullback-Leibler divergence (KL divergence). If $X \equiv \mathcal{P}^{m-1} := \{x \in \mathbb{R}_+^m : \sum_i x(i) = 1\}$and $d(x, y) = D(x\|y) := \sum_{j=1}^m x(j) \log \frac{x(j)}{y(j)}$, then

$$c(S) = \frac{1}{|S|} \sum_{i \in S} x_i.$$

  Proof. To solve $\min_{z \in \mathcal{X}} \left\{ \sum_{i \in S} d(x_i, z) \right\}$, consider its Lagrangian function

$$\mathcal{L}(z, \lambda) := \sum_{i \in S} d(x_i, z) + \lambda (\sum_j z(j) - 1). \tag{2.2}$$

  Differentiate $\mathcal{L}(z, \lambda)$ with respect to $z(j)$ gives that

$$\frac{\partial \mathcal{L}(z, \lambda)}{\partial z(j)} = - \sum_{i \in S} x_i(j) \frac{1}{z(j)} + \lambda.$$

  Set $\frac{\partial \mathcal{L}(z, \lambda)}{\partial z(j)} = 0$ gives that

$$z(j) = \frac{1}{\lambda} \sum_{i \in S} x_i(j), \quad \forall 1 \le j \le m.$$

  Since $\sum_j z(j) = 1$, it follows that $\lambda = |S|$ and hence the optimal z^* is given by $z^* = \frac{1}{|S|} \sum_{i \in S} x_i$. □

Note: Properties of $D(x\|y)$:

1. $D(x\|y) \geq 0$ with equality if and only if $x = y$.

   Proof.

   $$D(x\|y) = \sum_j y(j)\frac{x(j)}{y(j)} \log \frac{x(j)}{y(j)} \geq \left(\sum_j \frac{x(j)}{y(j)} y(j)\right) \log \left(\sum_j \frac{x(j)}{y(j)} y(j)\right) = 0,$$

   where the inequality follows from the convexity of $x \log x$ and Jensen's inequality, and it becomes equality if and only if $\frac{x(j)}{y(j)}$ does not depend on j, i.e., $x = y$. □

2. $D(x\|y) \neq D(y\|x)$ in general (convince yourself by constructing examples).
3. $D(x\|y)$ is convex in (x, y).

   Proof. By definition, one can check that for any convex function $f : \mathbb{R}_+ \to \mathbb{R}$, $(p, q) \to qf\left(\frac{p}{q}\right)$ is convex on $\mathbb{R}_+^2$. Let $f(x) = x \log x$. It follows that $(p, q) \to p \log \frac{p}{q}$ is convex on $\mathbb{R}_+^2$. Hence, $D(x\|y)$ is jointly convex in x and y. □

Here is an alternative proof of (2.2) using the non-negativity of $D(x\|y)$. Let $y = \frac{1}{|S|}\sum_{i\in S} x_i$. Then for any $z \in \mathcal{P}^{m-1}$,

$$\sum_{i\in S} (D(x_i\|z) - D(x_i\|y)) = \sum_{i\in S} \left(\sum_{j=1}^m x_i(j) \log \frac{y(j)}{z(j)}\right) = |S| \sum_{j=1}^m y(j) \log \frac{y(j)}{z(j)} = |S| D(y\|z) \geq 0.$$

2.2 Convergence of k-means

Proposition 2.1. *1. If $\underline{S}^{t+1} = \underline{S}^t$, then $\underline{S}^{t+\ell} = \underline{S}^t$ for all $\ell \geq 1$.*

2. The cost $c(\underline{S}^t)$ is non-increasing in t.

3. k-means halts after at most k^n iterations.

Proof. Claim 1 follows immediately from the algorithm description.

For Claim 2, define

$$\widetilde{C}(\underline{S}, c) = \sum_{a=1}^k \left\{\sum_{i\in S_a} d(x_i, c_a)\right\}.$$

Denote centers at iteration t by c^t. Then by the update step, $c(\underline{S}^t) = \min_c \widetilde{C}(\underline{S}^t, c)$ and by the assignment step, $\widetilde{C}(\underline{S}^{t+1}, c^t) = \min_{\underline{S}} \widetilde{C}(\underline{S}, c^t)$. It follows that

$$c(\underline{S}^{t+1}) \leq \widetilde{C}(\underline{S}^{t+1}, c^t) \leq \widetilde{C}(\underline{S}^t, c^t) \leq c(\underline{S}^t).$$

Claim 3 follows from Claim 1 and the fact that there are at most k^n different k-partitions of $[n]$. □

Note: From the proof of Claim 2, k-means algorithm can be viewed as an alternating minimization algorithm, which minimizes the cost function $\widetilde{C}(\underline{S}, c)$ over k-partitions $\underline{S}$ and cluster centers c in an alternating fashion.

Note: Although k-means algorithm halts after at most k^n iterations, the outcome of the algorithm depends on the initial condition. See Figure 20.4 Mackay's book [Mac03] for an example.

2.3 Failure case of k-means

- Cluster sizes are unbalanced. See Figure 20.5 in Mackay's book [Mac03].
- Distance metric d does not capture the shape of clusters well. See Figure 20.6 in Mackay's book [Mac03].

Note: To be precise, the two failure cases listed above are caused by improper choice of objective function in (2.1).

Chapter 3

Model-based Approach to Clustering

Outline

- Model-based approach to clustering
- Maximum likelihood estimation
- Expectation-Maximization (EM) algorithm
- Spectral relaxation

Announcement

- Scribing assignment (due next Wednesday)
- Office hour: Mon 4:00-6:00 pm

3.1 Probabilistic approach to clustering (Model-based approach)

3.1.1 Recall: k-means clustering from an optimization point of view

Example 3.1. (Cost function)

$$C(\underline{S}) = \sum_{a=1}^{k} \sum_{i \in S_a} d(x_i, C(S_a)) \tag{3.1}$$

- Second summation: the sum of variation within cluster
- First summation: the sum of all possible clusters
- $C(\underline{S})$ can be viewed as the total variation within clusters.

Optimization-based approach

– props: make few assumptions on data $\{x_i\}$

– cons: proposed objective function is heuristic without theoretical roots.

3.1.2 Model-based approach (general recipe)

- come up a probabilistic (statistical) model for $\{X_i\}_{i=1}^n$
- principled approach to estimate the parameters from data
  - props: not heuristic→, and do not need guess
  - cons: need assumptions to the generating process of data. In particular, assume that data come from some particular model

3.1.3 Mixture density for clustering

Definition 3.1 (Mixture density for clustering)**.** Given data points $X_1, \ldots, X_n \sim P(X \mid C_1, \ldots, C_k)$ (i.i.d.),

$$P(X \mid C_1, \ldots, C_k) = \frac{1}{k} \sum_{a=1}^{k} q(X \mid C_a) \tag{3.2}$$

where $q(X \mid C_a)$ is a distribution for one cluster. Therefore, the mixture density for clustering is defined as an average of all the possible k density functions. Here we assume each cluster have equal weights $1/k$ for simplicity, and it is straightforward to generalize it to unequal weights.

Example 3.2 (Gaussian Mixture Model)**.**

$$q(x \mid c) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma^2}\|x-c\|_2^2} \tag{3.3}$$

where $x, c \in \mathbb{R}^d$. This function is the d-dimensional Gaussian density function.

Note: An equivalent way to describe mixture model

- For each data point X_i, first generate its cluster label y_i uniformly at random from $[k]$.
- Conditional on the cluster label y_i, draw $X_i \sim q(\cdot|C_{y_i})$.

Proof of the equivalence: Suppose x_i is drawn as above. Then

$$\begin{aligned} P(X_i \mid C_1, \ldots, C_k) &= \sum_{y_i} P(X_i, y_i \mid C_1, \ldots, C_k) \\ &= \sum_{y_i} P(y_i) P(X_i \mid y_i, C_1, \ldots, C_k) \\ &= \frac{1}{k} \sum_{y_i} q(X_i \mid C_{y_i}) \ (\because y_i \text{ is uniformly random}) \end{aligned} \tag{3.4}$$

which equals the probability density function of X_i in Definition 3.1.
Note:

- X_i: can be observed
- $C_1, \ldots, C_k, y_i$: cannot be observed

Clustering is to estimate $\left(\{y_i\}_{i=1}^n, \{C_a\}_{a=1}^k\right)$ from $\{X_i\}_{i=1}^n$.

3.2 Estimation - Maximum likelihood estimation

3.2.1 Maximum likelihood estimation

We have

- observation: $(X_1, \ldots, X_n)$ data points
- parameters of interest: $(C_1, \ldots, C_k, y_1, \ldots, y_k)$

Likelihood function is defined as follows.

$$\begin{aligned} \mathcal{L}\left(\underline{\mathrm{C}}, \underline{\mathrm{y}}\right) &\triangleq P\left(X_1, \ldots, X_n \mid \underline{\mathrm{C}}, \underline{\mathrm{y}}\right) \\ &= \prod_{i=1}^{n} P\left(X_i \mid \underline{\mathrm{C}}, \underline{\mathrm{y}}\right) \end{aligned} \tag{3.5}$$

Maximum likelihood estimation is to maximize this likelihood function $\mathcal{L}\left(\underline{\mathrm{C}}, \underline{\mathrm{y}}\right)$ over $\underline{\mathrm{C}}, \underline{\mathrm{y}}$:

$$\max_{\underline{\mathrm{C}}, \underline{\mathrm{y}}} \mathcal{L}\left(\underline{\mathrm{C}}, \underline{\mathrm{y}}\right)$$

By taking log function to the both sides of (3.5), we get log-likelihood function

$$\log P\left(X_1, \ldots, X_n \mid \underline{\mathrm{C}}, \underline{\mathrm{y}}\right) = \sum_{i=1}^{n} \log P\left(X_i \mid \underline{\mathrm{C}}, \underline{\mathrm{y}}\right). \tag{3.6}$$

Example 3.3 (Gaussian mixture model)**.** Recall that the mixture density $P\left(X_i \mid \underline{\mathrm{C}}, \underline{\mathrm{y}}\right)$ is as follows:

$$P\left(X_i \mid \underline{\mathrm{C}}, \underline{\mathrm{y}}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma^2}\left\|x_i - C_{y_i}\right\|_2^2} \tag{3.7}$$

Note that $\frac{1}{\left(2\pi\sigma^2\right)^{\frac{d}{2}}}$ does not depend on partition. Therefore, it is just constant in terms of $\underline{\mathrm{C}}$ and $\underline{\mathrm{y}}$. Taking log function to (3.7), we have

$$\log P\left(X_1, \ldots, X_n \mid \underline{\mathrm{C}}, \underline{\mathrm{y}}\right) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\|x_i - C_{y_i}\right\|_2^2 + \text{Const} \tag{3.8}$$

Maximum likelihood estimation for (3.8) is equivalent for solving the following minimization problem.

$$\begin{aligned} \text{MLE} &\Leftrightarrow \min_{\underline{\mathrm{C}}, \underline{\mathrm{y}}} \sum_{i=1}^{n} \left\|x_i - C_{y_i}\right\|_2^2 \\ &\Leftrightarrow \min_{\underline{\mathrm{y}}} \min_{\underline{\mathrm{C}}} \sum_{i=1}^{n} \left\|x_i - C_{y_i}\right\|_2^2 \end{aligned} \tag{3.9}$$

Fix the cluster label. Then (3.9) is equivalent to,

$$\min_{y_i} \sum_{i=1}^{n} \left\|x_i - C\left(S_a\right)\right\|_2^2 \tag{3.10}$$

where $S_a = \{j : y_j = a\}, C(S_a) = \frac{1}{S_a} \sum_{i \in S_a} X_i$, which is the average of data points in the cluster. Hence, under the Gaussian mixture model, MLE is equivalent to the optimization-based approach, where we aim to minimize the cost function $c(\underline{\mathrm{S}})$.

3.2.2 Another formulation of MLE

We have

- observation: $(X_1, \ldots, X_n)$
- parameters: $(C_1, \ldots, C_n)$
- Treat $(y_1, \ldots, y_n)$ as missing data.

Consider Gaussian mixture model for simplicity.

$$\begin{aligned} P(X_1, \ldots, X_n \mid C_1, \ldots, C_k) &= \prod_{i=1}^{n} P(X_i \mid C_1, \ldots, C_k) \\ &= \prod_{i=1}^{n} \left(\frac{1}{k} \sum_{a=1}^{k} q(x_i \mid C_a) \right) \\ &= \prod_{i=1}^{n} \left(\frac{1}{k} \sum_{a=1}^{k} \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma^2}\|x_i - C_{y_i}\|_2^2} \right) \end{aligned} \tag{3.11}$$

By taking log function, we have the following log-likelihood function,

$$\log P(X_1, \ldots, X_n \mid \underline{C}) = \sum_{i=1}^{n} \log \left(\frac{1}{k} \sum_{a=1}^{k} e^{-\frac{1}{2\sigma^2}\|x_i - C_{y_i}\|_2^2} \right) + \text{Const}, \tag{3.12}$$

where the constant term does not depend on the parameters $\underline{C}$. Note that (3.12) is not concave in cluster centers ($\underline{C}$).

Exercise: Show that $f(x) = \log\left(1 + e^{-x^2}\right)$ is not concave in X.

Note: When $k = 1$, the exponential term is eliminated and only the quadratic term remains, so it is a convex function. But when $k > 1$, it is neither convex or concave.

3.3 Expectation-Maximization (EM) algorithm

Note: We use the concept of covariance discussed in last lecture.

3.3.1 General problem formulation

We are interested in parameter $\theta = (\theta_1, \ldots, \theta_k)$, the cluster center.
Suppose that we observe one data point $(X_1, y_1) \sim P_\theta$, where X_1 is an observation, whereas y_1 is a hidden observation. Denote this y_1 as a hidden cluster label. There is no loss of generality in this assumption, since we generated data points independently.
Define loss function $l(\theta)$ as follows.

$$\begin{aligned} l(\theta) \triangleq l(\theta \mid X_1) &= -\log P(X_1 \mid \theta) \\ &= -\log \left(\sum_{y_1} P(X_1, y_1 | \theta) \right) \end{aligned} \tag{3.13}$$

Note:

- We will minimize the loss function for all the possible θ.
- Due to the existence of summation inside the log in the last term, the loss function is nonconvex as well as nonconcave.

How can we deal with (3.13)? To do this, the idea is to move $\sum$ to the outside of log.

$$\begin{aligned} l(\theta) &= -\log\left(\sum_{y_1} P(X_1, y_1(\theta))\right) \\ &= -\log \sum_{y_1} P\left(y_1 \mid X_1, \theta^t\right) \frac{P(X_1, y_1 \mid \theta)}{P\left(y_1 \mid X_1, \theta^t\right)} \\ &= -\log\left(\mathbb{E}_{Y_1 \sim P(\cdot \mid X_1, \theta^t)}\left[\frac{P[X_1, Y_1 \mid \theta]}{P\left[Y_1 \mid X_1, \theta^t\right]}\right]\right) \end{aligned} \tag{3.14}$$

Here, $P\left(y_1 \mid X_1, \theta^t\right)$ is the probability of hidden cluster label y_1 conditional on X_1 and the current estimate θ^t of θ. Also, since $-\log$ is a convex function for $\mathbb{R}_+$, by Jensen's inequality,

$$\begin{aligned} l(\theta) &= -\log\left(\mathbb{E}_{Y_1 \sim P(\cdot \mid X_1, \theta^t)}\left[\frac{P[X_1, Y_1 \mid \theta]}{P\left[Y_1 \mid X_1, \theta^t\right]}\right]\right) \\ &\leq -\mathbb{E}_{Y_1 \sim P(\cdot \mid X_1, \theta^t)} \log\left[\frac{P[X_1, Y_1 \mid \theta]}{P\left[Y_1 \mid X_1, \theta^t\right]}\right] \\ &= -\sum_{y_1} P\left(y_1 \mid X_1, \theta^t\right) \log \frac{P[X_1, y_1 \mid \theta]}{P\left[y_1 \mid X_1, \theta^t\right]} \\ &\triangleq Q\left(\theta \mid \theta^t\right) \end{aligned} \tag{3.15}$$

Here, $Q\left(\theta \mid \theta^t\right)$ is a function of θ, which gives an upper bound of $l(\theta)$ for all θ.

Remark If $\log P(X_1, Y_1 \mid \theta)$ is concave in θ, then $Q\left(\theta \mid \theta^t\right)$ is convex in θ.

Question Why Q? (Why should we choose particular Q function?)

Claim

$$Q\left(\theta \mid \theta^t\right) = l(\theta) + D\left(P\left(\cdot \mid y_1, \theta^t\right) \| P(\cdot \mid X_1, \theta)\right)$$

(Note that $D\left(P\left(\cdot \mid y_1, \theta^t\right) \| P(\cdot \mid X_1, \theta)\right)$ is the KL divergence between probability distribution $P\left(\cdot \mid y_1, \theta^t\right)$ and $P(\cdot \mid X_1, \theta)$).

Implication

- $Q\left(\theta \mid \theta^t\right) \geq l(\theta)$ (always upper bound)
- $Q\left(\theta \mid \theta^t\right)|_{\theta=\theta^t} = l(\theta)|_{\theta=\theta^t}$ ($\because$KL divergence is zero at θ^t)

Recall $D(x \| y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i}$

Note that KL divergence is not symmetric(i.e. $D(x \| y) \neq D(y \| x)$). That's why here we use $\|$ instead of , .

- $\nabla_\theta Q\left(\theta \mid \theta^t\right)|_{\theta=\theta^t} = \nabla_\theta l\left(\theta\right)|_{\theta=\theta^t}$, which is equivalent to

$$\nabla_\theta D\left(P\left(\cdot \mid y_1, \theta^t\right) \| P\left(\cdot \mid X_1, \theta\right)\right)|_{\theta=\theta^t} = 0. \tag{3.16}$$

  Proof: Since $D\left(P\left(\cdot \mid y_1, \theta^t\right) \| P\left(\cdot \mid X_1, \theta\right)\right)$ is nonnegative, we have a local minimum at $\theta = \theta^t$, and thus its derivative at $\theta = \theta^*$ must be zero.

Proof of claim. Revisit the definition of $Q\left(\theta \mid \theta^t\right)$.

$$\begin{aligned} Q\left(\theta \mid \theta^t\right) &\triangleq -\sum_{y_1} P\left(y_1 \mid X_1, \theta^t\right) \log \frac{P\left[X_1, y_1 \mid \theta\right]}{P\left[y_1 \mid X_1, \theta^t\right]} \\ &= l\left(\theta\right) - \sum_{y_1} P\left(y_1 \mid X_1, \theta^t\right) \log \frac{P\left[y_1 \mid X_1, \theta\right]}{P\left[y_1 \mid X_1, \theta^t\right]} \end{aligned} \tag{3.17}$$

In the first row, $P\left(X_1, y_1 \mid \theta\right)$ is equal to $P\left(X_1 \mid \theta\right) P\left(y_1 \mid X_1, \theta\right)$. Since $P\left(X_1 \mid \theta\right)$ does not depend on y_1, it can be moved out of summation. Also, in the second row, the minus$(-)$ can enter inside the summation, and then it becomes KL divergence. That is,

$$Q\left(\theta \mid \theta^t\right) = l\left(\theta\right) + D\left(P\left(\cdot \mid X_1, \theta^t\right) \| P\left(\cdot \mid X_1, \theta\right)\right) \tag{3.18}$$

□

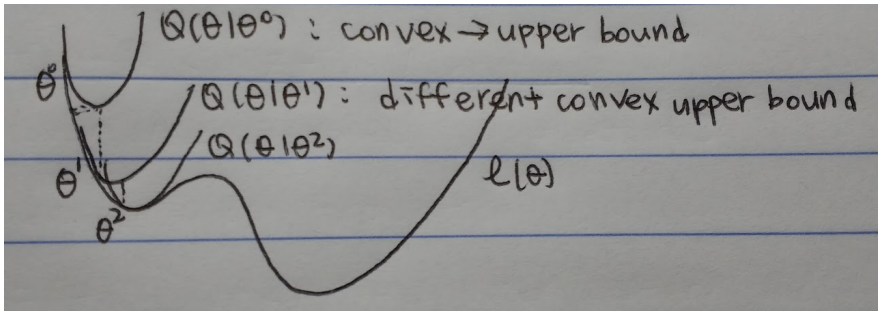


Figure 3.1: Iterative process of computing new θ^t

3.3.2 EM algorithm

- Input: Data $\{X_i\}_{i=1}^n$, initial estimate of θ, θ^0
- Output: estimate of θ, $\theta^\star$

For $t \in \{0, 1, \ldots\}$ do

E-step: Compute $P\left(y_1 \mid X_1, \theta^t\right)$

M-step: $\theta^{t+1} = \arg\min_\theta Q\left(\theta \mid \theta^t\right)$

Note:

- E-step is the expectation step. By doing this step, $Q\left(\theta \mid \theta^t\right)$ is computed.

- Through M-step, minimization step, we get new θ^{t+1}.

From Fig 3.1, we can see that EM algorithm is a special case of Majorization-Minimization(MM).

Proposition 3.1 (Convergence of EM)**.** *We have the following properties.*

- $l(\theta^0) \geq l(\theta^1) \geq ... \geq l(\theta^t) \geq ... \Rightarrow l(\theta^0) \geq 0$ *and thus* $\lim_{t\to\infty} l(\theta^t)$ *exists.*
- *if* $\theta^{t+1} = \theta^t \Rightarrow \nabla_\theta l(\theta^t) \mid_{\theta=\theta^t} = 0$.
- *Let* θ^* *be a limiting point of* θ^t*. Under some regularity conditions (see for details), we have* $\nabla_\theta l(\theta^t) \mid_{\theta=\theta^*} = 0$

Note: EM algorithm is very similar to k-means in that way that its quality depends on where the initial estimate is.

3.3.3 EM in Gaussian Mixture Clustering

- $\theta = (C_1, C_2, ..., C_k)$
- y_i is cluster label of x_i

$$p(y_i|x_i, \theta^t) = \frac{p(x_i, y_i, \theta^t)}{p(x_i, \theta^t)} = \frac{p(x_i, y_i, \theta^t)}{\sum_y (x_i, y, \theta^t)} \tag{3.19}$$

With Gaussian mixture, it gives that

$$p(y_i|x_i, \theta^t) = \frac{\frac{1}{k} e^{-\frac{1}{2\sigma^2}\|x_i - C_{y_i}^t\|_2^2}}{\sum_y \frac{1}{k} e^{-\frac{1}{2\sigma^2}\|x_i - C_y^t\|_2^2}}$$

And the $Q\left(\mathbb{C} \mid \mathbb{C}^t\right)$ is,

$$\begin{aligned} Q(\mathbb{C}|\mathbb{C}^t) &= -\sum_{i=1}^{n} \sum_{y_i} p(y_i|x_i, C^t) \log(\frac{p(x_i, y_i|c)}{p(y_i|x_i, c^t)}) \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{y_i} p(y_i|x_i, C^t) \|x_i - C_{y_i}\| + const, \end{aligned} \tag{3.20}$$

where the constant term here does not depend on $\mathbb{C}$. To compute the new center C_a^{t+1}, we take partial derivative of $Q(\mathbb{C}|\mathbb{C}^t)$ with respect to C_a, and get that

$$C_a^{t+1} = \frac{\sum\limits_i p(y_i = a|x_i, C^t) x_i}{\sum\limits_i p(y_i = a|x_i, C^t)}. \tag{3.21}$$

Note:

- θ^t is the current estimate of parameter.
- $p(y_i|x_i, C^t)$ is the probability that each data point X_i belongs to the cluster with cluster label y_i.
- C^{t+1} is a weighted average of data points x_i with weights given by $p(y_i|x_i, C^t)$.

3.3.4 Summary

E-Step: Calculate the following probability.

$$p(y_i|x_i, \underline{C}^t) = \frac{e^{-\frac{1}{2\sigma^2}\|x_i - C_{y_i}\|_2^2}}{\sum\limits_y exp^{-\frac{1}{2\sigma^2}\|x_i - C_y\|_2^2}} \tag{3.22}$$

M-step: Then the new center is decided as follows.

$$C_y^{t+1} = \frac{\sum\limits_i p(y_i|x_i, C^t)x_i}{\sum ip(y_i|x_i, C^t)} \tag{3.23}$$

Note:

- It is also known as "soft" k-means. At the E-step, we compute the probability that each data point belong to each cluster, whereas the "hard" k-means decide whether each data point belong to the cluster or not.
- If $\sigma^2 \to 0$, then $p(y_i|x_i, c^t) \to 1_{y_i \in \arg\min \|x_i - c_y^t\|_2}$ (assume it has a unique minimizer). In this case, EM algorithm acts like "hard" k-means.
- If we want to output what cluster a certain data point belongs to in the end, we could find y_i that maximizes the probability $p(y_i|x_i, c^t)$.

3.3.5 Failure Cases of k-means objective function

We mentioned two failure cases of the k-means objective function $c(\underline{S})$.

1. unequal cluster size. To deal with unequal cluster sizes, we could introduce
   - $\pi_1, \pi_2, ..., \pi_k$
     $p(y_i = a) = \pi_a$
   - $\sigma_1, \sigma_2, ..., \sigma_k$ are different for different clusters.

   Hence, all the unknown parameters to estimate are given by $(c_1, ..., c_k, \pi_1, ..., \pi_k, \sigma_1, ..., \sigma_k)$
2. shapes are not "spherical"
   - Note that previously we assume X_i are generated from a "spherical" Gaussian.

$$\begin{aligned} p(x_i|c_{y_i}, \sigma_{y_i}) &\propto \frac{1}{(2\pi\sigma_{y_i}^2)^{d/2}} e^{-\frac{1}{2\sigma_{y_i}^2}\|x - c_{y_i}\|_2^2} \\ E[X_i X_i^T] &= \sigma_{y_i}^2 I_{d\times d} \end{aligned} \tag{3.24}$$

   - As you can see, the "spherical" Gaussian assumption results in the sphere shape of the data. To deal with more general case, we can consider more general covariance. That is, the covariance matrix does not have to be proportional to the identity matrix any longer.

$$\mathbb{E}\left[X_i X_i^T\right] = \begin{pmatrix} (\sigma_{y_i}^{(1)})^2 & & & \\ & (\sigma_{y_i}^{(2)})^2 & & \\ & & \ddots & \\ & & & (\sigma_{y_i}^{(d)})^2 \end{pmatrix}$$

In this case, all the unknown parameters $(c_1, ..., c_k, \pi_1, ..., \pi_k, \sigma_1^{(1)}, ..., \sigma_1^{(d)}, ..., \sigma_k^{(1)}, ..., \sigma_k^{(d)})$

Caveat: As above, we have a total of $2k + kd$ unknown parameters. In the high-dimensional regime where d is comparable to or larger than the number of samples n, then the number of unknown parameters may exceed the amount of data available, and we might run into the problem of *over-fitting*. Intuitively speaking, with so many unknown parameters, we have enough freedom to perfectly fit the observed data, and this perfect fit may not have any predictive power.

In the clustering setting, put one cluster exactly on one data point and let its variance go to zero you can obtain an arbitrarily large likelihood. Hence, the maximum likelihood methods including EM will break down by finding highly tuned models that fit part of the data perfectly. Please refer to Section 22.4 in Mackay's book [Mac03] for more details.

To deal with clustering in high-dimensional, we are going to introduce a new method called spectral clustering.

3.4 Spectral Clustering

We first derive a spectral method by starting from k-means objective function.

3.4.1 Specture Relaxation of k-means

Recall: Quadratic Distance Setting

$$\min_{k-partition} \sum_{a=1}^{k} \sum_{i \in S_a} \|x_i - c(S_a)\|_2^2 \tag{3.25}$$

The objective function of (3.25) can be rewritten as:

$$\begin{aligned} \sum_{a=1}^{k} \sum_{i \in S_a} \|x_i - c(S_a)\|_2^2 &= \sum_{a=1}^{k} \sum_{i \in S_a} \|x_i - \frac{1}{|S_a|} \sum_{j \in S_a} x_j\|_2^2 \\ &= \sum_{a=1}^{k} \sum_{i \in S_a} \left(\|x_i\|^2 - \frac{1}{|S_a|} \sum_{j \in S_a} x_j \|^2 \right) \\ &= \sum_{a=1}^{k} \sum_{i \in S_a} \|x_i\|^2 - \sum_{a=1}^{k} \frac{1}{|S_a|} \| \sum_{i \in S_a} x_j \|_2^2 \end{aligned} \tag{3.26}$$

Since the former summation of the last row of (3.26) is not related to how the clustering comes, this problem is the same as

$$\max_{k-partition} \sum_{a=1}^{k} \frac{1}{|S_a|} \| \sum_{j \in S_a} x_j \|_2^2. \tag{3.27}$$

We define

$$X_{n \times d} = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in R^{n \times d}. \tag{3.28}$$

Denote 1_{S_a} as an indicated vector of cluster a. That is,

$$1_{S_a}(i) = \begin{cases} 1 & \text{if } i \in S_a, \\ 0 & \text{otherwise.} \end{cases} \tag{3.29}$$

So we have,

$$\sum_{i \in S_a} x_i = X^\top \cdot 1_{S_a} = (x_1, ..., x_n) \cdot 1_{S_a}.$$

With this result, we have

$$\begin{aligned} \sum_{a=1}^{k} \frac{1}{|S_a|} \| \sum_{j \in S_a} x_j \|_2^2 &= \sum_{a=1}^{k} \frac{1}{|S_a|} \| x^T \cdot 1_{S_a} \|_2^2 \\ &= \sum_{a=1}^{k} \frac{1}{|S_a|} \langle x^T \cdot 1_{S_a}, x^T \cdot 1_{S_a} \rangle \\ &= \sum_{a=1}^{k} \frac{1}{|S_a|} 1_{S_a}^T x \cdot x^T 1_{S_a} \\ &= \sum_{a=1}^{k} \left(\frac{1_{S_a}}{\sqrt{|S_a|}} \right)^T x \cdot x^T \left(\frac{1_{S_a}}{\sqrt{|S_a|}} \right) \end{aligned} \tag{*}$$

Note:

- $\left(\frac{1_{S_a}}{\sqrt{|S_a|}} \right)$ acts like a normalized indicator vector, so its norm is 1.
- Consider XX^T.

$$\begin{aligned} XX^T &= \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \begin{pmatrix} X_1 \cdots X_n \end{pmatrix} \\ &= \begin{pmatrix} < X_1, X_1 > & < X_1, X_2 > & \cdots & < X_1, X_n > \\ & & \ddots & \\ & & & < X_n, X_n > \end{pmatrix} \end{aligned}$$

  As you can see, it can be interpreted as a pairwise similarity matrix. As two points get closer, the inner product becomes larger.

Recall $Tr(A) = \sum_{i=1}^{n} A_{ii}$. If A and B are two $n \times d$ matrix, we have

$$\langle A, B \rangle \triangleq Tr(A^T B) = Tr(B A^T).$$

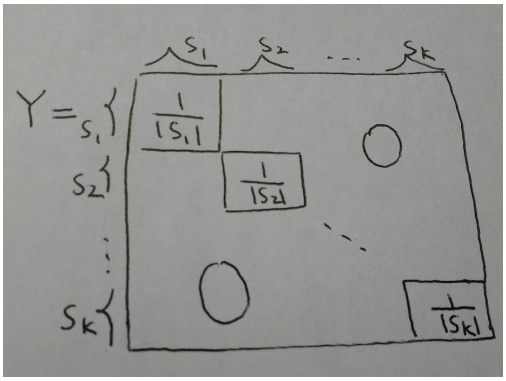
Let's continue to see (*).

$$\begin{aligned} (*) &= \sum_{a=1}^{k} Tr \left(\frac{1_{S_a}}{\sqrt{|S_a|}} x \cdot x^T \frac{1_{S_a}^T}{\sqrt{|S_a|}} \right) \\ &= \sum_{a=1}^{k} Tr \left(x \cdot x^T, \frac{1_{S_a} 1_{S_a}^T}{\sqrt{|S_a|}\sqrt{|S_a|}} \right) \\ &= \langle x \cdot x^T, \sum_{a=1}^{k} \frac{1_{S_a} 1_{S_a}^T}{|S_a|} \rangle \end{aligned}$$

Define Y as follows.

$$Y := \sum_{a=1}^{k} \frac{1_{S_a} 1_{S_a}^T}{|S_a|} \tag{3.30}$$

We call Y *cluster matrix* as it can be viewed as a matrix representation of cluster structure. When the rows (columns) from the same clusters are arranged together, Y is a diagonal-block matrix: In general, Y is a randomly permuted diagonal-block matrix. Using Y, maximizing the k means



objective is equivalent to

$$\begin{aligned} &max \quad \langle XX^T, Y \rangle \\ &s.t. \quad Y \text{ is indeed a valid cluster matrix.} \end{aligned} \tag{3.31}$$

In plain language, we would like to maximize the sum of total similarities within clusters among all valid k-partitions.

Definition 3.2. (Spectral relaxation)

$$\begin{aligned} &max \quad \langle XX^T, Y \rangle \\ &s.t. \quad Y = \sum_{a=1}^{k} U_a U_a^T \\ &\text{where} \quad \|U_a\|_2 = 1 \text{ and } U_a \perp U_b \text{ for } a \neq b \end{aligned} \tag{3.32}$$

Note:

- The first constraint, $\sum_{a=1}^{k} U_a U_a^T$, is expressed as an eigen-value decomposition. Note that $rank(Y) = k$.
- To tell that (3.32) is a relaxation of (3.31), we need to show that (3.32) has a larger feasible set. In particular, suppose Y is a valid cluster matrix, then because $Y = \sum_{a=1}^{k} \frac{1_{S_a} 1_{S_a}^T}{|S_a|}$, certainly Y satisfy the constraints in (3.32).

To see how should we solve (3.32), consider the following two cases.

- case 1: $k = 1$

$$\max_{\|U\|_2=1} \langle XX^T, UU^T \rangle = \max_{\|U\|_2=1} U^T XX^T U = \lambda_1(XX^T) \tag{3.33}$$

Here, λ_1 is the largest eigenvalue, and we use u_1 to denote the corresponding eigen-vector. So, the maximizer is $u_1(XX^T)$.

- case 2: $k > 1$
  Suppose $XX^\top$ admits the following eigenvalue decomposition:

$$XX^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T, \tag{3.34}$$

  where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, and u_i is an eigenvector corresponding to λ_i. Then the optimal solution of spectral relaxation problem is

$$Y^* = \sum_{i=1}^{k} u_i u_i^T \tag{3.35}$$

Note: Let $\widehat{Y}$ denote the output of the spectral relaxation. If $\widehat{Y}$ is exactly a valid cluster matrix, then we are done. However, in general, $\widehat{Y}$ will not be a valid cluster matrix. To get a valid clustering from $\widehat{Y}$, we will run a second step. We will discuss it later in this course.

In then next lecture, we will rigorously show the optimal solution of the spectral relaxation program is indeed given by Y^*.

Chapter 4

Singular value decomposition and Spectral clustering

Recap:

- EM algorithm for Gaussian mixture model
- Spectral relaxation of K means

Plan ahead:

- SVD
- Spectral clustering under Gaussian mixture
- Davis-Kahan $\sin\theta$ theorem
- Concentration inequalities

4.1 Singular Value Decomposition

Recap of eigenvalue decomposition:

Let A denote an $n \times n$ symmetric matrix A. We say that $u \in \mathbb{R}^n$ is an eigenvector of A corresponding to the eigenvalue λ if $Au = \lambda u$.

The eigenvalue decomposition of A is given as follows:

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$$
$$\lambda_1 \geq \cdots \lambda_n \in \mathbb{R}$$

where $u_i \in \mathbb{R}^n, \|u_i\|_2 = 1,\ u_i \perp u_j,\ \forall i \neq j$. We can also write the decomposition succinctly as

$$A = U\Lambda U^T,\ U^T U = I,$$

where $U = [u_1, u_2, \ldots, u_n]$, and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$.

If $A \in \mathbb{R}^{m \times n}$ with $m \neq n$, then the definitions of eigenvalues and eigenvectors will not apply. In this case, we need to introduce the singular value decomposition.

Definition 4.1 (SVD)**.** We sat $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ are left singular vector and right singular vector, respectively, and $\sigma \in \mathbb{R}_+$ is the corresponding singular value, if they together satisfy

$$Av = \sigma u$$
$$u^T A = \sigma v^T$$

In plain english, when we right multiply A by v, we get back u scaled by σ. Similarly, when we left multiply A by $u^\top$, we get back v scaled by σ.
Note: According to the definition, $u^\top Av = \sigma\|v\|_2^2 = \sigma\|u\|_2^2$. Hence, one can assume $\|u\|_2 = \|v\|_2 = 1$ without loss of generality in the definition of singular vectors.

4.1.1 Best Fit Vector

This subsection gives the geometric meanings of singular vectors and singular values.

Consider $v^* \in \arg\max_{\|v\|_2=1} \|Av\|_2^2$, where

$$Av = \begin{pmatrix} A_1^T \\ \vdots \\ A_m^T \end{pmatrix} v = \begin{pmatrix} < A_1, v > \\ \vdots \\ < A_m, v > \end{pmatrix},$$

where A_i denotes the i-th row of A. Note that the projection of A_i on v is given by $< A_i, v > v$, and the squared length of the projection is $< A_i, v >^2$.
Therefore, $\|Av\|_2^2$ is the sum of the squared length of projections of rows of A on v and thus v^* is the direction which maximizes the sum of the squared length of projections. Also, by Pythagorean theorem,

$$\|A_i\|_2^2 =< A_i, v >^2 + dist^2(A_i, v)$$

Hence, equivalently, v^* is the closest one-dimensional linear subspace to the rows of A in terms of the sum of the squared distances. For this reason, we call v^* the best-fit vector of rows of A.

The following proposition gives a connection between singular vectors and the best fit vector.

Proposition 4.1. *The vector v^* is an eigenvector of $A^\top A$ corresponding to the largest eigenvalue. Moreover, v^* is a singular vector of A and $\|Av^*\|_2$ is the largest singular value of A.*

Proof.

$$\begin{aligned} \|Av^*\|_2^2 &=< Av, Av > \\ &= v^T A^T Av, \end{aligned} \tag{4.1}$$

Notice that $A^T A \in \mathbb{R}^{n\times n}$ and it's symmetric. So we can apply the eigenvalue decomposition:

$$A^\top A = \sum_{i=1}^{n} \lambda_i v_i v_i^\top,$$

where λ_i are real eigenvalues of $A^T A$ and $\lambda_1 \geq \cdots \geq \lambda_n$. Also, $A^\top A$ is positive semi-definite and thus $\lambda_n \geq 0$.

Plugging the eigenvalue decomposition of $A^\top A$ into (4.1), we get that

$$\begin{aligned} v^T A^T A v &= \sum_{i=1}^{n} \lambda_i v^T v_i v_i^T v \\ &= \sum_{i=1}^{n} \lambda_i (v^T v_i)^2 \end{aligned}$$

So to get v^*, it's equivalent to solve

$$\begin{aligned} &\max \sum_{i=1}^{n} \lambda_i (v^T v_i)^2 \\ &s.t. \sum_{i=1}^{n} (v^T v_i)^2 = 1 \end{aligned}$$

Since all λ_i's are non-negative, the optimal solution is given by $v^* = v_1$, and the optimal value is given by $\|Av^*\|_2^2 = \lambda_1$. Therefore,

$$A^T A v^* = \lambda_1 v^*$$

Define $u^* = \frac{Av^*}{\sqrt{\lambda_1}}$ and $\sigma_1 = \sqrt{\lambda_1} = \|Av^*\|$.
It follows from the last displayed equation that

$$A^T u^* = \sigma_1 v^*$$

By definition of u^*

$$Av^* = \sigma_1 u^*$$

Therefore, σ_1 is a singular value of A, and u^* and v^* are the corresponding left and right singular vectors, respectively.

To see σ_1 is the largest singular value, consider any singular value τ of A, and its corresponding left and right singular values x and y. Then by definitions of singular vectors and values, we have $A^\top A y = \sigma^2 y$. Hence, $\sigma^2 \leq \lambda_1$, i.e., $\tau \leq \sigma_1$.

□

4.1.2 Best Fit Subspace

From best fit vector to best fit subspace. In this section, we find the best fit subspace of dimension k for rows of A via a greedy algorithm.

Theorem 4.1. *Define* $V_k = span\{v_1, v_2, \cdots, v_k\}$, *where*

$$\begin{aligned} v_1 &\in \arg \max_{\|v\|_2=1} \|Av\|_2^2 \\ v_2 &\in \arg \max_{\|v\|_2=1, v \perp v_1} \|Av\|_2^2 \\ &\vdots \\ v_k &\in \arg \max_{\|v\|_2=1, v \perp v_1, \cdots, v_{k-1}} \|Av\|_2^2 \end{aligned}$$

with ties broken arbitrarily. Let $\lambda_i = \|Av_i\|_2^2$ *for* $1 \le i \le k$. *Then*

1. $V_k \in \arg\max_{V:dim(V)\le k} \sum_{i=1}^n dist^2(A_i, V)$, *where* $dist(x, V) = \min_{y\in V} \|x - y\|_2$.

2. $A^\top A = \sum_{i=1}^n \tau_i x_i x_i^\top$ *denote its eigenvalue decomposition with* $\tau_1 \ge \tau_2 \ge \cdots \ge \tau_n \ge 0$. *Then* $v_1, \ldots, v_k$ *can be chosen so that* $v_i = x_i$ *and* $\lambda_i = \tau_i$ *for all* $1 \le i \le k$.

Proof. We first prove Claim 1. Base case $k = 1$ has already been proved in Proposition 4.1.
Assume V_{k-1} is indeed optimal (best fit subspace with dimension $\le k-1$).
Suppose V_k' is an optimal subspace of $dim(V_k') = k$. There exists an orthonormal basis of V_k', given by $w_1, \cdots, w_k$ such that $w_k \perp V_{k-1}$ (Convince yourself such a basis indeed exists).
By definition of v_k,

$$\|Aw_k\|_2 \le \|Av_k\|_2$$

By optimality of V_{k-1} (induction hypothesis),

$$\sum_{i=1}^{k-1} \|Aw_i\|_2 \le \sum_{i=1}^{k-1} \|Av_i\|_2$$

By optimality of V_k',

$$\sum_{i=1}^{k} \|Aw_i\|_2 \le \sum_{i=1}^{k} \|Av_i\|_2$$

By the above three inequalities, we can see they are all equalities. In particular, $\sum_{i=1}^k \|Aw_i\|_2 = \sum_{i=1}^k \|Av_i\|_2$, and thus V_k is also optimal.

Next, we prove claim 2 using again the induction method. Base case $k = 1$ follows from Proposition 4.1. Suppose the claim is true for $k-1$. We prove it also holds for k. Consider $\max_{\|v\|_2=1, v\perp v_1,\ldots,v_{k-1}} \|Av\|_2^2$. By induction hypothesis, $v_i = x_i$ for $1 \le i \le k-1$, the maximization problem is equivalent to

$$\begin{aligned} &\max_v \sum_{i=k}^n \tau_i (v^T x_i)^2 \\ &s.t. \sum_{i=k}^n (v^T x_i)^2 = 1 \end{aligned}$$

Therefore, x_k is an optimal solution and τ_2 is the optimal value. Hence, we could choose $v_k = x_k$ and $\lambda_k = \tau_k$. □

Proposition 4.2. *Let* $v_1, \cdots, v_n$ *be formed by the greedy algorithm above, then the following holds,*

1.

$$\begin{aligned} \sigma_1 \triangleq \|Av_1\|_2 &\ge \sigma_2 \triangleq \|Av_2\|_2 \\ &\vdots \\ &\triangleq \sigma_2 = \|Av_2\|_2 \\ &> \sigma_{r+1} \triangleq \|Av_r\|_2 = \cdots = \sigma_n \triangleq \|Av_n\|_2 = 0, \end{aligned}$$

where r *is some integer in* $[n]$.

2.

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

where $u_i \triangleq \frac{Av_i}{\sigma_i}$ *for* $\sigma_i > 0$. *Moreover,* $u_i \perp u_j, \forall i \neq j$ *and* $\|u_i\|_2 = 1$. *In particular,* u_i *and* v_i *are the left and right singular vectors corresponding to* σ_i, *respectively.*

Proof. Claim 1 follows from the construction of $\{v_1, \ldots, v_n\}$.

We prove Claim 2. Since $\{v_1, \cdots, v_n\}$ form an orthonormal basis for $\mathbb{R}^n$, we have

$$\begin{aligned} A_i &= \sum_{l=1}^{n} < A_i, v_l > v_l \\ &= \sum_{l=1}^{r} < A_i, v_l > v_l, \end{aligned}$$

where the second equality holds because $\sigma_{r+1} = \cdots = \sigma_n = 0$. Hence, we can rewrite A as

$$\begin{aligned} A &= \begin{pmatrix} \sum_{l=1}^{n} < A_1, v_l > v_l^T \\ \vdots \\ \sum_{l=1}^{n} < A_m, v_l > v_l^T \end{pmatrix} \\ &= \begin{pmatrix} < A_1, v_1 > & \cdots & < A_1, v_r > \\ \vdots & \ddots & \vdots \\ < A_m, v_1 > & \cdots & < A_m, v_r > \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_r^T \end{pmatrix} \end{aligned}$$

Also by definition of u_i,

$$u_i = \frac{Av_i}{\sigma_i} = \frac{1}{\sigma_i} \begin{pmatrix} < A_1, v_i > \\ \vdots \\ < A_m, v_i > \end{pmatrix}$$

We can further rewrite A as

$$\begin{aligned} A &= \begin{pmatrix} u_1\sigma_1 & \cdots & u_r\sigma_r \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_r^T \end{pmatrix} \\ &= \begin{pmatrix} u_1 & \cdots & u_r \end{pmatrix} \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_r^T \end{pmatrix} \\ &\triangleq U \Sigma V^T \end{aligned}$$

To prove $u_i \perp u_j, \forall i \neq j$, recall that Theorem 4.1 shows that

$$A^T A v_i = \lambda_i v_i$$

Thus we have

$$\begin{aligned} < u_i, u_j > &= \frac{1}{\sigma_i \sigma_j} v_i^T A^T A v_j \\ &= \frac{\lambda_j}{\sigma_i \sigma_j} v_i^T v_j \\ &= 0 \end{aligned}$$

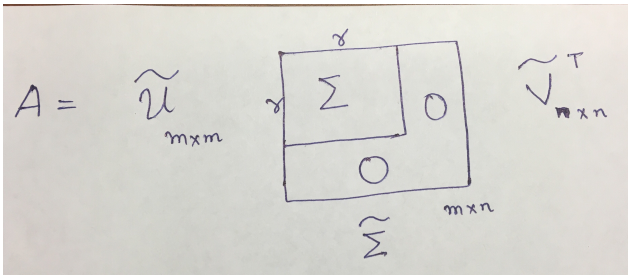
Moreover, $< u_i, u_i >= 1$. It follows that $u_i^\top A = \sigma_i v_i^\top$. Hence, u_i and v_i are left and right singular vectors corresponding to singular values σ_i, respectively. □

4.1.3 Compact form and Full version of SVD

Proposition 4.2 shows that $A = U\Sigma V^T$ for $\Sigma \in \mathbb{R}^{r\times r}$, $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{r\times n}$. Notice that $r \leq \min\{m, n\}$ (convince yourself it is indeed the case). Such a decomposition is known as the the singular value decomposition of A in **compact form**.

Full version of SVD$\longrightarrow$ Take $(u_1, u_2, ...u_r)$ and concatenate it to get $\widetilde{U} = (u_1, u_2, ...u_m)$ which forms an orthonormal basis for $\mathbb{R}^m$. Similarly $\widetilde{V} = (v_1, v_2, ...v_n)$ is obtained which is an orthonormal basis for $\mathbb{R}^n$.

$$\widetilde{U} \in \mathbb{R}^{m\times m},\ \ \widetilde{V} \in \mathbb{R}^{n\times n},\ \ \widetilde{U}^T\widetilde{U} = \widetilde{U}\widetilde{U}^T = \mathbb{I}_{m\times m},\ \ \widetilde{V}^T\widetilde{V} = \widetilde{V}\widetilde{V}^T = \mathbb{I}_{n\times n}.$$



4.1.4 Norms of A

1. Frobenius Norm of A : $\|A\|_F^2 \triangleq \langle A, A\rangle = \sum_{ij} A_{ij}^2$.
2. Spectral Norm of A : $\|A\|_2 \triangleq \sigma_1$ where σ_i are s.t. $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$; $\sigma_{r+1} = ... = \sigma_n = 0$.

Note: One can show that $\|A\|_F^2 = \sum_{j=1}^n \sigma_j^2$. To see this, note that $\|A\|_F^2 = \sum_{i=1}^m \|A_i\|_2^2$ and $\sigma_j^2 = \|Av_j\|_2^2$ for $j = 1, \ldots, n$. Hence, $\sum_{j=1}^n \sigma_j^2$ equals the sum of the squared length of the projections of $\{A_i\}_{i=1}^m$ to the space spanned by $\{v_1, \ldots, v_n\}$, which is $\mathbb{R}^n$. Clearly, the projection of A_i to $\mathbb{R}^n$ is itself. Hence, $\sum_{i=1}^m \|A_i\|_2^2 = \sum_{j=1}^n \sigma_j^2$.

4.1.5 Spectral relaxation of k-means problem:

Recalling from the spectral relaxation of k-means:

$$\max \langle XX^T, Y\rangle$$

$$s.t.\ \ (1)\ \ Y = \sum_{i=1}^k U_i U_i^T$$

(2) $\|U_i\|_2 = 1 \ \forall i$ and $U_i \perp U_j \ \forall i \neq j$.

Proposition 4.3. *Suppose* $XX^T = \sum_{i=1}^{n} \lambda_i \widetilde{U}_i \widetilde{U}_i^T$ *such that* $\lambda_1 \geq \lambda_2 ... \geq \lambda_n$. *Then the optimal solution* $Y^* = \sum_{i=1}^{k} \widetilde{U}_i \widetilde{U}_i^T$.

Proof.

$$\begin{aligned}\langle XX^T, Y \rangle &= \sum_{i=1}^{k} \langle XX^T, U_i U_i^T \rangle \\ &= \sum_{i=1}^{k} U_i^T XX^T U_i \\ &= \sum_{i=1}^{k} \left\| X^T U_i \right\|_2^2\end{aligned}$$

Thus, the spectral relaxations of k-means aims to find the k-dimensional, best-fit subspace for rows of $X^\top$. Hence, the optimal solutions are given by the top k right singular vectors of $X^\top$, or equivalently, the top k left singular vectors of X, or equivalently, the top k eigenvectors of $XX^\top$ (proved by Theorem 4.1). □

Here is another way to view the spectral-relaxations of k-means.

Given a data matrix X and a partition $S = (S_1, S_2, , , S_k)$ where $C(S_a) = \frac{1}{|S_a|} \sum_{i \in S_a} X_i$ is the cluster center of the cluster S_a, the cost function associated is given as:

$$C(S) = \sum_{a=1}^{k} \left\{ \sum_{i \in S_a} \|X_i - C(S_a)\|_2^2 \right\}$$

Let W denote the subspace given by $span\{C(S_1), C(S_2), ...C(S_k)\} \subseteq \mathbb{R}^d$.

Then: Cost function associated with the partition S :

$$\begin{aligned}C(S) &\geq \sum_{i=1}^{n} (dist(X_i, W))^2 \\ &\geq \sum_{i=1}^{n} (dist(X_i, V))^2 \\ &= \sum_{j=k+1}^{r} \sigma_j^2,\end{aligned}$$

where the fist inequality follows from the definition of the $dist$; $V = (v_1, v_2, ...v_k)$ are the top-k right singular vectors of the data matrix

$$X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix},$$

and hence the second inequality follows from that V is a k-dimensional best-fit subspace of $\{X_i\}_{i=1}^n$; the last inequality follows because

$$\begin{aligned}\sum_{i=1}^{n}(dist(X_i,V))^2 &= \sum_{i=1}^{n}\|X_i\|_2^2 - \sum_{j=1}^{k}\|Xv_j\|_2^2 \\ &= \sum_{j=1}^{n}\sigma_j^2 - \sum_{j=1}^{k}\sigma_j^2.\end{aligned}$$

This method is a relaxation of the k-means problem, since a lower bound on the cost function $C(S)$ is obtained: $C(S) \geq \sum_{j=k+1}^{r}\sigma_j^2$.

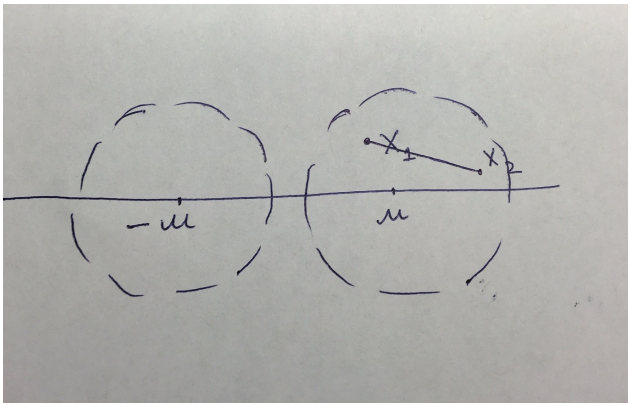
4.2 Spectral clustering under Gaussian mixture model

In this section, we derive the spectral clustering from a model-based perspective. We shall focus on a special case of Gaussian mixture model with **2 clusters**, i.e., $k = 2$. The first cluster center is taken to be $\mu_1 = \mu$ and the second cluster center is $\mu_2 = -\mu$. The same idea applies to more general settings.

Let

$$X \triangleq \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^{n\times d}$$

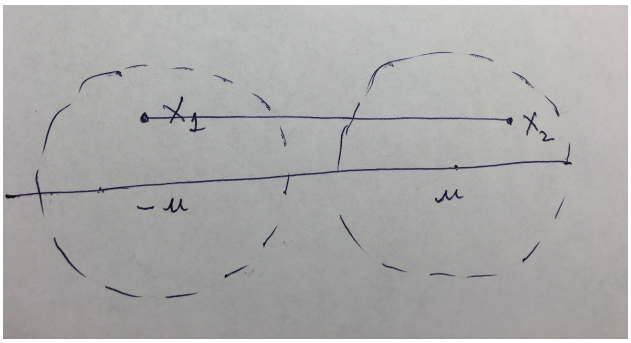
be the data matrix ($X_i \in \mathbb{R}^d$), where $X_i \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\mathcal{N}(\mu, \sigma^2\mathbb{I}_{d\times d}) + \frac{1}{2}\mathcal{N}(-\mu, \sigma^2\mathbb{I}_{d\times d})$.



Suppose X_1, X_2 are 2 data points from the same cluster with mean μ i.e. with the same cluster center μ. Then the expected squared L_2 distance between the two points is calculated as:

$$\begin{aligned}E[\|X_1 - X_2\|_2^2] &= E[\|(X_1 - \mu) - (X_2 - \mu)\|_2^2] \\ &= E[\|X_1 - \mu\|_2^2] + E[\|X_2 - \mu\|_2^2] \\ &= 2\sigma^2 d\end{aligned}$$

Now suppose X_1, X_2 are from 2 different clusters say $X_1 \sim \mathcal{N}(-\mu, \sigma^2\mathbb{I}_{d\times d})$ and $X_2 \sim \mathcal{N}(\mu, \sigma^2\mathbb{I}_{d\times d})$.



Then the expected squared L_2 distance between the two points in this case is :

$$\begin{aligned} E[\| X_1 - X_2 \|_2^2] &= E[\| (X_1 + \mu) - (X_2 - \mu) - 2\mu \|_2^2] \\ &= 4 \| \mu \|_2^2 + 2\sigma^2 d \end{aligned}$$

Here the additional term of $4 \| \mu \|_2^2$ captures the separation between the two cluster centers. Intuitively speaking, if the cluster center separation $\|\mu\|^2$ is much larger than the typical deviation $\sigma^2 d$, then two data points are far away if they are from two different clusters, and close by otherwise. This suggests a simple algorithm to cluster data points by thresholding the pairwise distances.

4.3 Simple Thresholding Algorithm

Given $\{X_i\}_{i=1}^n$ and a **threshold value** τ. This algorithm computes $\|X_i - X_j\|_2$. If $\|X_i - X_j\|_2 \leq \tau$ then X_i, X_j are assigned to the same cluster. Else they are assigned to different clusters. There could be instances when the assignment is inconsistent and at those times the algorithm fails.

As we point out earlier, the algorithm works fine if the separation between the clusters (i.e. $4 \| \mu \|_2^2$) is much larger than the typical deviations. The following HW question rigorously proves this.

HW question: Let $\tau_0 = 2\sigma \left(\sqrt{d} + \sqrt{2}(1+\epsilon)^{1/4} d^{1/4} \log^{1/4} n \right)$ for a small constant $\epsilon > 0$. If $\|\mu\|_2 \geq \tau_0$ and $\tau \in [\tau_0, 2\|\mu\|_2 - \tau_0]$, then the simple thresholding algorithm outputs the correct clustering with probability at least $1 - ne^{-\epsilon^2 d/4}$.

Notice that the sufficient condition for the simple thresholding algorithm to succeed needs the cluster center separation $\|\mu\|_2$ to scale as $\sqrt{d}$, while often implementation of clustering of data points is done on very high dimensions. Therefore one would like to get rid of the dependence on dimension d.

Idea to get rid of dependence on d: Suppose there is a genie who reveals the true cluster center μ to the algorithm. Then each of the data points can be projected on the 1-dimensional subspace spanned by μ; this reduces the effective dimension of the problem. Note that the projection of X_i on the subspace spanned by μ is given by $\langle X_i, \mu \rangle$.

If X_i lies in the first cluster (mean -μ), then $\langle X_i, \mu \rangle$ follows $\mathcal{N}(-||\mu||_2^2, \sigma^2)$ and if it is in the second cluster (with mean μ) then $\langle X_i, \mu \rangle$ follows $\mathcal{N}(\|\mu\|_2^2, \sigma^2)$. To see this, suppose that X_i is in the first cluster. Then we have $\langle X_i, \mu \rangle = -\|\mu\|_2^2 + \langle Z_i, \mu \rangle$, where $Z_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{d \times d})$.

HW question: If μ is known and $\|\mu\|_2 \geq 2\sigma\sqrt{2\log(n)}$, then the simple thresholding algorithm applied on $\langle X_i, \mu \rangle$ with a proper choice of threshold τ outputs the correct clustering with high probability. (Recall that n is the total number of data points).

In order to get rid of the dependence of the algorithm on d, it was assumed that the cluster centers are somehow known, but in reality the cluster center μ, i.e., the direction in which X_i needs to be projected, is not known. In such a case, we might want to try *random projections*, i.e., projecting data points to some random directions.

4.3.1 Random projections

Originally $X_i \in \mathbb{R}^d$ and the goal is to project these X_i on $\mathbb{R}^l$, where $1 \leq l < d$.

Take a matrix $A \in \mathbb{R}^{l \times d}$ s.t. $1 \leq l < d$. Suppose $A_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d})$ and let $X_i' = AX_i \in \mathbb{R}^l$, which represents the data point after projection. Similarly, let $\mu' = A\mu \in \mathbb{R}^l$, which represent the centers after projection.

Conditional on μ' and A : X_i' is approximately distributed as $\frac{1}{2}\mathcal{N}(\mu', \sigma^2 \mathbb{I}_{l \times l}) + \frac{1}{2}\mathcal{N}(-\mu', \sigma^2 \mathbb{I}_{l \times l})$.

Now the separation between the cluster centers is given as:

$$\begin{aligned}
\mathbb{E}\left[\| \mu'\|_2^2 \right] &= \mathbb{E}\left[\|A\mu\|_2^2 \right] \\
&= E[\sum_{j=1}^{l} \langle A_{j.}, \mu \rangle^2] \\
&= \sum_{j=1}^{l} E[\langle A_{j.}, \mu \rangle^2] \\
&= \sum_{j=1}^{l} \frac{\|\mu\|_2^2}{d} \\
&= \frac{l \|\mu\|_2^2}{d}
\end{aligned}$$

$A_{j.}$ is the j^{th} row of A. It is seen that cluster center separation shrinks by a factor of $\frac{l}{d}$.

For threshold algorithm : $\|\mu'\|_2 \geq 2\sigma\sqrt{l}$ (after projection).

$$\begin{aligned}
&\iff \sqrt{\tfrac{l}{d}} \|\mu\|_2 \geq 2\sigma\sqrt{l} \\
&\iff \|\mu\|_2 \geq 2\sigma\sqrt{d}
\end{aligned}$$

$\implies$ Random projection does not help in getting ride of dependency on d. In the next lecture, we

will see that spectral projection, which projects data points to the top singular vectors of X, will help.

Chapter 5

Analysis of Spectral Clustering

Outline

- Review of Singular Value Decomposition
- Spectral Clustering under Gaussian Mixture Model (continued from previous lecture)
- Analysis of spectral clustering

5.1 Review of Singular Value Decomposition (SVD)

Recall from previous lecture that we give singular value decomposition of a matrix A as $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$ where $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_n$ are singular values and $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$ are corresponding left and right singular vectors respectively. Below we present a summary of some of the results related to geometric interpretation of SVD that we discussed in previous lecture:

- We can interpret leading right singular vector (v_1) of A as the best-fit vector for rows of A. It is also the leading eigenvector of $A^T A$. Leading singular value (σ_1) can be viewed as sum of the length of the projections of rows of A onto the linear subspace spanned by v_1 i.e. $\|Av_1\|_2 = \sigma_1$.
- Previous result extends to higher dimensions i.e. best-fit $k-$dimensional subspace for rows of A is given by $span\{v_1, v_2, \ldots, v_k\}$ where

$$\begin{aligned} v_1 &\in \arg \max_{\|v\|_2=1} \|Av\|_2^2 \\ v_2 &\in \arg \max_{\substack{v \perp v_1 \\ \|v\|_2=1}} \|Av\|_2^2 \\ &\vdots \\ v_k &\in \arg \max_{\substack{v \perp v_1, \ldots v \perp v_{k-1} \\ \|v\|_2=1}} \|Av\|_2^2 \end{aligned}$$

- Collection of $v_1, v_2, \ldots v_k$ can be chosen as the top-k eigenvectors of $A^T A$.
- u_i's are defined as $u_i = \frac{Av_i}{\sigma_i}$. This combined with the previous property implies that $u_i \perp u_j$ if $i \neq j$ and $\|u_i\|_2 = 1$.

- $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ for some $1 \leq r \leq n$ assuming $\sigma_{r+1} = \cdots = \sigma_n = 0$. In this case, let row$(A)$denote the row space of A. Then row$(A) = \text{span}\{v_1, v_2, \ldots, v_r\}$.
- Frobenius norm ($\|A\|_F$) is defined as $\|A\|_F^2 = \sum_{j=1}^{n} \sigma_j^2 = \sum_{j=1}^{n} \|Av_j\|_2^2 = \sum_{j=1}^{m} \|A_{i.}\|_2^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2$ where $A_{i.}$ is ith row of A. Second equality holds because $\sigma_j = \|Av_j\|_2$ and third equality holds because of the previous property.

5.2 Spectral Clustering under Gaussian Mixture Model

Recall that if we know the cluster mean μ a priori then we can project our sample points to $span\{\mu\}$. This can help us in reducing the dimension of the problem (in the simple example of 2 clusters we can reduce $d-$ dimensions to $1-$dimension). However, in general we don't have prior knowledge of μ. As discussed in the previous lecture, we could try a random projection but we showed that it doesn't help. Here, we'll discuss another projection scheme called "Spectral projection".

5.2.1 Spectral Projection

We'll start with our basic example of 2 clusters centered at μ and $-\mu$ and variance σ. Then, we'll extend the model to more general case.

Idea We have been given below information:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

$$X_i \overset{i.i.d.}{\sim} \frac{1}{2}\mathcal{N}(\mu, \sigma^2 I_{d\times d}) + \frac{1}{2}\mathcal{N}(-\mu, \sigma^2 I_{d\times d})$$

Based on this we can say that

$$\mathbb{E}[X] = \begin{matrix} \text{First Cluster} \\ \\ \\ \\ \text{Second Cluster} \\ \\ \\ \end{matrix} \begin{cases} \begin{bmatrix} -\mu^T \\ -\mu^T \\ \vdots \\ -\mu^T \end{bmatrix} \\ \begin{bmatrix} \mu^T \\ \vdots \\ \mu^T \end{bmatrix} \end{cases} = \begin{matrix} \text{Left Singular Vector} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{cases} \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \end{cases} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{cases} \mu^T \end{cases} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \text{Right Singular vector} \\ \\ \\ \\ \\ \\ \\ \end{matrix}$$

Observe that points from first cluster contribute $-\mu^T$ in first expected value matrix and points from second cluster contribute μ^T in first expected value matrix. This matrix can be further decomposed into vectors of $\{-1, 1\}^n$ and μ^T which can be treated as left singular vector and right singular vector (upon normalization) of $\mathbb{E}[X]$ respectively. This gives us intuition that if X is "close" to $\mathbb{E}[X]$ then we would expect the leading right singular vector of X to be close to μ. Note that $\mathbb{E}[X]$ is rank 1 matrix however X may not be rank 1. Now suppose,

$$X = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

$$Xv_1 = \sigma_1 u_1$$

If X is close to $\mathbb{E}[X]$ then u_1 would be close to $\begin{bmatrix} -1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$. We can treat the problem of clustering X_i as problem of clustering u_1 and this gives us an algorithm.

5.2.2 Spectral clustering algorithm for $k = 2,\ \mu_1 = -\mu,\ \mu_2 = \mu$

1. Compute the leading left singular vector of X say it is given by u_1.
2. If
   (a) $u_{1,i} < 0$ assign X_i to the first cluster
   (b) $u_{1,i} > 0$ assign X_i to the second cluster
   (c) $u_{1,i} = 0$ assign X_i an arbitrarily chosen cluster

We can easily generalize our results for a general clustering problem following Gaussian mixture model with k clusters centered at $\mu_1, \mu_2, \ldots, \mu_k$ respectively. We can again check that,

$$\mathbb{E}[X] = \begin{matrix} \text{First Cluster} \\ \text{Second Cluster} \\ \\ \text{k-th Cluster} \end{matrix} \begin{Bmatrix} \begin{bmatrix} \mu_1^T \\ \vdots \end{bmatrix} \\ \begin{bmatrix} \mu_2^T \\ \vdots \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \mu_k^T \\ \vdots \end{bmatrix} \end{Bmatrix} = \text{Left Singular Vectors} \begin{Bmatrix} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \mu_1^T \\ \mu_2^T \\ \vdots \\ \mu_k^T \end{bmatrix} \end{Bmatrix} \text{Right Singular vectors}$$

Column i of the matrix containing left singular vectors (up to normalization) acts as an indicator vector ($\mathbf{1}_{S_i}$) for cluster i. The matrix itself is known as "membership matrix". It is easy to extend our previous algorithm to deal with general case.

5.2.3 Spectral clustering algorithm in general case

1. Compute SVD of X i.e. $X = \sum_{i=1}^{r} \sigma_i u_i v_i^T$.
2. $U = [u_1, u_2, \ldots u_k] \in \mathbb{R}^{n \times k}$.
3. Run $k-$means on the rows of U.

Clustering U is easy if U is close to membership matrix.
Note:

- We are treating $[\mu_1^T, \mu_2^T, \ldots, \mu_k^T]^T$ as right singular vectors up to normalization. However, they may not be orthogonal to each other. Our assumption still works because we are only interested in the space spanned by them.
- Recall our spectral relaxation of k-means problem:

$$\widehat{Y} = \arg\max < XX^T, Y > \ \textit{such that}\ Y = \sum_{i=1}^{r} w_i w_i^T,\ \|w_i\|_2 = 1, w_i \perp w_j \forall\ i \neq j$$

Optimal solution for above is $Y^* = \sum_{i=1}^{k} u_i u_i^T = UU^T$ where $u_1, \ldots, u_k$ are top k left singular vectors of X and $U = [u_1, u_2, \ldots, u_k]$. Notice how U appears in the spectral clustering algorithm as well.

5.3 Analysis of Spectral Clustering

Algorithms mentioned in previous section depend on our assumption that X is close to $\mathbb{E}[X]$. In this section, we'll try to quantify this closeness. We will be using Davis-Kahan's $\sin -\theta$ theorem to analyze spectral clustering. But before we move there, we'll define some notation. Lets say we have two matrix A and B such that

$$B = A + \Delta$$

where Δ is called perturbation. Suppose that A and B have a decomposition which is similar to SVD and is given by :

$$A = E \begin{bmatrix} A_0 & 0 \\ 0 & A_1 \end{bmatrix} G = \begin{bmatrix} E_0 & E_1 \end{bmatrix} \begin{bmatrix} A_0 & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} G_0^T \\ G_1^T \end{bmatrix}$$

$$B = F \begin{bmatrix} B_0 & 0 \\ 0 & B_1 \end{bmatrix} H = \begin{bmatrix} F_0 & F_1 \end{bmatrix} \begin{bmatrix} B_0 & 0 \\ 0 & B_1 \end{bmatrix} \begin{bmatrix} H_0^T \\ H_1^T \end{bmatrix}$$

where

$A \in \mathbb{R}^{m \times n}, E \in \mathbb{R}^{m \times m}, G \in \mathbb{R}^{n \times n}, E_0 \in \mathbb{R}^{m \times k}, E_1 \in \mathbb{R}^{m \times m-k}, G_0 \in \mathbb{R}^{k \times n}, G_1 \in \mathbb{R}^{n-k \times n}$

$B \in \mathbb{R}^{m \times n}, F \in \mathbb{R}^{m \times m}, H \in \mathbb{R}^{n \times n}, F_0 \in \mathbb{R}^{m \times k}, F_1 \in \mathbb{R}^{m \times m-k}, H_0 \in \mathbb{R}^{k \times n}, H_1 \in \mathbb{R}^{n-k \times n}$

$A_0 \in \mathbb{R}^{k \times k}, A_1 \in \mathbb{R}^{m-k \times n-k}, B_0 \in \mathbb{R}^{k \times k}, B_1 \in \mathbb{R}^{m-k \times n-k}$

assume

$EE^T = E^T E = I_{m \times m}$

$GG^T = G^T G = I_{n \times n}$

$F^T F = FF^T = I_{m \times m}$

$H^T H = HH^T = I_{n \times n}$

clearly

$A = E_0 A_0 G_0^T + E_1 A_1 G_1^T$

$B = F_0 B_0 H_0^T + F_1 B_1 H_1^T$

In our case, we can view $A = \mathbb{E}[X]$ and $B = X$. Our goal would be to define a distance $d(E_0, F_0)$ between E_0 and F_0 and upper bound it as a function of Δ. Davis-Kahan's $\sin -\theta$ theorem helps us in doing that. But before we move to actual theorem we'll define some specific distances and look into their properties.

5.3.1 Projection distance

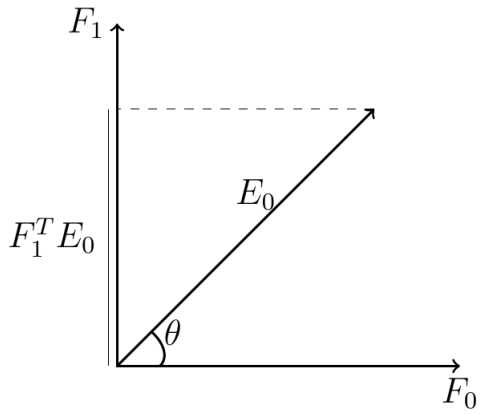
Definition 5.1.

$$d_p(E_0, F_0) \triangleq ||E_0 E_0^T - F_0 F_0^T||_2$$

Lemma 5.1. $d_p(E_0, F_0) = \|F_1^T E_0\|_2 = \|E_0^T F_1\|_2$

Proof. Left for homework.

To get the intuition behind above lemma we can take a simple example where E_0 and F_0 are one dimensional and we know that $F_1 \perp F_0$. Hence, the arrangement looks like below:



It is easy to see that $\|E_0 E_0^T - F_0 F_0^T\|_2 = \|F_1^T E_0\|_2 = \sin\theta$. Notice, how we can denote projection distance in terms of $\sin\theta$. We'll generalize this notion and present a way to view the projection distance in terms of principal angles. Let,

$$E_0^T F_0 = U \cos\Theta V^T$$

where

$$\Theta = \begin{bmatrix} \theta_1 & & & \\ & \theta_2 & & \\ & & \ddots & \\ & & & \theta_k \end{bmatrix}$$

and

$$\cos\Theta = \begin{bmatrix} \cos\theta_1 & & & \\ & \cos\theta_2 & & \\ & & \ddots & \\ & & & \cos\theta_k \end{bmatrix}$$

with

$$0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k \leq \frac{\pi}{2}$$

We can do this because $E_0 F_0$ are basis with singular values less than or equal to 1. Also, note that $U, V \in O(k)$ where $O(k)$ is set of $k \times k$ orthonormal matrices. In our one dimensional example above $E_0^T F_0 = \cos\theta$.

Lemma 5.2.

$$\|F_1^T E_0\|_2 = \|\sin\Theta\|_2 = \sin\theta_k$$

Proof.

$$\begin{aligned}
\|F_1^T E_0\|_2 &= \|E_0^T F_1 F_1^T E_0\|_2^{\frac{1}{2}} \\
&= \|E_0^T (I - F_0 F_0^T) E_0\|_2^{\frac{1}{2}} \\
&= \|E_0^T E_0 - E_0^T F_0 F_0^T E_0\|_2^{\frac{1}{2}} \\
&= \|I_{k\times k} - U\cos\Theta V^T V \cos\Theta U^T\|_2^{\frac{1}{2}} \\
&= \|I_{k\times k} - U\cos^2\Theta U^T\|_2^{\frac{1}{2}} \\
&= \|I_{k\times k} - \cos^2\Theta\|_2^{\frac{1}{2}} \\
&= \|\sin^2\Theta\|_2^{\frac{1}{2}} \\
&= \|\sin\Theta\|_2 \\
&= \sin\theta_k
\end{aligned}$$

Second equality holds because $FF^T = I_{m\times m}$, third and fourth equalities use SVD of $E_0^T F_0$ and fifth and sixth equalities hold because left and right multiplication by U and U^T respectively only causes rotation which doesn't affect the spectral norm.

5.3.2 Spectral distance

Definition 5.2 (Spectral distance)**.**

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

The equality holds because we can interpret Q and R as rotation matrices. Let E_0, F_0 be any vectors in $\mathbb{R}^2$, then we only need to multiply one of the two vectors by -1 to get the quantity to be minimized.

Lemma 5.3.

$$d_s(E_0, F_0) = \|2\sin\frac{\Theta}{2}\|_2 = 2\sin\frac{\theta_k}{2}$$

Proof.

$$\begin{aligned}
d_s^2(E_0, F_0) &= \min_{R\in O(k)} \|E_0 - F_0 R\|_2^2 \\
&= \min_{R\in O(k)} \|(E_0 - F_0 R)^\top (E_0 - F_0 R)\|_2 \\
&= \min_{R\in O(k)} \|E_0^\top E_0 - R^\top F_0^\top E_0 - E_0^\top F_0 R + R^\top F_0^\top F_0 R\|_2 \\
&= \min_{R\in O(k)} \|I - R^\top F_0^\top E_0 - E_0^\top F_0 R + I\|_2 \\
&= \min_{R\in O(k)} \|2I - R^\top V\cos\Theta U^\top - U\cos\Theta V^\top R\|_2 \\
&= \min_{R\in O(k)} \|U^\top (2I - R^\top V\cos\Theta U^\top - U\cos\Theta V^\top R) U\|_2 \\
&= \min_{R\in O(k)} \|2I - U^\top R^\top V\cos\Theta - \cos\Theta V^\top R U\|_2
\end{aligned}$$

Let $R' \triangleq V^\top RU$. Since the product of two orthogonal matrices is also an orthogonal matrix, we have $R' \in O(k)$. Next, we bound the quantity $d_s^2(E_0, F_0)$ on the both sides. On the one hand, we have

$$\begin{aligned} d_s^2(E_0, F_0) &= \min_{R' \in O(k)} \|2I - (R')^\top \cos\Theta - \cos\Theta R'\|_2 \\ &\leq \|2I - 2\cos\Theta\|_2 \\ &= 2(1 - \cos\Theta_k) \\ &= 4\sin^2\frac{\theta_k}{2} \end{aligned}$$

The inequality holds by letting R' be a feasible solution, i.e. $I_{k\times k}$. On the another hand, we have

$$\begin{aligned} d_s^2(E_0, F_0) &= \min_{R' \in O(k)} \|2I - (R')^\top \cos\Theta - \cos\Theta R'\|_2 \\ &= \min_{R' \in O(k)} \left(\max_{\|x\|_2 = 1} x^\top (2I - (R')^\top \cos\Theta - \cos\Theta R')x \right) \\ &\geq \min_{R' \in O(k)} x^\top (2I - (R')^\top \cos\Theta - \cos\Theta R')x \\ &\geq \min_{R' \in O(k)} 2 - 2e_k^\top (R')^\top \cos\Theta e_k^\top \\ &= \min_{R' \in O(k)} 2 - 2R'_{kk}\cos\theta_k \\ &= 2 - 2\cos\theta_k \\ &= 4\sin^2\frac{\theta_k}{2} \end{aligned}$$

The second inequality is true by letting $x \triangleq e_k$. □

Corollary 5.1.

$$d_p(E_0, F_0) \leq d_s(E_0, F_0) \leq \sqrt{2} d_p(E_0, F_0)$$

Proof. By Lemma 5.2, we have $d_p(E_0, F_0) = \sin\theta_k = 2\sin\frac{\theta_k}{2}\cos\frac{\theta_k}{2}$. From Lemma 5.3 , we have $d_s(E_0, F_0) = 2\sin\frac{\theta_k}{2}$. Since $0 \leq \theta_k \leq 1$, then $\frac{1}{\sqrt{2}} \leq \cos\frac{\theta_k}{2} \leq 1$. Therefore we have

$$d_p(E_0, F_0) \leq d_s(E_0, F_0) \leq \sqrt{2} d_p(E_0, F_0).$$

□

5.3.3 Davis-Kahan sin-Θ Theorem

Theorem 5.1 (Davis-Kahan sin-Θ Theorem)**.** *Let* $\mathrm{Sval}(A_0)$ *and* $\mathrm{Sval}(B_1)$ *be the set of singular values of* A_0 *and* B_1*, respectively. If* $\mathrm{Sval}(A_0) \subseteq [0, \alpha]$ *and* $\mathrm{Sval}(B_1) \subseteq [\alpha + \delta, \infty)$ *for some* $\alpha \in R$ *and* $\delta > 0$*, then we have*

$$d_p(E_0, F_0) \leq \frac{\|\triangle\|_2}{\delta} \tag{5.1}$$

In the theorem, δ is called the spectral gap. Before going to prove the theorem, we discuss an application of Davis-Kahan sin-Θ Theorem in spectral clustering.

Example 5.1 (Application of D-K sin-Θ theorem in spectral cluster)**.** Recall in the two clusters setting, cluster one centers at $-\mu \in \mathbb{R}^d$, cluster two at $\mu \in \mathbb{R}^d$, the matrix $X \in \mathbb{R}^{n\times d}$ is the data matrix. Let $A \triangleq X$, $B \triangleq \mathbb{E}[X]$, and $\triangle \triangleq X - \mathbb{E}[X]$. Then, we have a SVD of A

$$A = \sigma_1 u_1 v_1^\top + \sum_{i=2}^{r} \sigma_i u_i v_i^\top;$$

and

$$B = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \mu = \begin{bmatrix} \frac{-1}{\sqrt{n}} \\ \frac{-1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{bmatrix} (\sqrt{n}\|\mu\|_2) \frac{\mu^\top}{\|\mu\|_2} \triangleq \beta(\sqrt{n}\|\mu\|_2) \frac{\mu^\top}{\|\mu\|_2},$$

where

$$\beta \triangleq \begin{bmatrix} \frac{-1}{\sqrt{n}} \\ \frac{-1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}}. \end{bmatrix}$$

The goal is to derive an upper bound of the distance between β and singular vector u_1 in term of $X - \mathbb{E}[X]$. We can apply the Davis-Kahan sin-Θ Theorem (Theorem 5.1) ,with $E_0 = \beta$, $F_0 = u_1$, $A_1 = \begin{bmatrix} \sigma_2 & & & \\ & \sigma_3 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix}$ and $B_0 = \sqrt{n}\|\mu\|_2$ and obtain

$$d_p(\beta, u_1) \leq \frac{\|X - \mathbb{E}[X]\|_2}{\delta},$$

where a lower bound of δ is given as below. We need to obtain a upper bound of the singular value set $\{\sigma_2, \ldots, \sigma_r\}$. From the Weyl's Theorem, we know $|\sigma_i(A) - \sigma_i(B)| \leq \|A - B\|_2$. Thus the singular value set $\{\sigma_2, \ldots, \sigma_r\}$ is bounded by $\|\triangle\|_2$. Hence, $\delta \geq \sqrt{n}\|\mu\|_2 - \|\triangle\|_2$ and we have

$$d_p(\beta, u_1) \leq \frac{\|X - \mathbb{E}[X]\|_2}{\delta} \leq \frac{\|\triangle\|_2}{\sqrt{n}\|\mu\|_2 - \|\triangle\|_2}.$$

We need one more lemma to prove the Davis-Kahan sin-Θ Theorem.

Lemma 5.4. *Let $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{m\times m}$, $X \in \mathbb{R}^{n\times m}$ and $Y \in \mathbb{R}^{m\times n}$. Assume $\|P\|_2 \leq \alpha$ and $\|Q^{-1}\|_2 \leq \frac{1}{\alpha+\delta}$ from some $\alpha \in \mathbb{R}_+$ and $\delta \in \mathbb{R}_+$. Let $C \triangleq XQ - PY$, then we have*

$$\|C\|_2 \geq (\alpha + \delta)\|X\|_2 - \alpha\|Y\|_2$$

Proof. First, we have $\|C\|_2 = \|XQ - PY\|_2 \geq \|XQ\|_2 - \|PY\|_2$ by the subadditivity of a norm.

Then, we derive a lower bound of $\|XQ\|_2$:

$$\begin{aligned}\|X\|_2 &= \|XQQ^{-1}\|_2 \\ &\leq \|XQ\|_2\|Q^{-1}\|_2 \\ &\leq \|XQ\|_2\frac{1}{\alpha+\delta},\end{aligned}$$

where the second inequality holds because for any two matrices A, B, $\|AB\|_2 \leq \|A\|_2\|B\|_2$. Thus, $\|XQ\|_2 \geq (\alpha+\delta)\|X\|_2$. We also have a upper bound of $\|PY\|_2 \leq \|P\|_2\|Y\|_2 \leq \alpha\|Y\|_2$. Hence, $\|C\|_2 \geq (\alpha+\delta)\|X\|_2 - \alpha\|Y\|_2$.

□

Proof of Davis-Kahan sin-Θ Theorem. Recall

$$\begin{aligned}A &= \begin{bmatrix}E_0 & E_1\end{bmatrix}\begin{bmatrix}A_0 & 0\\ 0 & A_1\end{bmatrix}\begin{bmatrix}G_0^\top \\ G_1^\top\end{bmatrix} \\ B &= \begin{bmatrix}F_0 & F_1\end{bmatrix}\begin{bmatrix}B_0 & 0\\ 0 & B_1\end{bmatrix}\begin{bmatrix}H_0^\top \\ H_1^\top\end{bmatrix} \\ \triangle &= B - A\end{aligned}$$

Then since $E, F \in O(m)$ and $G, H \in O(n)$

$$\begin{aligned}E_0^\top \triangle H_1 &= E_0^\top(B-A)H_1 \\ &= E_0^\top BH_1 - E_0^\top AH_1 \\ &= E_0^\top F_1 B_1 - A_0 G_0 H_1\end{aligned}$$

Let $E_0^\top F_1$ be X, B_1 be Q, A_0 be P, G_0H_1 be Y, by Lemma 5.4, we have

$$\|\triangle\|_2 \geq \|E_0^\top \triangle H_1\|_2 \geq (\alpha+\delta)\|E_0^\top F_1\|_2 - \alpha\|G_0^\top H_1\|_2.$$

Similarly, we have

$$\|\triangle\|_2 \geq \|F_1^\top \triangle G_0\|_2 \geq (\alpha+\delta)\|G_0^\top H_1\|_2 - \alpha\|E_0^\top F_1\|_2.$$

Let $t_1 = \|G_0^\top H_1\|_2$ and $t_2 = \|E_0^\top F_1\|_2$. Thus, $t_1 \leq \frac{\alpha t_2 + \|\triangle\|_2}{\alpha+\delta}$ and $t_2 \leq \frac{\alpha t_1 + \|\triangle\|_2}{\alpha+\delta}$. Therefore $\max\{t_1, t_2\} \leq \frac{\|\triangle\|_2}{\delta}$. By Lemma 5.1, $d_p(E_0, F_0) \leq \frac{\|\triangle\|_2}{\delta}$.

□

Chapter 6

Concentration Inequalities

Outline

- Recap of Spectral Clustering
- Concentration inequalities

6.1 Spectral Clustering (cont'd.)

Recap: Under the Gaussian mixture clustering with two clusters of cluster centers given by μ and $-\mu$, respectively, we observe a data matrix $A \in \mathbb{R}^{n \times d}$.

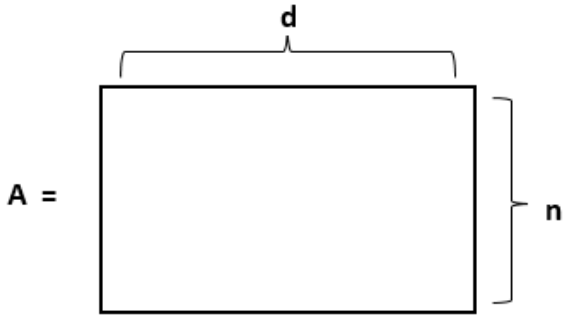


Figure 6.1: Dimension of A.

After proper arrangement of rows and columns, the mean of A has the following decomposition:

$$E[A] = \begin{bmatrix} -1 \\ -1 \\ -1 \\ . \\ . \\ . \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu^T = \begin{bmatrix} \frac{-1}{\sqrt{n}} \\ \frac{-1}{\sqrt{n}} \\ \frac{-1}{\sqrt{n}} \\ . \\ . \\ . \\ \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n}} \end{bmatrix} (\sqrt{n}\|\mu\|_2) \frac{\mu^T}{\|u\|_2}$$

Define $\Delta = A - E[A]$.
Let u_1 denote the leading (left) singular vector of A and let $\bar{u}_1$ denote the leading (left) singular vector of $\mathbb{E}[A]$. Notice that $\sqrt{n}\bar{u}_1$ is a vector in $\{\pm\}^n$.

Applying Davis-Kahan sin-θ theorem, we have

$$d_p(u_1, \bar{u}_1) \leq \frac{\|\Delta\|_2}{\sqrt{n}\|\mu\|_2 - \|\Delta\|_2}$$

Since $\sigma_2(E[A]) = 0$, by Weyl's theorem we get tgat

$$\sigma_2(A) \leq \sigma_2(E[A]) + \|\Delta\|_2 = \|\Delta\|_2.$$

Hence we have,

$$\begin{aligned} d_s(u_1, \bar{u}_1) &\leq \sqrt{2} \cdot d_p(u_1, \bar{u}_1) \\ &\leq \sqrt{2}\frac{\|\Delta\|_2}{\sqrt{n}\|\mu\|_2 - \|\Delta\|_2} \end{aligned}$$

where $d_s(u_1, \bar{u}_1) = \min\{\|u_1 - \bar{u}_1\|_2, \|u_1 + \bar{u}_1\|_2\}$.
Recall that in spectral clustering,

$$\widehat{u}_1 \triangleq Sign(u_1). \tag{6.1}$$

Without loss of generality, assume that $d_s(u_1, \bar{u}_1) = \|u_1 - \bar{u}_1\|_2$. If not we can take $\widetilde{u}_1 = -u_1$, then $d_s(\widetilde{u}_1, \bar{u}_1) = \|\widetilde{u}_1 - \bar{u}_1\|_2$ and clustering through the sign of $\widetilde{u}_1$ is equivalent to clustering through the signs of u_1.

Our goal is to bound the number of misclassified data points. We have,

$$\begin{aligned} d_H(\sqrt{n}\bar{u}_1, \widehat{u}_1) &\triangleq \sum_{i=1}^{n} 1\{\sqrt{n}\bar{u}_{1,i} \neq \widehat{u}_{1,i}\} \quad \text{(number of different coordinates)} \\ &\leq \sum_{i=1}^{n} \left(\sqrt{n}\bar{u}_{1,i} - \sqrt{n}u_{1,i}\right)^2 \\ &= n\|\bar{u}_1 - u_1\|_2^2 \\ &= n \cdot d_s^2(u_1, \bar{u}_1), \end{aligned}$$

where the second inequality holds because if $\sqrt{n}\bar{u}_{1,i} \neq \widehat{u}_{1,i}$, then $\sqrt{n}\bar{u}_{1,i}$ and $\sqrt{n}u_{1,i}$ have different signs, and thus $|\sqrt{n}\bar{u}_{1,i} - \sqrt{n}u_{1,i}| \geq |\sqrt{n}\bar{u}_{1,i}| = 1$.

Therefore,

$$\frac{1}{n}d_H(\sqrt{n}\bar{u}_1, \widehat{u}_1) \leq \frac{2\|\Delta\|_2^2}{\left(\sqrt{n}\|u\|_2 - \|\Delta\|_2\right)^2} \tag{6.2}$$

The expression in the left hand side of the inequality above can be interpreted as the fraction of misclassified data points. Inequality (6.2) implies that $\frac{1}{n}d_H(\sqrt{n}\bar{u}_1, \widehat{u}_1) \to 0$ if $\frac{\sqrt{n}\ \|u\|_2}{\|\Delta\|_2} \to \infty$.
Notice that A is a random matrix and thus Δ is a random matrix In order to bound $\|\Delta\|_2$, we will next introduce some concentration inequalities results.

6.2 Concentration Inequalities

6.2.1 Markov Inequality

Theorem 6.1 (Markov's inequality)**.** *Given a non negative random variable X with $E[X] < \infty$, we have*

$$P(X \geq t) \leq E[t]/t.$$

The proof of the Markov inequality is left to the reader. Although the Markov inequality is simple, it turns out that it is the "mother" of many concentration inequalities that we are going to derive.

6.2.2 Moment Method

The Markov inequality only involves the first moment of X. When higher moments of X are available, we could derive tighter concentration bounds:

Theorem 6.2 (Moment method)**.** *Suppose X is a random variable with $E[|X - E[X]|^k] < \infty$. Then for any $t > 0$,*

$$P\{|X - E[X]| \geq t\} \leq \frac{E[|X - E[X]|^k]}{t^k}.$$

Recall that the **moment generating function** (MGF) of a random variable X is defined as

$$\begin{aligned}\varphi_x(t) &\triangleq E[e^{tX}] \\ &= E\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right] \\ &= \sum_{k=0}^{\infty} \frac{E[X^k]t^k}{k!},\end{aligned}$$

where we here interchange the summation and expectation (needs to be justified). As we can see, the moment generating function involves the moments $E[X^k]$ of the random variable X for all $k \in \mathbb{N}$. It turns out that we can derive a concentration inequality in terms of the MGF of X.

Theorem 6.3 (Chernoff's Bound)**.** *Suppose X has $\varphi_x(\lambda) < \infty$ for $|\lambda| \leq b$ and $\mathbb{E}[X] = \mu$. Then,*

$$\mathbb{P}[X - \mu \geq t] \leq \exp\left(-\sup_{\lambda \in [0,b]} \{\lambda(t + \mu) - M_X(\lambda)\}\right)$$

$$\mathbb{P}[X - \mu \leq -t] \leq \exp\left(-\sup_{\lambda \in [-b,0]} \{\lambda(\mu - t) - M_X(\lambda)\}\right)$$

where $M_X(\lambda) \triangleq \log E[e^{\lambda(x-u)}]$ is the log moment generating function of X.

Proof. For any $\lambda \in [0, b]$,

$$\begin{aligned}P\{X - \mu > t\} &= P\{e^{\lambda(X-\mu)} > e^{\lambda t}\} \text{ apply Markov's inequality to get} \\ &\leq e^{-\lambda t} E[e^{\lambda(X-\mu)}] \quad \forall t \geq 0 \\ &= \exp\left(-\left(\lambda t + \lambda\mu - M_X(\lambda)\right)\right),\end{aligned}$$

Minimizing the right hand side of the last display over $\lambda \in [0, b]$ gives the desired bound. The desired bound for $\mathbb{P}[X - \mu \leq -t]$ follows similarly. □

Example 6.1 (Normal random variable)**.** Let $X \sim N(\mu, \sigma^2)$. We have $E[e^{\lambda(X-\mu)}] = e^{\frac{1}{2}\lambda^2\sigma^2}$ and

$$P\{X - \mu \geq t\} \leq \exp\left(-\sup_{\lambda \in [0,\infty)} \{\lambda t - \lambda^2 \frac{\sigma^2}{2}\}\right) = \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

and similarly, $P\{X - \mu \leq -t\} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$. Thus, by the union bound,

$$P\{|X - \mu| \geq t\} \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Recall that $P\{X - \mu \geq t\} = Q(t/\sigma)$, where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy$. In homework 1, we introduce a lower bound to $Q(x)$ as $Q(x) \geq \left(1 - \frac{1}{x^2}\right) \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$. Hence, we have that

$$P\{|X - \mu| \geq t\} \geq 2\left(1 - \frac{\sigma^2}{t^2}\right) \frac{\sigma}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Hence, the tail probability $P\{|X - \mu| \geq t\}$ decays in t as $\exp(-t^2)$. Such a decaying tail is called "Gaussian tail". In the next subsection, we will identify a class of random variables which has a fast decaying tail than Gaussian tail.

6.2.3 Sub-gaussian Random Variables

Definition 6.1 (Sub-gaussian random variable)**.** A random variable X is sub-gaussian with mean μ, if $\exists\ \sigma \in \mathbb{R}$ s.t $E[e^{\lambda(X-\mu)}] \leq e^{\frac{1}{2}\sigma^2\lambda^2}$.

Note: In the definition above if X is normal then $\sigma = \text{sd(X)}$.

Proposition 6.1. *If X is sub-gaussian (μ, σ^2) then,*

$$P\{X - \mu \geq t\} \leq e^{\frac{-t^2}{2\sigma^2}} \tag{6.3}$$

and

$$P\{X - \mu \leq t\} \leq e^{\frac{-t^2}{2\sigma^2}} \tag{6.4}$$

Combining inequalities 6.3 and 6.4 and union bound yields,

$$P\{|X - \mu| \leq t\} \leq 2e^{\frac{-t^2}{2\sigma^2}}.$$

Proof. The proof directly follows from the Gaussian example. □

Note: Proposition 6.1 implies that sub-gaussian random variables decay at least as fast as a Gaussian tail.

Example 6.2 (Rademarker random variable)**.** Let X be a Rademarker random variable defined as

$$X \sim \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2. \end{cases}$$

Then we get,

$$\begin{aligned}
E[e^{\lambda(X-\mu)}] &= E[e^{\lambda X}] \\
&= \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} \qquad \text{(using taylor expansion we get)} \\
&\leq \frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{\lambda^k}{k!} + \frac{(-\lambda)^k}{k!}\right) \quad \text{(the odd terms cancel to give)} \\
&= \sum_{k=even}\frac{\lambda^k}{k!} \\
&= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} \quad \text{(since } (2k)! \geq 2^k\ k!\ \ \forall\ \ k \geq 1) \\
&= 1 + \sum_{k=1}^{\infty}\frac{(\frac{\lambda^2}{2})^k}{k!} \\
&= e^{\frac{\lambda^2}{2}}.
\end{aligned}$$

Hence the Rademarker random variable X is sub-gaussian with $\sigma = 1$.

Example 6.3 (Bounded Random Variables)**.** Suppose X is zero mean random variable, taking values over $[a, b]$. Then X is sub-gaussian with $(\mu = 0, \sigma = b - a)$.

Note: The proof of the claim above involves using the so-called *symmetrization* technique, which consists of creating an independent copy of the random variable X, then using Jensen's inequality to pull out the expectation.

Proof. Since X is a zero mean random variable, $E[e^{\lambda(X-\mu)}] = E[e^{\lambda X}]$.

$$\begin{aligned}
E[e^{\lambda(X-\mu)}] &= E[e^{\lambda(X-E[X'])}] \text{ where } X' \text{ is an independent copy of the random variable } X \\
&= E_X[e^{\lambda E_{X'}(X-X')}] \\
&\leq E_X E_{X'}[e^{\lambda(X-X')}] \text{ by Jensen's inequality} \\
&= E_{XX'}E_{\epsilon}[e^{\lambda\epsilon(X-X')}] \text{ where } \epsilon \text{ is a Rademarker variable independent of } X \text{ and } X' \\
&\leq E_{XX'}[e^{\frac{1}{2}\lambda^2(X-X')^2}] \\
&\leq e^{\frac{1}{2}\lambda^2(b-a)^2} \text{ since } X \text{ and } X' \text{ are bounded.}
\end{aligned}$$

It follows then that X is sub-gaussian with $(\mu = 0, \sigma = b - a)$. □

Note: The above proof based on symmetrization may not yield the smallest value of σ. To see this, recall that for Rademacher random variable, we showed that it is subgaussian with $\sigma = 1$, while the above symmetrization proof only shows that it is subgaussian with $\sigma = 2$. However, for most of time, we care about the scaling instead of the precise constants, the symmetrization technique will be sufficient and very useful.

Recall that if $x_1, \ldots, x_n$ are independent Gaussian, then the sum $\sum_{i=1}^n x_i$ is also Gaussian. The following result generalizes the above property to sub-gaussian random variables.

Proposition 6.2. *Suppose $x_1, \ldots, x_n$ are independent and x_i is sub-gaussian (μ_i, σ_i), then $\sum_{i=1}^n x_i$ is sub-gaussian $(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i)$*

Proof.

$$\begin{aligned}
&E[e^{\lambda(\sum_{i=1}^n x_i - \sum_{i=1}^n \mu_i)}] \\
&= E[e^{\lambda(\sum_{i=1}^n (x_i - \mu_i)}] \\
&= \prod_{i=1}^n E(e^{\lambda(x_i - \mu_i)}) \qquad (x_i's \text{ are independent}) \\
&\leq \prod_{i=1}^n e^{\frac{1}{2}\lambda^2\sigma_i^2} \\
&= e^{\frac{1}{2}\lambda^2(\sum_{i=1}^n \sigma_i^2)}
\end{aligned}$$

□

Proposition 6.2 immediately implies a concentration inequality for sum of independent sub-gaussian random variables, which is known as *Hoeffding's inequality* when they are bounded random variables.

Theorem 6.4 (Hoeffding's inequality)**.** *Suppose $x_1, \ldots, x_n$ are independent and x_i is sub-gaussian(μ_i, σ_i), then*

$$P\{\sum_{i=1}^n (x_i - \mu_i) \geq t\} \leq e^{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}}$$

In particular, if x_i is supported over $[a_i, b_i]$, then

$$P\{\sum_{i=1}^n (x_i - \mu_i) \geq t\} \leq e^{-\frac{t^2}{2\sum_{i=1}^n (b_i - a_i)^2}}$$

The following theorem gives equivalent conditions for a random variable to be sub-gaussian.

Theorem 6.5. *Suppose X is sub-gaussian with mean 0, then the following are equivalent.*

1. *$E[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for a constant σ.*
2. *$\exists$ a constant c and $Z \sim N(0, \tau^2)$, such that*
$$P\{|X| \geq s\} \leq cP\{|Z| \geq s\}, \forall s \geq 0$$
3. *$\exists$ a constant $\theta \geq 0$, such that $E[X^{2k})] \leq \frac{(2k)!}{2^k k!}\theta^{2k}$.*
4. *$E[e^{\frac{\lambda X^2}{2\sigma^2}}] \leq \frac{1}{\sqrt{1-\lambda}}, \forall \lambda \in [0, 1]$*

The proof is left as homework. The second property says that Z decays as fast as a Gaussian tail. The third property yields that

$$\Big(E[X^{2k})]\Big)^{\frac{1}{2k}} \asymp \theta\sqrt{2k}.$$

Motivated by this scaling behavior, we can define a sub-Gaussian norm of X.

Definition 6.2 (Sub-Gaussian norm)**.** Thus sub-gaussian norm of X, denoted by $\|X\|_{\psi_2}$, is defined as

$$\|X\|_{\psi_2} = \sup_{p \geq 1} \frac{(\mathbb{E}\,[X])^{1/p}}{\sqrt{p}}.$$

Note: It can be easily verified that sub-gaussian norm is indeed a valid norm. Since properties 1 and 3 and equivalent, it follows that

$$\|X\|_{\psi_2} \asymp \sigma.$$

6.2.4 Sub-exponential random variables

Although the class of sub-gaussian random variables is quite wide, it leaves out some useful random variables which have tails decaying slower than Gaussian (heavier tails than Gaussian tail). One such an example is a standard exponential random variable.

Example 6.4. Exponential random variable $X \sim \exp(\lambda)$:

$$P\{X \geq s\} = e^{-\lambda s}, \quad \forall s > 0.$$

We see that the tail probability $P\{X \geq s\}$ decays in s exponentially as $\exp(-s)$ as opposed to the Gaussian tail $\exp(-s^2)$. Such a exponential decaying tail is called "exponential" tail.

To cover random variables with exponential tails, we consider a class of sub-exponential random variables which have tails decaying at least as fast as an exponential tail.

Definition 6.3 (Sub-exponential)**.** Random variable X with mean μ is sub-exponential if $\exists\ \nu > 0$ and $\alpha > 0$ such that

$$E[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2\lambda^2}{2}} \text{for all } |\lambda| \leq \frac{1}{\alpha}.$$

Note: The MGF of a sub-exponential random variable X has similar upper bound as in the sub-Gaussian case; the only difference is that the upper bound only holds in a small neighborhood of 0. Clearly, sub-gaussian(μ, σ) is sub-exponential with mean μ, $\nu = \sigma$, and $\alpha = 0$

Example 6.5 (χ^2 distribution)**.** If $Z \sim N(0,1)$ and $X = Z^2$, then X is called χ^2 random variable with $E[X] = E[Z^2] = 1$.

$$\begin{aligned}
&E[e^{\lambda(X-\mu)}] \\
&= E[e^{\lambda(Z^2-1)}] \\
&= e^{-\lambda} E[e^{\lambda Z^2}] \\
&= \begin{cases} e^{-\lambda}\frac{1}{\sqrt{1-2\lambda}} & \text{if } \lambda < \frac{1}{2} \\ \infty & \text{if } \lambda \geq \frac{1}{2}. \end{cases}
\end{aligned}$$

Since $e^{-\lambda}\frac{1}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2}$ when $|\lambda| \leq \frac{1}{4}$, X is sub-exponential with $\nu^2 = 4$, and $\alpha = 4$.

Theorem 6.6 (Sub-exponential tail bound)**.** *Suppose X is sub-exponential (ν, α), then*

$$P\{X \geq \mu + t\} \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha} \\ e^{-\frac{t}{2\alpha}} & \text{if } t > \frac{\nu^2}{\alpha}. \end{cases}$$

Note: Sub-exponential random variables satisfy large deviation inequalities similar to ones for sub-gaussian. The difference is that two tails have to appear here: They have Gaussian tails when the deviation t is relatively small comparing to ν^2/α, and have exponential-tail when t is relatively large.

Proof.

$$\begin{aligned}
&P\{X \geq \mu + t\} \\
&\leq \exp\left(- \sup_{\lambda \in [0, \frac{1}{\alpha}]} \{\lambda t - \log E[e^{\lambda(X-\mu)}]\} \right) \qquad \text{(Chernoff's bound)} \\
&\leq \exp\left(- \sup_{\lambda \in [0, \frac{1}{\alpha}]} \{\lambda t - \frac{1}{2}\lambda^2 \nu^2\} \right) \qquad \text{(X is sub-exponential)}
\end{aligned}$$

Define $g(\lambda, t) = \lambda t - \frac{1}{2}\lambda^2\nu^2$
To get the maximum value of $g(\lambda, t)$, let $\frac{\partial g(\lambda,t)}{\partial \lambda} = t - \lambda\nu^2 = 0$, then we have $\lambda = \frac{t}{\nu^2}$.
Case 1. $\frac{t}{\nu^2} \leq \frac{1}{\alpha}, i.e.,\ 0 \leq t \leq \frac{\nu^2}{\alpha}$

$$\sup_{\lambda \in [0, \frac{1}{\alpha}]} g(\lambda, t) = g(\frac{t}{\nu^2}, t) = \frac{t^2}{2\nu^2}$$

thus

$$\exp\left(- \sup_{\lambda \in [0, \frac{1}{\alpha}]} \{\lambda t - \frac{1}{2}\lambda^2 \nu^2\} \right) = e^{-\frac{t^2}{2\nu^2}}$$

Case 2. $\frac{t}{\nu^2} > \frac{1}{\alpha}, i.e.,\ t > \frac{\nu^2}{\alpha}$

$$\sup_{\lambda \in [0, \frac{1}{\alpha}]} g(\lambda, t) = g(\frac{1}{\alpha}, t) = \frac{t}{\alpha} - \frac{1}{2\alpha^2}\nu^2 = \frac{t}{2\alpha}$$

thus

$$\exp\left(- \sup_{\lambda \in [0, \frac{1}{\alpha}]} \{\lambda t - \frac{1}{2}\lambda^2 \nu^2\} \right) = e^{-\frac{t}{2\alpha}}$$

□

Idea: If I have bounds on moments of X, then can I derive tail bounds? This motivates us to consider the following Bernstein's condition on moments of X.

Definition 6.4 (Bernstein's condition)**.** Given a random variable X with $\mathbb{E}[X] = \mu$ and $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$, we say it satisfies Bernstein's condition with parameter $b > 0$, if $E[|X - \mu|^k] \leq \frac{1}{2}k!\sigma^2 b^{k-2}$.

Proposition 6.3. *If X with $\mathbb{E}[X] = \mu$ and $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$ satisfies Bernstein's condition with parameter $b > 0$. Then X is sub-exponential with $(\sqrt{2}\sigma, 2b)$.*

Proof.

$$
\begin{aligned}
&E[e^{\lambda(X-\mu)}] \\
&= \sum_{k=0}^{\infty} \frac{\lambda^k E[(X-\mu)^k]}{k!} \qquad \text{(Taylor's expansion)} \\
&= 1 + \frac{\lambda^2 E[(X-\mu)^2]}{2!} + \sum_{k=3}^{\infty} \frac{\lambda^k E[(X-\mu)^k]}{k!} \\
&\leq 1 + \frac{\lambda^2}{2!}\sigma^2 + \sum_{k=3}^{\infty} \frac{\lambda^k \sigma^2 b^{k-2}}{2} \qquad \text{(Apply Bernstein's condition)} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (\lambda b)^{k-2} \quad (*)
\end{aligned}
$$

If $\lambda < \frac{1}{b}, i.e., \lambda b < 1$ then $\sum_{k=3}^{\infty}(\lambda b)^{k-2} = \frac{\lambda b}{1-\lambda b}$. Therefore,

$$
\begin{aligned}
(*) &\leq 1 + \frac{\lambda^2\sigma^2}{2}(1 + \frac{\lambda b}{1-\lambda b}) \\
&= 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{1-\lambda b} \\
&\leq e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-\lambda b}} \qquad \text{(Since } 1 + X \leq e^X\text{)} \quad (**)
\end{aligned}
$$

Assume $\lambda \leq \frac{1}{2b}$, then $1 - \lambda b \geq \frac{1}{2}$. It follows that

$$
(**) \leq e^{\lambda^2\sigma^2} = e^{\frac{\lambda^2}{2}(\sqrt{2}\sigma)^2},
$$

so X is sub-exponential with $(\sqrt{2}\sigma, 2b)$. □

Note: Combining the above proposition and Theorem 6.6, we immediately get tail bounds for random variables satisfying the Bernstein's condition. It turns out that we can sharp the tail bounds a bit by directing plugging the upper bound $(**)$ on MGF of X into the Chernoff bound, which gives the so-called Bernstein's inequality.

Theorem 6.7 (Bernstein's inequality)**.** *If X with $\mathbb{E}[X] = \mu$ and $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$ satisfies Bernstein's condition with parameter $b > 0$, then*

$$
P\{|X-\mu| \geq t\} \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \leq \begin{cases} 2e^{-\frac{t^2}{2(c+1)\sigma^2}} & \text{if } t \leq \frac{c\sigma^2}{b} \\ 2e^{-\frac{t}{2(1+1/c)}} & \text{if } t \geq \frac{c\sigma^2}{b}, \end{cases}
$$

where c is a constant.

Proof. By plugging the upper bound $(**)$ on MGF of X into the Chernoff bound, we get that

$$
P\{X-\mu \geq t\} \leq \exp\left(-\sup_{\lambda\in[0,1/b)}\left\{\lambda t - \frac{\lambda^2\sigma^2}{2(1-\lambda b)}\right\}\right) = \exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right),
$$

where the last equality holds because by setting $\lambda = \frac{t}{bt+\sigma^2}$. □

An important example of a random variable satisfying the Bernstein's condition is the bounded random variable.

Example 6.6 (Bounded random variable)**.** Assume $E[X] = \mu$, $E[|X-\mu|^2] = \sigma^2 \le b^2$, and $|X-\mu| \le b$. Then

$$E[|X-\mu|^k]] \le E[|X-\mu|^2]b^{k-2} = \sigma^2 b^{k-2}, \quad \forall k \ge 3.$$

Therefore, X satisfies the Bernstein's condition with parameter b. It follows from Bernstein's inequality that

$$P\{|X-\mu| \ge t\} \le 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \le \begin{cases} 2e^{-\frac{t^2}{2(c+1)\sigma^2}} & \text{if } t \le \frac{c\sigma^2}{b} \\ 2e^{-\frac{t}{2(1+1/c)}} & \text{if } t \ge \frac{c\sigma^2}{b}, \end{cases}$$

Note: It is instructive to compare the Bernstein's inequality with the Hoeffding's inequality for bounded random variables. Recall that by Hoeffding's inequality,

$$P\{X-\mu \ge t\} \le e^{-\frac{t^2}{8b^2}}.$$

Notice that Bernstein's inequality involves both σ and b, while Hoeffding's inequality only involves b. If $t \le \frac{c\sigma^2}{b}$, then Bernstein's inequality shows that X has Gaussian tail $\exp(-t^2/\sigma^2)$, while Hoeffding's inequality shows that X has Gaussian tail $\exp(-t^2/b^2)$. In the case where $\sigma^2 \ll b^2$, Bernstein's inequality gives substantially tighter bounds than Hoeffding's inequality. Intuitively, that is because Bernstein's inequality utilizes the variance information, while Hoeffding's inequality does not not.

Chapter 7

Matrix Concentration Inequalities

Plan ahead:

- Sub-exponential random variable
- Gaussian concentration inequality
- Slepian's comparison inequality and concentration inequality for Gaussian random matrices

7.1 Review of Sub-exponential random variables

Definition 7.1 (Sub-exponential (ν, α))**.** A random variable X with $\mathbb{E}[X] = \mu$ is sub-exponential with parameter (ν, α) if:

$$\mathbb{E}[e^{(\lambda(X-\mu))}] \leq e^{\frac{\nu^2\lambda^2}{2}}; \;\; for \;\; |\lambda| \leq \frac{1}{\alpha}$$

By plugging the above upper bound to MGF of X into Chernoff's bound, and optimizing over the exponent, we get the following tail bounds (The detailed derivation is given in the last lecture).

Sub-exponential tail bounds :

$$\mathbb{P}[X \geq \mu + t] \leq \begin{cases} e^{\frac{-t^2}{2\nu^2}} & \text{if } t \in [0, \frac{\nu^2}{\alpha}] \\ e^{\frac{-t}{2\alpha}} & \text{if } t \geq \frac{\nu^2}{\alpha}. \end{cases}$$

Note: X has either Gaussian tail or exponential tail depending on how large the deviation t is.

Definition 7.2 (Bernstein's condition)**.** Given a random variable with $\mathbb{E}[X] = \mu$ and $var(X) = \sigma^2$. It is said that X satisfies Bernstein's condition (BC) with parameter $b > 0$ if for $k \geq 3$:

$$\mathbb{E}[|X - \mu|^k] \leq \frac{1}{2} k! \sigma^2 b^{(k-2)}$$

Proposition 7.1. *If X satisfies Bernstein's condition with parameter $b > 0$, then X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$. It further follows that*

$$\mathbb{P}[X \geq \mu + t] \leq \begin{cases} e^{\frac{-t^2}{4\sigma^2}} & \textit{if } t \in [0, \frac{\sigma^2}{b}] \\ e^{\frac{-t}{4b}} & O.W \end{cases}$$

Proof. The proof is given in the last lecture. □

We can get a slightly tighter bound by directly invoking the Chernoff's bound.

Proposition 7.2 (Bernstein's inequality)**.** *If X satisfies BC with $b > 0$, then*

$$\mathbb{P}[X \geq \mu + t] \leq e^{\frac{-t^2}{2(\sigma^2+bt)}}.$$

Note: One can easily check that

$$e^{\frac{-t^2}{2(\sigma^2+bt)}} \leq \begin{cases} e^{\frac{-t^2}{4\sigma^2}} & \text{if } t \in [0, \frac{\sigma^2}{b}] \\ e^{\frac{-t}{4b}} & O.W \end{cases}$$

Proof. Recall that in the last lecture, we have showed that if X satisfies BC with parameter $b > 0$, then

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-\lambda b)}\right).$$

Directly plugging in the above bound to Chernoff's bound, and optimizing over the exponent, we get the desired result. □

Example 7.1 (Bounded random variable.)**.** Give a bounded random variable X with $\mathbb{E}[X] = \mu$, $var(X) = \sigma^2$, and $|X - \mu| \leq b$. We have shown in the last lecture that X satisfies BC with parameter b. By Bernstein's inequality,

$$\mathbb{P}[X - \mu \geq t] \leq e^{\frac{-t^2}{2(\sigma^2+bt)}}$$

We can also invoke Hoeffding's inequality to get that

$$\mathbb{P}[X - \mu \geq t] \leq e^{\frac{-t^2}{8b^2}}.$$

Notice that if $0 \leq t \leq \frac{\sigma^2}{b}$ and $\sigma \ll b$, then Bernstein's inequality implies a Gaussian tail with a faster decaying rate than Hoeffding's inequality.

7.1.1 Sum of independent sub-exponential random variables

Recall that in the last lecture, we have shown that sum of independent sub-gaussian random variables is also sub-gaussian. The following theorem gives the similar conclusion for sub-exponential random variables.

Theorem 7.1. *If $X_1 \dots X_n$ are independent random variables where X_i's are sub-exponential (ν_i, α_i) with mean μ_i. Then $\sum_{i=1}^n X_i$ is sub-exponential with (ν_*, α_*), where $\nu_*^2 = \sum_{i=1}^n \nu_i^2$ and $\alpha_* = max_{1 \leq i \leq n} \alpha_i$.*

Proof. Note that

$$\begin{aligned} \mathbb{E}[e^{\lambda \sum_{i=1}^n (X_i - \mu_i)}] &= \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i-\mu_i)}] \\ &\leq \prod_{i=1}^n e^{\frac{\lambda^2\nu_i^2}{2}} \quad \forall |\lambda| < \frac{1}{\alpha_*} \\ &= e^{\frac{\lambda^2 \sum_{i=1}^n (\nu_i^2)}{2}} \quad \forall |\lambda| < \frac{1}{\alpha_*}, \end{aligned}$$

where the first equality holds because X_i's are independent. □

Note: Theorem 7.1 together with sub-exponential tail bounds immediately imply that

$$\mathbb{P}[\sum_{i=1}^{n}(X_i - \mu_i) \geq t] \leq \begin{cases} e^{\frac{-t^2}{2\nu_*^2}} & \text{if } t \in [0, \frac{\nu_*^2}{\alpha_*}] \\ e^{\frac{-t}{2\alpha_*}} & O.W \end{cases}$$

Example 7.2 (χ^2 distribution)**.** . Suppose X_i are i.i.d random variables with $X_i \sim \mathcal{N}(0,1), 1 \leq i \leq n$. Let $Y = \sum_{i=1}^{n} X_i^2 = \|X\|_2^2 \sim \chi^2(n)$. Recall X_i^2 is sub-exponential $(2,4)$ (check last lecture). Then Theorem 7.1 implies that $Y = \sum_{i=1}^{n} X_i^2$ is sub-exponential(ν_*, α_*), where $\nu_*^2 = 4n$ and $\alpha_* = 4$.

$$\implies \mathbb{P}[Y - n \geq t] \leq \begin{cases} e^{\frac{-t^2}{8n}} & \text{if } t \in [0, n] \\ e^{\frac{-t}{8}} & \text{if } t \geq n \end{cases}$$

Note: Recall that in the first homework, we outline a derivation to show that $\mathbb{P}[\|X\|_2^2 \geq \frac{n}{1-\epsilon}] \leq e^{\frac{-\epsilon^2 n}{4}}$ for any $\epsilon \in [0,1]$.

Example 7.3 (Binomial distribution)**.** Suppose X_i are i.i.d random variables with $X_i \sim \text{Bern}(p)$ for $1 \leq i \leq n$. Let $Y = \sum_{i=1}^{n} X_i^2$, then $Y \sim \text{Binom}(n,p)$. Since X_i is a bounded random variable, it satisfies BC with $b = 1$. It follows that X_i is sub-exponential with parameter $(\sqrt{2}\sigma, 2b)$ Hence Y is sub-exponential and thus

$$\mathbb{P}[Y - np \geq t] \leq \begin{cases} e^{\frac{-t^2}{4np(1-p)}} & \text{if } t \in [0, np(1-p)] \\ e^{\frac{-t}{4}} & O.W \end{cases}$$

7.1.2 Application: maximum degree in Erdős - Rényi random graph $G(n,p)$

Given n nodes and for each pair of nodes we connect them with probability p.

Definition 7.3 (Erdős - Rényi random graph $G(n,p)$)**.** Suppose there are n vertices indexed by $[n]$. We generate a random graph G in the following way. For each pair of two vertices i and j, they are connected by an edge independently at random with probability p.

If there is an edge between i and j, then we say i is neighbor of j and j is neighbor of i. Let d_i denote the number of neighbors of node i. Then $d_i \sim Binomial(n-1,p)$. We are interested in deriving a high probability bound to $d_{\max} = \max_{1 \leq i \leq n} d_i$, i.e., to find a threshold τ_n s.t. $d_{\max} \leq \tau(n)$ with probability tending to 1 as $n \longrightarrow \infty$. We assume p is bounded away from 1 and consider the following three cases.

Case 1: $np = \omega(\log(n)) \iff \frac{np}{\log(n)} \to \infty$ as $n \to \infty$. Since $d_i \sim \text{Binom}(n-1,p)$, it follows that

$$\implies \mathbb{P}[d_i - (n-1)p \geq t] \leq \begin{cases} e^{\frac{-t^2}{4(n-1)p(1-p)}} & \text{if } t \in [0, (n-1)p(1-p)] \\ e^{\frac{-t}{4}} & O.W \end{cases}$$

Notice that d_i are identically distributed (but not independent). Using the union bound, we have

$$\begin{aligned} \mathbb{P}[d_{\max} > \tau(n)] &= \mathbb{P}[\max_{1 \leq i \leq n}(d_i) > \tau_n] \\ &\leq \sum_{i=1}^{n} \mathbb{P}[d_i > \tau_n] \\ &= n\mathbb{P}[d_1 > \tau(n)]. \end{aligned}$$

We would like to have $n\mathbb{P}[d_1 > \tau_n] \to 0$ as $n \to \infty$; hence we could choose τ_n so that $\mathbb{P}[d_1 > \tau(n)] \leq \frac{1}{n^{1+\epsilon}}$. Therefore, let

$$e^{\frac{-t^2}{4(n-1)p(1-p)}} = \frac{1}{n^{1+\epsilon}} \iff t = 2\sqrt{(1+\epsilon)(n-1)p(1-p)\log(n)}$$

It remains to check if the above value of t is in $[0, (n-1)p(1-p)]$. This is indeed true because $\frac{np}{\log(n)} \to \infty$ as $n \to \infty$. Therefore,

$$\tau_n = (n-1)p + 2\sqrt{(1+\epsilon)(n-1)p(1-p)\log(n)} \sim (1+o(1))np.$$

Case 2 : $np = a\log(n)$ for a constant $a > 0$. Notice that $p = \frac{alog(n)}{n} \longrightarrow 0$ as $n \longrightarrow \infty$. If we stick to our previous choice of τ_n, i.e., $\tau_n = 2\sqrt{(1+\epsilon)(n+1)p(1-p)log(n)}$, then we have:

$$t \leq (n-1)p(1-p)$$

$$\iff 2\sqrt{(1+\epsilon)log(n)} \leq \sqrt{(n-1)p(1-p)} \sim \sqrt{a\log(n)}$$

$$\iff a > 4(1+\epsilon)$$

Hence, our previous choice of τ_n only works if $a > 4(1+\epsilon)$.

Instead, we look at the sub-exponential tail of Binomial distribution. We need to ensure $\frac{t}{4} \geq (1+\epsilon)\log(n)$. and $t > (n-1)p(1-p) \sim a\log(n)$. Therefore, we choose $t = \max\{4(1+\epsilon), a\}\log(n)$. Hence,

$$Pr\{\max_{1\leq i\leq n} d_i > (n-1)p + t\} \leq \frac{1}{n^\epsilon} \to 0$$

It follows that

$$\tau(n) = (n-1)p + \max\{4(1+\epsilon), a\}\log(n) \sim (a + \max\{4(1+\epsilon, a\})\log(n).$$

Case 3 : $p = \frac{a}{n}$, where a is constant. This case is left as a homework problem.

Finally, we remark that if we use Hoeffding's inequality, we ge that

$$Pr\{d_i - np \geq t\} \leq e^{-\frac{t^2}{2n}}$$

and thus $\tau(n) = (n-1)p + \sqrt{2(1+\epsilon)n\log(n)}$. We see that Bernstein's inequality implies tighter bound than Hoeffding's inequality. The reason is that Hoeffding's inequality does not utilize the variance information.

7.2 Gaussian Concentration Inequality, Slepian Comparison inequality, and Gaussian random matrix

Recall the Gaussian concentration inequality.

Theorem 7.2. *Suppose $X_i \overset{iid}{\sim} N(0,1)$ for $1 \leq i \leq d$. let f to be 1-Lipschitz function $R^d \to R$ with respect to the Euclidean Norm. Then for all $t > 0$,*

$$\mathbb{P}\left[|f(X) - \mathbb{E}\left[f(X)\right]| \geq t\right] \leq 2e^{-\frac{t^2}{2}}$$

Applying the Gaussian concentration inequality, we have the following concentration inequality for the spectrum norm of a Gaussian random matrix.

Theorem 7.3. *Suppose $A \in \mathbb{R}^{n\times d}$ is a Gaussian random matrix with $A_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Then*

$$\mathbb{P}\left[||A||_2 - \mathbb{E}\left[||A||_2\right] \geq t\right] \leq 2e^{-\frac{t^2}{2}}.$$

Proof. It suffices to check that $f(A) \triangleq ||A||_2$ is 1-Lipschitz. Indeed, $||A-B||_2 \leq ||A-B||_F$. □

To derive a high-probability bound on $\|A\|_2$, it remains to bound $\mathbb{E}\left[||A||_2\right]$. We need *Slepian's Comparison Inequality* for this.

Theorem 7.4 (Slepian's comparison inequality)**.** *Consider two centered Gaussian Process $(X_S)_{S\in T}$ and $(Y_S)_{S\in T}$, where T is a given set. Suppose $E[(X_s - X_t)^2] \leq E[(Y_s - Y_t)^2]$ for all $(s,t) \in T \times T$, Then*

$$\mathbb{E}\left[\sup_{s\in T} X_s\right] \leq \mathbb{E}\left[\sup_{s\in T} Y_s\right].$$

Intuitively speaking, $E[(X_s - X_t)^2] \leq E[(Y_s - Y_t)^2]$ implies that Y is more volatile than X, and hence we would expect $\mathbb{E}\left[\sup_{s\in T} Y_s\right]$ is larger than $\mathbb{E}\left[\sup_{s\in T} X_s\right]$.

Applying Slepian's comparison inequality, we have the following bound on $\mathbb{E}\left[\|A\|_2\right]$.

Theorem 7.5. *Suppose $A \in \mathbb{R}^{n\times d}$ is a Gaussian random matrix with $A_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Then $E[||A||_2] \leq \sqrt{n} + \sqrt{d}$.*

Note: By combining Theorem 8.1 and Theorem 8.2, we get that $Pr\{||A||_2 \geq \sqrt{n} + \sqrt{d} + t\} \leq 2e^{-\frac{t^2}{2}}$.

Proof. Recall that

$$||A||_2 = \sup_{||v||_2=1} ||Av||_2 = \sup_{||v||_2=1, ||u||_2=1} u^T A v.$$

Define $X_{uv} \triangleq u^T A v$ for all $\|u\|_2 = 1$ and $\|v\|_2 = 1$. Fix $u, v, \widetilde{u}, \widetilde{v}$ with unit norms. Then

$$\begin{aligned}
E[(X_{uv} - X_{\widetilde{u}\widetilde{v}})^2] &= E[(u^T A v - \widetilde{u}^T A \widetilde{v})^2] \\
&\overset{(a)}{=} E[(\langle u^T v, A\rangle - \langle \widetilde{u}^T \widetilde{v}, A\rangle)^2] \\
&= E[(\langle u^T v - \widetilde{u}^T \widetilde{v}, A\rangle)^2] \\
&= E[\sum_{i,j} A_{i,j}(u^T v - \widetilde{u}^T \widetilde{v})_{i,j}^2] \\
&= E[\sum_{i,j}\sum_{i',j'} A_{i,j}A_{i',j'}(u^T v - \widetilde{u}^T \widetilde{v})_{i,j}(u^T v - \widetilde{u}^T \widetilde{v})_{i',j'}] \\
&= \sum_{i,j} E(A_{i,j}^2)(u^T v - \widetilde{u}^T \widetilde{v})_{i,j}^2 \\
&= ||u^T v - \widetilde{u}^T \widetilde{v}||_F^2 \\
&\leq ||u - \widetilde{u}||_2^2 + ||v - \widetilde{v}||_2^2.
\end{aligned}$$

where (a) holds because $u^T A v =< uv^T, A >= \mathrm{Tr}(vu^T A) = \mathrm{Tr}(u^T A v)$.

Define $Y_{u,v} = \langle u, g\rangle + \langle v, h\rangle$, where $g \sim \mathcal{N}(0, I_{n\times n})$ and $h \sim \mathcal{N}(0, I_{d\times d})$, and g is independent of h. Then

$$\begin{aligned} E[(Y_{uv} - Y_{\widetilde{u}\widetilde{v}})^2] &= E[(\overbrace{< u-\widetilde{u}, g > + < v-\widetilde{v}, h >}^{N(0,\, ||u-\widetilde{u}||_2^2 + ||v-\widetilde{v}||_2^2)})^2] \\ &= ||u-\widetilde{u}||_2^2 + ||v-\widetilde{v}||_2^2, \end{aligned}$$

where the second equality holds because $\langle g, u-\widetilde{u}\rangle \sim N(0, ||u-\widetilde{u}||_2^2)$ and $\langle h, v-\widetilde{v}\rangle \sim N(0, ||v-\widetilde{v}||_2^2)$, and they are independent to each other. Hence, we have $E[(X_{uv} - X_{\widetilde{u}\widetilde{v}})^2] \le E[(Y_{uv} - Y_{\widetilde{u}\widetilde{v}})^2]$.

Furthermore,

$$\begin{aligned} E[\sup_{||v||_2=1, ||u||_2=1} Y_{uv}] &= E[\sup_{||v||_2=1, ||u||_2=1} \{\langle u, g\rangle + \langle v, h\rangle\}] \\ &= E[||g||_2 + ||h||_2], \\ &\le \sqrt{E[||g||_2^2]} + \sqrt{E[||h||_2^2]} \\ &\le \sqrt{n} + \sqrt{d} \end{aligned}$$

where the second equality holds because the optimal u and v are given by $u = \frac{g}{||g||_2}$ and $v = \frac{h}{||h||_2}$.

The desired conclusion follows by applying Slepian's comparison inequality.

□

Chapter 8

Spectral clustering and Laplacian matrices

Outline

- Concentration inequality for Gaussian random matrix
- Spectral clustering algorithm based on Laplacian matrix

8.1 Concentration inequality for Gaussian random matrix (cont'd)

8.1.1 Brief review

Theorem 8.1. *Given* $A_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$*, we have*

$$Pr\left\{\mid \|A\|_2 - \mathbb{E}\left[\|A\|_2\right] \mid > t\right\} \leq 2e^{\frac{-t^2}{2}}.$$

It implies that $\|A\|_2$ *has a Gaussian tail behavior.*

Theorem 8.2. *Given Gaussian random matrix* $A \in \mathbb{R}^{n\times d}$*, we have*

$$\mathbb{E}\left[\|A\|_2\right] \leq \sqrt{n} + \sqrt{d}.$$

Theorem 8.1 and Theorem 8.2 imply that

$$Pr\left\{\|A\|_2 \geq \sqrt{n} + \sqrt{d} + t\right\} \leq 2e^{-t^2/2}.$$

Therefore, when the deviation t gets bigger, the probability becomes smaller. In particular, let $t = t_n$, and assume that $t_n \to \infty$ as $n \to \infty$. Then we get that with probability converging to 1,

$$\|A\|_2 \leq \sqrt{n} + \sqrt{d} + t_n,$$

In plain english, $\|A\|_2$ is likely to be smaller than $\sqrt{n} + \sqrt{d} + t_n$.

Remark Is $\sqrt{n} + \sqrt{d} + t_n$ a tight upper bound? The following simple analysis shows that the upper bound is tight up to constant factors. Consider $\|A\|_2$ as the sum of squared length of projections of rows of A to best-fit vector v. That is,

$$\begin{aligned} \|A\|_2^2 &= \max_{\|v\|_2=1} \|Av\|_2^2 \\ &\geq \|Ae_1\|_2^2, \end{aligned} \tag{8.1}$$

where $e_1 \in \mathbb{R}^d$ is a unit vector whose first element is equal to 1 and the other elements are all zero. In fact, we can use any other basis vector e_i. Then $\|Ae_1\|_2^2 = \|A_{\cdot 1}\|_2^2$, which is the squared length of the first column of A, and follows chi-squared distribution with n degree of freedom ($\chi^2(n)$) (Recall that chi-squared random variable is expressed as the summation of squared normal random variables). In addition, we have shown in the previous lectures that chi-squared random variable is also highly concentrated on its mean, and thus $\|A\|_2 \gtrsim \sqrt{n}$ with high probability. Note that the same argument is applied for A^T. Hence, we have

$$\|A\|_2 \gtrsim \max\left\{\sqrt{n}, \sqrt{d}\right\}$$

with high probability. Thus, we see that the upper and lower bounds of $\|A\|_2$ differ by at most a constant factor. In plain english, $\|A\|_2$ is roughly on the same order as the length of the row or column vector of A. Notice that we have assumed that A has independent entries. It is easy to see that without such independence assumption, $\|A\|_2$ could be much larger than the length of the row or column vector of A, for example, when all rows or columns of A are identical.

8.2 Spectral clustering under Gaussian mixture model (revisited)

In the previous class, we learned that the fraction of misclassified nodes is bounded as,

$$\text{(Fraction of misclassified nodes)} \leq \frac{2\|\Delta\|_2^2}{\left(\sqrt{n}\|\mu\|_2 - \|\Delta\|_2\right)^2},$$

where $\Delta = X - \mathbb{E}[X]$. Note that $X \in \mathbb{R}^{n \times d}$ is the data matrix, and μ is the cluster center under Gaussian mixture model. Notice that $\Delta_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. Then by plugging in the previous result, we have

$$\|\Delta\|_2 \leq \sigma\left(\sqrt{n} + \sqrt{d} + t_n\right),$$

and σ behaves as a scaling factor. From this result, if $\sqrt{n}\|\mu\|_2 \gg \sigma(\sqrt{n} + \sqrt{d} + t_n)$, then the fraction of misclassified nodes goes to zero with high probability. Hence, we say spectral clustering achieves approximate cluster recovery under sufficient condition $\sqrt{n}\|\mu\|_2 \gg \sigma(\sqrt{n} + \sqrt{d} + t_n)$.
If the dimension of dataset, d, is proportional to n, the number of data points (i.e. $d = \alpha n$, where α is a fixed constant), we have

$$\|\mu\|_2 \gg \sigma\left(1 + \sqrt{\alpha} + \frac{t_n}{\sqrt{n}}\right).$$

Note that t_n goes to ∞ arbitrarily slowly as $n \to \infty$. Then $\frac{t_n}{\sqrt{n}}$ goes to zero as $n \to \infty$. As a result, we see that the cluster separation $\|\mu\|_2$ does not have to scale with dimension d anymore. In contrast, recall that in the previous lectures, we have shown that a naive thresholding algorithm needs $\|\mu\|_2$ to scale with $\sqrt{d}$.

8.3 Spectral clustering based on Laplacian matrix

Plan

- In this lecture, we will introduce weighted similarity graph and its Laplacian matrices for spectral clustering
- In the next lecture, we will introduce stochastic block model, which is a simple probabilistic model to generate similarity (unweighted) graph.

8.3.1 Motivation for spectral clustering based on Laplacian matrix

Recall that we have learnt spectral clustering based on data matrix $X \in \mathbb{R}^{n\times d}$; it involves singular value decomposition of X or eigenvalue decomposition of $XX^\top$ and $X^\top X$.

Here, we introduce a more general setup for spectral clustering. Define a similarity matrix $S \in \mathbb{R}^{n\times n}$ between every pair of data points. The component S_{ij} characterizes similarity between data points $X_{i\cdot}$ and $X_{j\cdot}$, where $X_{i\cdot}$ denotes the ith row of X.

Example 8.1. 1. $S = XX^T$.

Each component S_{ij} is the inner product of x_i and x_j. Notice that the spectral clustering algorithm discussed in the previous lectures deals with this similarity matrix.

2. Define S with

$$S_{ij} = \exp\left\{-\frac{1}{2\sigma^2}\|X_{i\cdot} - X_{j\cdot}\|_2^2\right\}.$$

This function has the form of Gaussian density function, and it is called as *Gaussian kernel function* with σ. When the two data points $X_{i\cdot}$ and $X_{j\cdot}$ differ a lot, the corresponding similarity score S_{ij} gets smaller, and vice versa.
A crucial parameter is σ: when σ becomes larger, S_{ij} is less sensitive to the difference between x_i and x_j, and vice versa.

The similarity matrix can be equivalently viewed as a similarity graph.

Definition 8.1 (Similarity graph)**.** Given a similarity matrix S, let G denote a weighted graph, where each node i corresponds to a data point i; every pair of two nodes i and j are connected by an edge with edge weight S_{ij}. In this case S is called *weighted* adjacency matrix of graph G.

Note that the defined similarity graph G is a complete graph (every pair of two nodes are connected). Often, we may apply a truncation procedure to sparsify G. There are two possible truncation procedures. Let $A \in \mathbb{R}^{n\times n}$ denote the similarity matrix after truncation:

- $A_{ij} = \mathbf{1}_{S_{ij}\geq\epsilon}$, where ϵ is a given threshold.
- $A_{ij} = S_{ij}\mathbf{1}_{S_{ij}\geq\epsilon}$

Equivalently, by truncating, in graph G, two nodes i and j are connected if and only if $S_{ij} \geq \epsilon$. In the first case, A is a binary matrix, and thus graph G after truncation becomes an unweighted graph.

There are at least two reasons why truncation is used in practical.

1. It can potentially remove a bit of noise;
2. A is sparse and computation of A is faster. For instance, eigenvalue decomposition of A gets faster.

Example 8.2 (A clustering example for truncation)**.** Let us see why truncation could possibly remove noise by considering the following example, depicted in Fig. 8.1. In this example, the data distribution clearly does not come from Gaussian mixture model. There are two possible clusters, one given by the "out" ring, and the other given by the "inner ring". Two clusters have very close cluster centers. Hence, the spectral clustering developed under GMM does not work well any more.

To deal with this, we may apply truncation. Two data points i and j are connected, i.e., $A_{ij} > 0$, if and only if $S_{ij} \geq \epsilon$. By carefully picking the threshold ϵ, one can get that data points in the same cluster are densely connected, while data points in different clusters are loosely connected, even

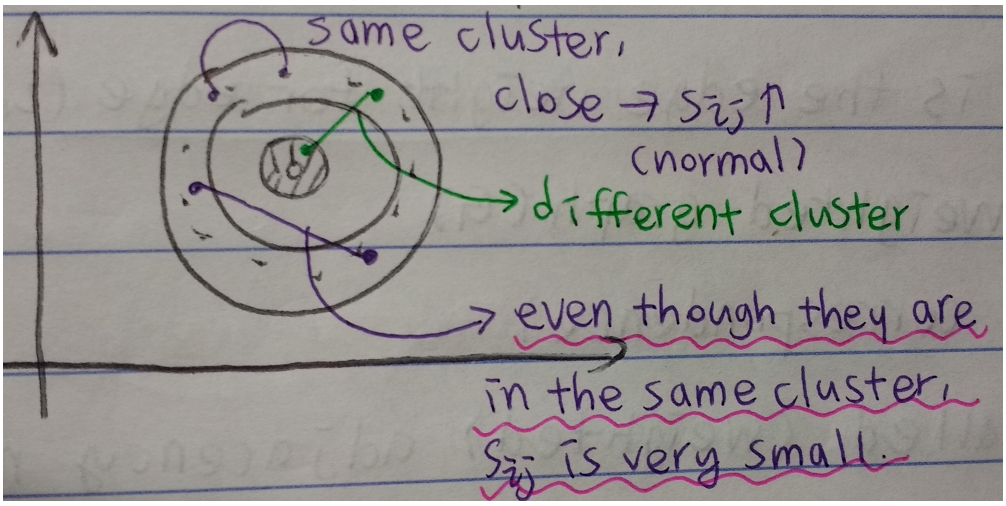


Figure 8.1: An example of two clusters: one consists of data points in the out ring; the other consists of data points in the inner ring.

though not every pair of data points in the same cluster is connected. We call edges which connect data points in two different clusters *cross links*. Fig. 8.2 shows that there are few cross-links after proper truncation.

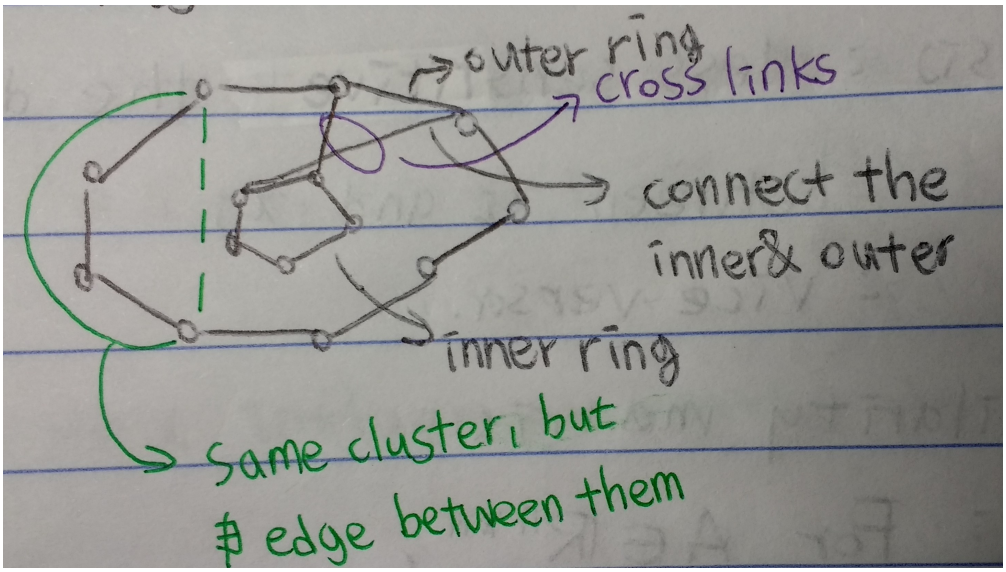


Figure 8.2: Construction of cross links between inner and outer ring

An important question is: can we recover clusters from A? Let us consider two cases:

1. There is no cros link. In this case, we can classifies the inner and outer ring very easily by finding the two connected components.

2. There are a few cross links. In this case, we will use spectral clustering based on *Laplacian matrix of a graph*.

8.3.2 Weighted graph and its Laplacian matrices

Consider a general symmetric matrix $A \in \mathbb{R}^{n\times n}$ with $A_{ii} = 0$ and all $A_{ij} \geq 0$. Let G denote a graph with n nodes, where nodes i and j are connected if and only if $A_{ij} > 0$, and when they are connected, A_{ij} is the edge weight for edge (i, j). Notice that there is a one-to-one correspondence between A

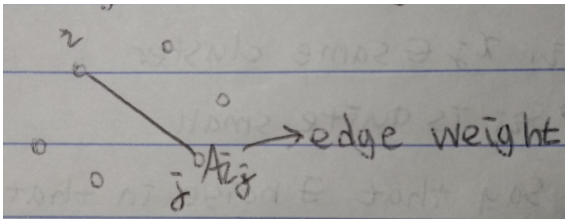


Figure 8.3: Node i and j are connected with edge weight A_{ij}

and G , and A is called (weighted) adjacency matrix of graph G. When A is a binary matrix, there is no weight on the edge and A only represents whether the two points are the neighbors. In this case, A is simply called adjacency matrix of G.

Consider the following diagonal matrix D:

$$D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix},$$

and its ith elements are defined as $d_i = \sum_{j=1}^{n} A_{ij}$ for $i \in \{1, \ldots, n\}$. If $A_{ij} \in \{0, 1\}$, then d_i represents the number of neighbors of node i. Hence d_i is called *degree* of node i, and D is called *degree matrix*.

Definition 8.2. (Laplacian matrix) There are three possible versions of Laplacian matrix.

1. unnormalized version:

$$L = D - A, \qquad \text{(unnormalized)}$$

   where A is a weighted adjacency matrix defined previously.

2. normalized version: Assume that D has no zero diagonal elements. With this D,

$$\begin{aligned} L_n &= D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \\ &= I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \end{aligned} \qquad \text{(normalized)}$$

   where

$$D^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{d_1}} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\sqrt{d_n}} \end{pmatrix}$$

3. random walk version:

$$\begin{aligned} L_{rw} &= D^{-1}L \\ &= I - D^{-1}A = D^{-\frac{1}{2}}L_n D^{\frac{1}{2}} \end{aligned} \tag{randomwalk}$$

Note For undirected graph G, A is symmetric (i.e., $A_{ij} = A_{ji}$). Hence, L and L_n are also symmetric, but L_{rw} may not be symmetric.

Question

1. Why is this called *Laplacian*?

   Recall: For a scalar function $f(x) : \mathbb{R} \to \mathbb{R}$, Laplacian operator is simply defined as,

   $$\nabla^2 f(x) \triangleq \frac{d^2 f(x)}{(dx)^2}. \tag{8.2}$$

   We consider a (unweighted) line graph G with n nodes. Each node is associated with a value x_i, and the values at any two neighboring nodes differ by δx, i.e., $x_{i+1} - x_i = \delta x$ for all $1 \leq i \leq n-1$.

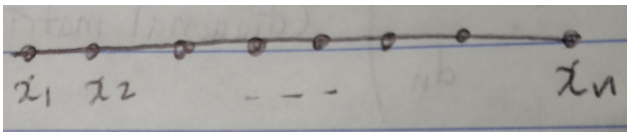


Figure 8.4: Each point in G has two neighbors

   Assume n is large and δx is small. With this notation, we can approximate the derivative of f at x_i and x_{i-1} respectively as,

$$\begin{aligned} f'(x_i) &\approx \frac{f(x_{i+1}) - f(x_i)}{\delta x}, \\ f'(x_{i-1}) &\approx \frac{f(x_i) - f(x_{i-1})}{\delta x}. \end{aligned}$$

   Then the second derivative at x_i can be approximated as,

$$\begin{aligned} \nabla^2 f(x_i) = f''(x_i) &\approx \frac{f'(x_i) - f'(x_{i-1})}{\delta x} \\ &\approx \frac{f(x_{i-1}) + f(x_{i+1}) - 2f(x_i)}{(\delta x)^2}. \end{aligned} \tag{8.3}$$

   Now consider the unnormalized Laplacian matrix $L = D - A$, where A is the adjacency matrix of the line graph G. We have

$$L\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} = \begin{pmatrix} \vdots \\ 2f(x_i) - f(x_{i-1}) - f(x_{i+1}) \\ \vdots \end{pmatrix}, \tag{8.4}$$

since every ith node has $(i-1)$th and $(i+1)$th nodes as its neighbors, so its degree is equal to 2 for $i = 2, \ldots, n$. By comparing (12.2) and (8.4), we can observe that the ith element of (8.4) is approximately $-\nabla^2 f(x_i)(\delta x)^2$.

2. Why is L_{rw} called random walk version?

   To answer for this question, let's see what the random walk is.

Definition 8.3. (Randomwalk on Graph G) We say $X_0, X_1, \ldots, X_t, \ldots$ is random walk on graph G with corresponding A if

$$P(X_{t+1} = j \mid X_t = i) \triangleq \frac{1}{d_i} A_{ij}.$$

In other words, the random walk on undirected graph is a Markov chain defined on the graph, which jumps over the nodes at each time setp, and the probability of which node jumping to only depends on the current node that it is at. Denote P_{ij} as the *transition probability matrix* for random walk. That is,

$$P_{ij} \triangleq \frac{1}{d_i} A_{ij}.$$

Since D is a diagonal matrix, it is easy to see that $L_{\text{rw}} = I - P$. Hence L is called random walk version of Laplacian.

8.3.3 Properties of Laplacian matrices

In this subsection, we characterize some properties of Laplacian matrices.

Proposition 8.1. *Recall $L = D - A$ and $L_n = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. Then, for any $v \in \mathbb{R}^n$, we have*

$$v^\top L v = \frac{1}{2} \sum_{i,j} A_{ij}(v_i - v_j)^2$$

and

$$v^\top L_n v = \frac{1}{2} \sum_{i,j} A_{ij} \left(\frac{v_i}{\sqrt{d_i}} - \frac{v_j}{\sqrt{d_j}} \right)^2.$$

Note: Given any $v \in \mathbb{R}^n$, let v_i denote the value associated with node i. Then, the proposition shows that L and L_n characterize the quadratice difference between node values weighted by the edge weights.

Proof.

$$\begin{aligned} v^\top L v = v^\top (D - A) v &= \sum_i d_i v_i^2 - \sum_{i,j} A_{ij} v_i v_j \\ &= \frac{1}{2} \left(\sum_{i=1}^n d_i v_i^2 - 2 \sum_{i,j} A_{ij} v_i v_j + \sum_{j=1}^n d_j v_j^2 \right) = \frac{1}{2} \sum_{i,j} A_{ij}(v_i - v_j)^2 \end{aligned}$$

□

The following proposition characterizes eigenvalues and eigenvector of Laplacian matrices.

Proposition 8.2. *Consider a* **connected** *graph G, and let L be the unnormalized Laplacian matrix and L_n be the normalized Laplacian matrix. Then*

1. *If G is connected, then $0 = \lambda_1(L_n) < \lambda_2(L_n) \leq \cdots \leq \lambda_n(L_n)$ and similarly, $0 = \lambda_1(L) < \lambda_2(L) \leq \cdots \leq \lambda_n(L)$.*
2. *Let u_1 be the eigenvector corresponding to $\lambda_1(L_n)$. Then $u_{1,i} \propto \sqrt{d_i}$. Similarly, let $\widetilde{u}_1$ be the eigenvector corresponding to $\lambda_1(L)$. Then $\widetilde{u}_1$ is parallel to the all-one vector.*
3. $$\begin{aligned} \lambda_2(L_n) = \min_z \ & \frac{1}{2}\sum_{i,j} A_{ij}(z_i - z_j)^2 \\ s.t. \ & \sum_{i=1}^n d_i z_i = 0 \\ & \sum_{i=1}^n d_i z_i^2 = 1 \end{aligned}$$

Proof. By *Proposition* 8.1, L_n is a positive semidefinite matrix and $u_1^\top L_n u_1 = 0$. Thus, $0 = \lambda_1(L_n) \leq \lambda_2(L_n) \leq \cdots \leq \lambda_n(L_n)$, and u_1 is an eigenvector corresponding to $\lambda_1(L_n)$. Similar conclusions hold for L.

We next show $\lambda_2(L_n) > 0$. It suffices to show that $v^\top L_n v > 0$ for all $v \perp u_1$. We prove it by contradiction. Suppose there exists a $v \perp u_1$ and $v^\top L_n v = 0$. Fix any node $i \neq 1$ in graph G. Since G is connected, then there must exist a path from 1 to i. Suppose the edges on the path are given by $(1 = i_1, i_2), (i_2, i_3), \ldots, (i_{k-1}, i_k = i)$. Then $A_{i_1,i_2} > 0, \ldots, A_{i_{k-1},i_k} > 0$. Thus, $v^\top L_n v = 0$ implies

$$\frac{v_1}{\sqrt{d_1}} = \frac{v_{i_1}}{\sqrt{d_{i_1}}}, \quad \cdots, \quad \frac{v_{i_{k-1}}}{\sqrt{d_{i_{k-1}}}} = \frac{v_{i_k}}{\sqrt{d_{i_k}}}.$$

Therefore $\frac{v_1}{\sqrt{d_1}} = \frac{v_i}{\sqrt{d_i}}$. Since i is arbitrarily chosen, we have that for all i,

$$\frac{v_i}{\sqrt{d_i}} \equiv \frac{v_1}{\sqrt{d_1}},$$

which contradicts the assumption that $v \perp u_1$.

Finally, we give a characterization of $\lambda_2(L_n)$. By definition, $\lambda_2(L_n) = \min_{\|v\|_2=1, v\perp u_1} v^\top L_n v$. We have

$$v \perp u_1 \Leftrightarrow \sum_{i=1}^n v_i u_{1i} = 0 \Leftrightarrow \sum_{i=1}^n v_i \sqrt{d_i} = 0.$$

Let $z_i = \frac{v_i}{\sqrt{d_i}}$. Then $v \perp u_1 \Leftrightarrow \sum_i z_i d_i = 0$. Moreover $\|v\|_2 = 1 \Leftrightarrow \sum_{i=1}^n d_i z_i^2 = 1$. Also, by *Proposition* 8.1, $v^\top L_n v = \frac{1}{2}\sum_{i,j} A_{ij}(z_i - z_j)^2$.

□

Next, we will derive an analogy of the Proposition 8.2 for a graph G with k connected components. Let $S_1, \ldots, S_k$ denote the k connected components of G. Then $[n] = S_1 \cup S_2 \cdots \cup S_k$, $S_i \cap S_j = \emptyset$, and $A_{S_i,S_j} = 0$ for $i \neq j$, where $A_{S,T} \triangleq (A_{ij})_{(i,j)\in S\times T}$. Thus, the normalized Laplacian matrix for G has the following diagonal-block structure:

$$L_n = \begin{pmatrix} L_n^{(1)} & & \\ & \ddots & \\ & & L_n^{(k)} \end{pmatrix};$$

where $L_n^{(j)}$ is the normalized Laplacian matrix for j^{th} component of the graph G. The following proposition immediately follows.

Proposition 8.3.

$$\{\lambda_i(L_n)\}_{i=1}^n = \left\{\lambda_i(L_n^{(1)})\right\}_{1\le i\le |S_1|} \bigcup \left\{\lambda_i(L_n^{(2)})\right\}_{1\le i\le |S_2|} \bigcup \cdots \left\{\lambda_i(L_n^{(k)})\right\}_{1\le i\le |S_k|}.$$

In particular, we have

$$\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0 < \lambda_{k+1} \le \cdots \le \lambda_n$$

and there are k orthogonal eigenvectors corresponding to zero eigenvalue, which are given by $u_1, \ldots, u_k$ with

$$u_{\ell,i} = c_\ell \mathbf{1}_{i\in S_\ell}\sqrt{d_i}, \quad 1 \le \ell \le k;$$

where c_ℓ is a normalization constant dependent on ℓ. In particular,

$$u_1 = \begin{pmatrix} c_1\sqrt{d_1} \\ \vdots \\ c_1\sqrt{d_{|S_1|}} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad u_2 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ c_2\sqrt{d_{|S_1|+1}} \\ \vdots \\ c_2\sqrt{d_{|S_1|+|S_2|}} \\ 0 \\ \vdots \end{pmatrix} \cdots \qquad u_k = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ c_k\sqrt{d_{n-|S_k|+1}} \\ \vdots \\ c_k\sqrt{d_n} \end{pmatrix}.$$

Let $U_{n\times k} \triangleq [u_1, u_2 \cdots, u_k]$. Then,

$$U_{i\cdot} = \begin{cases} (c_1\sqrt{d_i}), 0, 0, \cdots) & i \in S_1 \\ (0, c_2\sqrt{d_i}, 0, \cdots) & i \in S_2 \\ \cdots\cdots & \\ (0, \cdots, 0, c_k\sqrt{d_i}) & i \in S_k \end{cases}$$

We can normalize $U_{i\cdot}$ to have a unit norm:

$$Z_{i\cdot} \triangleq \frac{U_{i\cdot}}{\|U_{i\cdot}\|_2} = \begin{cases} (1, 0, 0, \cdots) & i \in S_1 \\ (0, 1, 0, \cdots) & i \in S_2 \\ \cdots\cdots & \\ (0, \cdots, 0, 1) & i \in S_k \end{cases}$$

Note

- By *Proposition* 8.3, the number of zero eigenvalues is exactly the same as the number of connected components.

- One can recover the k-connected components by clustering n rows $\{U_{i\cdot}\}_{i=1}^{n}$.
- Since k zero eigenvalues implies k multiplicity, eigenvectors are not uniquely determined. In particular,

  Claim 8.1. *Let $R \in O(k,k)$, and let $[\widetilde{u}_1, \cdots, \widetilde{u}_k] = [u_1, \cdots, u_k]R$. Then, $\{\widetilde{u}_1, \ldots, \widetilde{u}_k\}$ are also orthogonal eigenvectors of L_n corresponding to zero eigenvalue.*

Chapter 9

Spectral clustering in graphs

Outline

- Review of spectral clustering with Laplacian matrix
- Spectral Clustering Algorithm
- Spectral Clustering to identify k-dense clusters
- Analysis of spectral clustering with k-dense clusters
- Probablistic models for generating graphs

9.1 Review of spectral clustering with Laplacian matrix

In the last lecture we discussed that we can write normalized Laplacian matrix for a graph with k-connected components (Figure 9.1) as a block diagonal matrix.

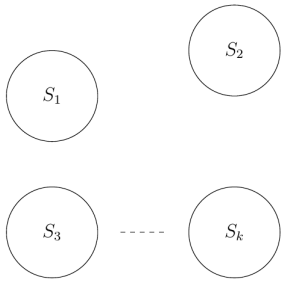


Figure 9.1: A k-connected graph (Each circle is a connected component)

In particular, by properly ordering rows/columns, the normalized Laplacian matrix for the above graph can be written in the form:

$$L_n = \begin{bmatrix} L_n^{(1)} & & & \\ & L_n^{(2)} & & \\ & & \ddots & \\ & & & L_n^{(k)} \end{bmatrix} \tag{9.1}$$

where $L_n^{(i)}$ is a normalized Laplacian for connected component S_i. As discussed in the last lecture, we get one zero eigenvalue from each of k connected components and correspondingly k-orthogonal eigenvectors can be written as $\{u_1, u_2, \ldots, u_k\}$. More precisely,

$$u_1 = \begin{bmatrix} c_1\sqrt{d_1} \\ c_1\sqrt{d_2} \\ \vdots \\ c_1\sqrt{d_{|S_1|}} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad u_k = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c_k\sqrt{d_{n-|S_k|+1}} \\ \vdots \\ \vdots \\ c_k\sqrt{d_n} \end{bmatrix}, \tag{9.2}$$

where c_ℓ is the normalization constant for cluster ℓ.

We can construct a matrix $U_{n\times k} = [u_1\ u_2\ \ldots\ u_k]$, where the rows of matrix U are given by

$$U_{i.} = \begin{cases} (c_1\sqrt{d_i}, 0, 0, \ldots, 0) & \text{if } \ i \in S_1 \\ \vdots & \\ (0, 0, \ldots, 0, c_k\sqrt{d_i}) & \text{if } \ i \in S_k \end{cases} \tag{9.3}$$

We can further normalize $U_{i.}$ and define $Z_{i.}$ as

$$Z_{i.} \triangleq \frac{U_{i.}}{\|U_{i.}\|_2} = \begin{cases} e_1^T & \text{if } \ i \in S_1 \\ \vdots & \\ e_k^T & \text{if } \ i \in S_k, \end{cases} \tag{9.4}$$

where $e_1, \ldots, e_k$ are standard basis vectors for $\mathbb{R}^k$. Interestingly, we observe that Z has k distinct rows, and $Z_{i.}$ is e_ℓ if i is from S_ℓ. Hence, we can easily recover the k connected components by putting identical rows of Z into one component.

9.2 Spectral Clustering algorithm

The analysis in the previous section suggests the following spectral clustering algorithm for recovering k connected components.

Proposition 9.1. *If G has k connected components, then Algorithm 2 exactly outputs these k-connected components.*

Algorithm 2 Spectral clustering based on normalized Laplacian matrix

1: Input: $A \in \mathbb{R}^{n\times n}$ or equivalently graph G, and k.
2: Output: k-partition of $[n]$.
3: Construct $L_n = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.
4: Compute $\{u_1, u_2, \ldots, u_k\}$ as k orthogonal eigenvectors corresponding to 0 eigenvalue of L_n
5: Let $U_{n\times k} = [u_1\ u_2\ \ldots\ u_k]$ and normalize rows of U by

$$Z_{i.} \triangleq \frac{U_{i.}}{\|U_{i.}\|_2} \tag{9.5}$$

6: Cluster the rows of Z using k-means and output the k-partition.

Proof by plot. Consider a graph with 2 connected components i.e. $k = 2$. The rows of Z are located at e_1 or e_2, depending on which connected component it belongs to. See figure 9.2

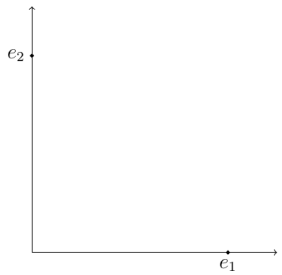


Figure 9.2: Spectral clustering with $k = 2$ in ideal case

Then if we pick e_1 and e_2 as initial cluster centers, then k-means exactly recover the 2 connected components. □

Note: There is a caveat. The eigenvectors corresponding to zero eigenvalue are not unique, i.e., the set $\{u_1, u_2, \ldots, u_k\}$ is not unique. We can have any

$$[\widetilde{u}_1, \widetilde{u}_2, \ldots, \widetilde{u}_k] \triangleq [u_1, u_2, \ldots, u_k]R \tag{9.6}$$

as eigenvectors corresponding to zero eigenvalue, where $R \in \mathbb{O}^{k\times k}$ is an orthogonal rotation matrix i.e. $RR^T = R^TR = I_{k\times k}$. More compactly,

$$\widetilde{U} = UR \tag{9.7}$$
$$\widetilde{Z} = ZR \tag{9.8}$$
$$\tag{9.9}$$

where rows of $\widetilde{Z}$ are given by:

$$\widetilde{Z}_{i.} = \begin{cases} e_1^TR & \text{if } \ i \in S_1 \\ \vdots & \\ e_k^TR & \text{if } \ i \in S_k \end{cases} \tag{9.10}$$

Notice that $\widetilde{Z}$ still has k distinct rows. All the rows of $\widetilde{Z}$ are exactly located at $e_1^\top R, \ldots, e_k^\top R$. In other words, the cluster centers are only rotated by rotation matrix R. We can still easily exactly recover the k-connected components by putting identical rows of $\widetilde{Z}$ into same component.

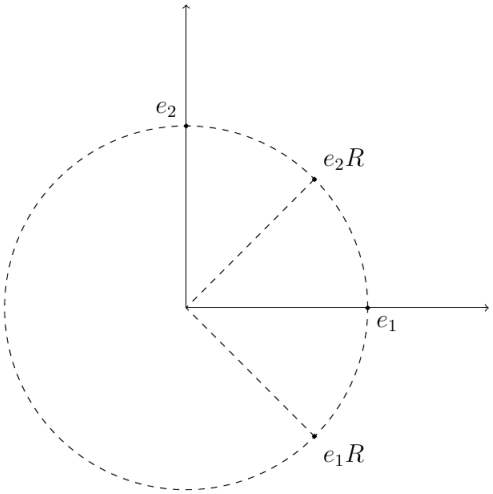


Figure 9.3: Spectral clustering with $k = 2$ in when eigenvectors are rotated by R

Note: We can see that we may not need the k-means to exactly recover the k-connected components. Instead, we can simply put those i's with same row vectors $\widetilde{Z}_{i,}$ into same cluster. Since rows of $\widetilde{Z}$ corresponding to nodes from cluster ℓ are exactly located at $e_\ell^\top R$ for $1 \leq \ell \leq k$, this simple procedure succeeds. As we will see in the next section, in the case where the graph G has k-dense clusters instead of k-connected components, rows of $\widetilde{Z}$ corresponding to nodes from cluster ℓ are clustered around but not exactly located at $e_\ell^\top R$, and in this case, we will use k-means to cluster rows of $\widetilde{Z}$.

9.3 Spectral clustering to identify k-dense clusters

Till now we have been working with the ideal case where we had a graph G with k connected components but now we'll deal with more challenging problem where G has k-dense clusters instead of k-connected components. The case with k-dense clusters can be viewed as a perturbation of the ideal case with k-connected components. In particular,

1. there exist cross-links between clusters
2. nodes inside a cluster may not be connected

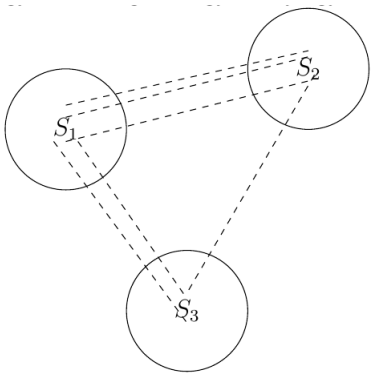


Figure 9.4: 3-dense clusters (each circle is a dense cluster but may not be connected)

Note: What happens when we apply the previous spectral algorithm to find k-dense cluster? If there exist few cross-links and clusters are well-connected, we expect that all the rows of $\widetilde{Z}$ are clustered around $e_1^\top R, \ldots, e_k^\top R$. In the simple case with $k = 2$, rows of $\widetilde{Z}$ corresponding to nodes from cluster 2 center around $e_2^\top R$, and rows of $\widetilde{Z}$ corresponding to nodes from cluster 1 center around $e_1^\top R$. An illustration is shown below (see also reference [Mon15]):

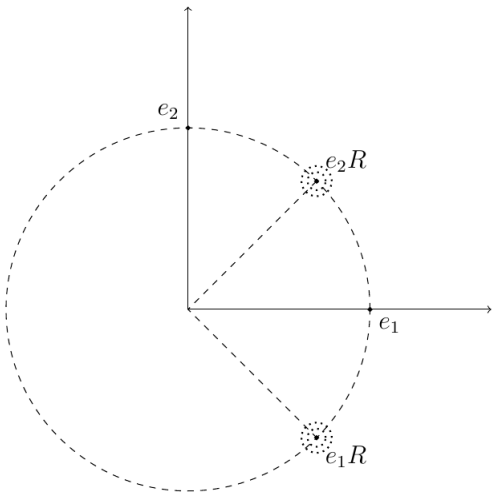


Figure 9.5: Spectral clustering to identify k-dense cluster (illustration)

Certainly, the amount of deviation from centers $e_1^\top R$ and $e_2^\top R$ will depend on how many cross-links there are and how well the clusters are connected inside. In the next section, we quantify such dependency using Davis-Kahan sin-θ theorem.

9.4 Analysis of spectral clustering with k-dense cluster

Given a graph G with k dense clusters. Suppose it is "close" to a graph $\bar{G}$ with exactly k connected components. More formally, let L_n and $\bar{L}_n$ denote the normalized Laplacian matrices of G and $\bar{G}$, respectively. We can measure the closeness between G and $\bar{G}$ by using the difference between their

Laplacian matrices, i.e.,

$$L_n = \bar{L}_n + \Delta \tag{9.11}$$

Suppose L_n has the following eigenvalue decomposition.

$$L_n = \sum_{i=1}^{n} \lambda_i u_i u_i^T \tag{9.12}$$

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n, \tag{9.13}$$

where $u_1, u_2, \ldots u_k$ are the eigenvectors of L_n corresponding to the k smallest eigenvalues. Similarly, let

$$\bar{L}_n = \sum_{i=1}^{n} \bar{\lambda}_i \bar{u}_i \bar{u}_i^T \tag{9.14}$$

$$0 = \bar{\lambda}_1 \leq \bar{\lambda}_2 \leq \cdots \leq \bar{\lambda}_n, \tag{9.15}$$

where $\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_k$ are eigenvectors of $\bar{L}_n$ corresponding to 0 eigenvalue. We can apply Davius-Kahan Theorem to measure the distance between the eigen-spaces U and $\bar{U}$, where

$$U_{n\times k} = [u_1 \ u_2 \ \ldots \ u_k] \tag{9.16}$$

$$\bar{U}_{n\times k} = [\bar{u_1} \ \bar{u_2} \ \ldots \ \bar{u_k}]. \tag{9.17}$$

$$\tag{9.18}$$

In particular,

$$d_s(U, \bar{U}) \triangleq \min_{Q\in O(k,k)} \|U - \bar{U}Q\|_2 \leq \frac{\sqrt{2}\|\Delta\|_2}{\delta} \tag{9.19}$$

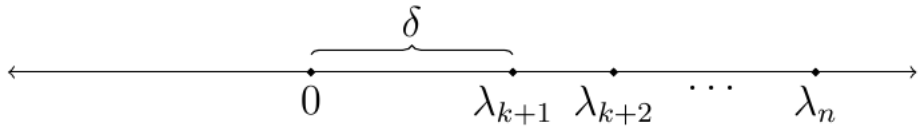
$$\tag{9.20}$$

where $\delta = \lambda_{k+1}(L_n) - \lambda_k(\bar{L}_n)$ is called *spectral gap*. Since $\bar{G}$ has k connected components,

$$\bar{\lambda}_1 = \bar{\lambda}_2 = \cdots = \bar{\lambda}_k = 0 \tag{9.21}$$

and thus $\delta = \lambda_{k+1}(L_n)$.

Pictorially,



From Weyl's theorem,

$$\delta = \lambda_{k+1}(L_n) \geq \lambda_{k+1}(\bar{L}_n) - \|\Delta\|_2 \tag{9.22}$$

$$\Rightarrow \min_{Q\in O(k,k)} \|U - \bar{U}Q\|_2 \leq \frac{\sqrt{2}\|\Delta\|_2}{\lambda_{k+1}(\bar{L}_n) - \|\Delta\|_2} \tag{9.23}$$

Note: It is evident that for U and $\bar{U}Q$ to be close, we need $\lambda_{k+1}(\bar{L}_n)$ to be much larger than $\|\Delta\|_2$. In HW3, we have shown that λ_{k+1} is related to the edge expansion of graph $\bar{G}$ through *Cheeger's inequality*. A larger $\bar{\lambda}_{k+1}$ means the connected components are more well connected. Also, $\|\Delta\|_2$ characterizes the perturbation. If graph G is close to $\bar{G}$, then $\|\Delta\|_2$ tends to be small.
Note: We can further derive a bound on the difference between U and $\bar{U}Q$ in terms of Frobenius norm. In particular, note that U and $\bar{U}Q$ both are at most rank k matrices and hence $U - \bar{U}Q$ is at most rank-$2k$ matrix. This means that,

$$\|U - \bar{U}Q\|_F \leq \sqrt{2k}\|U - \bar{U}Q\|_2, \tag{9.24}$$

as now we can have at most $2k$ non-zero singular values and each is bounded above by the largest singular value. Therefore,

$$\min_{Q\in O(k,k)} \|U - \bar{U}Q\|_F \leq \frac{\sqrt{2k}\sqrt{2}\|\Delta\|_2}{\lambda_{k+1}(\bar{L}_n) - \|\Delta\|_2} \tag{9.25}$$

$$\Rightarrow \min_{Q\in O(k,k)} \frac{1}{n}\|U - \bar{U}Q\|_F^2 \leq \frac{2k\|\Delta\|_2^2}{n(\lambda_{k+1}(\bar{L}_n) - \|\Delta\|_2)^2} \tag{9.26}$$

$$\Rightarrow \min_{Q\in O(k,k)} \frac{1}{n}\sum_{i=1}^{n} \|U_i - (\bar{U}Q)_i\|_2^2 \leq \frac{2k\|\Delta\|_2^2}{n(\lambda_{k+1}(\bar{L}_n) - \|\Delta\|_2)^2} \tag{9.27}$$

Note that $\|U_i - (\bar{U}Q)_i\|_2$ measures the deviation of U_i from the cluster center $(\bar{U}Q)_i$. A small $\frac{1}{n}\sum_{i=1}^{n}\|U_i - (\bar{U}Q)_i\|_2^2$ means that the average deviation of U_i's from the k cluster centers is small, as shown in Fig. 9.3 for $k = 2$.

We are left to quantify $\lambda_{k+1}(\bar{L}_n)$ and $\|\Delta\|_2$. In the next section, we will introduce probabilistic models for graph G, so that $\lambda_{k+1}(\bar{L}_n)$ and $\|\Delta\|_2$ can be characterized using concentration inequalities.

9.5 Random graph models for graph clustering

Analogues to Gaussian mixture model for clustering, we will introduce some probabilistic model to generate graph G. In particular, we will introduce *Stochastic Block model*, also known as *Planted Partition model*. Before that, we first briefly mention the motivations for graph clustering.

9.5.1 Motivations for graph clustering

There are at least two motivations to study graph clustering. One is that as we just discussed, the problem of data clustering can be turned into a problem of clustering (weighted) graph by constructing similarity matrix between data points. The other is network analysis. In practice, we often observe various forms of networks, such as social networks and biological networks. In many cases, these networks have community structure, where nodes form clusters, and there are more edges inside the clusters than across clusters. For example, in social networks, the communities may correspond to group of people who share similar interests, and people with similar interests are more likely to be friends. In protein-protein interaction network, the communities may correspond to functional groups of proteins, and proteins with similar biological functions are more likely to interact with each other. It is thus of great interest to discover the hidden community structure (clusters) from observation of graph.

9.5.2 Inhomogeneous Random Graph

Recall that we have already introduced *Erdős-Rényi random Graphs* (ER Random Graphs) which are named after Paul Erdős and Alfred Rényi. This model is also denoted by $\mathcal{G}(n,p)$, where there are n vertices in total and every pair of two vertices is connected independently at random with probability p. Observe that in $\mathcal{G}(n,p)$, edge probabilities are homogeneous and it does not capture the community structure.

We next consider an inhomogeneous model denoted by $\mathcal{G}(n, P_{n\times n})$, where there are n vertices and every pair of two different vertices i and j is connected by an edge independently at random with probability P_{ij}. Note that P is a symmetric matrix. In the special case where $P_{ij} = p$ for $i \neq j$, it reduces to $\mathcal{G}(n,p)$. The inhomogeneous model $\mathcal{G}(n,P)$ is quite general, but it still does not capture the community structure in networks.
To deal with this issue, we may assume P has the following property:

$$P_{ij} = \begin{cases} p, & \text{if } i,j \text{ from the same cluster} \\ q, & \text{if } i,j \text{ from two different cluster} \\ 0, & \text{if } i = j \end{cases}$$

where $0 \leq q \leq p \leq 1$. Thus the edge probabilities within clusters are p and edge probabilities across clusters are q. Since $p \geq q$, it captures the fact that nodes from the same cluster are more likely to be connected than nodes from different clusters. As a result, the graph G will have clusters that are densely connected inside and loosely connected across. Some special cases for matrix P are of interest:

1. $p = 1$ and $q = 0$: The graph G are formed by disjoint cliques.

2. $p = q$: It reduces to the E-R random graph model

Edge probabilities within clusters may not all be the same; similarly, edge probabilities across clusters may not all be the same as well. Thus we have the following definition of SBM in the more general case.

Definition 9.1 (Stochastic block model)**.** Let $C_1^*, C_2^*, \dots C_k^*$ denote a k-partition of $[n]$. Let $U \in [0,1]^{n\times k}$ denote the cluster membership matrix, where $U_{is} = 1$ if and only if $i \in C_s^*$. Let $B \in [0,1]^{k\times k}$ denote the symmetric, connectivity matrix between clusters, where B_{st} is the edge probability between nodes in C_s^* and C_t^*. Define $P = UBU^\top$, i.e., $P_{ij} = B_{s,t}$ if $i \in C_s^*$ and $j \in C_t^*$. Conditional on $C_1^*, C_2^*, \dots C_k^*$, we generate a random graph G with adjacency matrix A, where $A_{ii} = 0$ for all i, and $A_{ij} = A_{ji} \overset{\text{i.i.d.}}{\sim} \text{Bern}(P_{ij})$ for $i < j$.

Note that if $B_{st} = p$ if $s = t$ and $B_{st} = q$ for $s \neq t$, then it reduces to the previous simple (p,q) model.

We have not specified how to generate cluster partition $C_1^*, \dots, C_k^*$ yet. Let us first focus on the special case where $k = 2$. Denote $x_i \in \{\pm 1\}$ as the cluster label of node i, where $x_i = 1$ means node i belongs to C_1^*, and $x_i = -1$ means node i belongs to C_2^*. We can consider the following two possible distributions for x.

1. Binomial-sized partition model: $x_i \overset{\text{i.i.d.}}{\sim} \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$, where δ_x denote the delta measure at point x. In other words, x_i is independently and uniformly drawn from $\{\pm\}$.

2. Equal-sized partition model: x is uniformly drawn from $\{y \in \{\pm 1\}^n : \sum_{i=1}^n y_i = 0\}$.

Note: In the first case, x_i's are i.i.d., while in the second distribution, x_i's are not independent due to the equal-size constraint. Moreover, in the first case, the size of any one of the cluster is distributed as $\text{Binom}(n, 1/2)$. That is why we call the first case Binomial-sized partition model. It follows from concentration inequality for binomial distribution that the cluster sizes are bounded between $n/2 - \sqrt{n\rho_n}$ and $n/2 + \sqrt{n\rho_n}$ with high probability, where $\rho_n \to \infty$ arbitrarily slowly.

We can generalize the cluster label to $k \geq 2$ communities as follows. Denote $x_i \in [k]$ as the cluster label of node i, where $x_i = s$ means node i belongs to C_s^* for $1 \leq s \leq k$. We can consider the following two possible distributions for x:

1. Binomial-sized partition model: $x_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\{1, 2, \ldots, r\}$.
2. Equal-sized partition model: x is uniformly drawn from $\{y \in [k]^n : |\{i : y_i = s\}| = n/k, \forall 1 \leq s \leq k\}$.

Remark 9.1 (Conditional independent but marginally dependent)**.** Notice that $\{A_{ij}\}$ are independently drawn conditional on x, however, $\{A_{ij}\}$ are not independent marginally. In particular, one can check that if $p > q$, then

$$P(A_{jk} = 1 | A_{ij} = 1, A_{ik} = 1) > P(A_{jk} = 1),$$

which shows $\{A_{ij}\}$ are not independent.

9.5.3 Spectral clustering for binary symmetric SBM

In this section, we introduce the spectral clustering fro binary symmetric SBM. For ease of exposition, we focus on the binary symmetric case where $k = 2$, and $B = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$. We are interested in recovering the underlying clusters based on observation of G and knowledge of k, p, q.

We assume x is generated according to the equal-sized partition model, i.e., $|C_1^*| = |C_2^*| = \frac{n}{2}$. Then

$$P_{n\times n} = E[A] = \left(\begin{array}{c|c} p\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} & q\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} \\ \hline q\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} & p\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} \end{array}\right) - p\mathbf{I}_{n\times n},$$

where $\mathbf{J}$ denotes the all-one matrix, and $\mathbf{I}$ denotes the identity matrix. If we know $E[A]$, then the problem is trivial. However, the point is we do not know $E[A]$, but only observe A.
The key observation is that spectrum of $E[A]$ contains cluster information. More specifically,

$$\begin{aligned} E[A] - \frac{p+q}{2}\mathbf{J} &= \left(\begin{array}{c|c} \frac{p-q}{2}\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} & \frac{q-p}{2}\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} \\ \hline \frac{q-p}{2}\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} & \frac{p-q}{2}\mathbf{J}_{\frac{n}{2}\times\frac{n}{2}} \end{array}\right) - p\mathbf{I} \\ &= \frac{p-q}{2}Y^* - p\mathbf{I}, \end{aligned}$$

where $Y^* = xx^\top$ is the *partition matrix* and x is the underlying true label vector. In particular, $Y_{ij}^* = 1$ if i and j are from the same cluster and $Y_{ij}^* = -1$ if i and j are from two different clusters.

Note that Y^* has rank 1 with the leading eigenvalue given by n and the leading eigenvector given by $x/\sqrt{n}$. Since A is a noisy version of $\mathbb{E}[A]$, we hope that the leading eigenvector of $\mathbb{E}[A] - \frac{p+q}{2}\mathbf{J}$ is close to $x/\sqrt{n}$. This leads to the following simple spectral algorithm.

Algorithm 3 Spectral algorithm for binary symmetric SBM

1: Compute the leading eigenvector of $A - \frac{p+q}{2}\mathbf{J}$ denoted by u.
2: Take $\operatorname{sign}(u)$ as the cluster label.

Remark 9.2. Instead of applying spectral algorithm on adjacency matrix A, we could also apply spectral clustering on Laplacian matrix. The two versions all belong to the general class of *spectral method*. In the relatively dense graph case, the two versions are almost equivalent. To see this, assume $n(p+q) \gg \log n$, then we have shown that node degrees d_i are concentrated around the average degree $n\frac{p+q}{2}$ with high probability, and thus D is close to $n\frac{p+q}{2}\mathbf{I}$. Thus the eigenvectors of Laplacian matrices are almost the same as eigenvectors of A.

Chapter 10

Concentration of random graphs

Outline

- Review of spectral clustering under binary symmetric SBM
- Analysis of spectral clustering
- Concentration of random graph

10.1 Review of Spectral clustering under binary symmetric stochastic block model (SBM)

First we talk about the setup of binary symmetric SBM. We have two communities denoted by C_1^* and C_2^*, where C_1^* represents the set of nodes in the first community and C_2^* is the set of nodes in the second community. Let's assume that these two communities have equal size $|C_1^*| = |C_2^*| = n/2$, where n is the total number of nodes in the graph. We also assume that (C_1^*, C_2^*) is uniformly chosen over all possible equal-sized partitions at random. Conditional on the underlying true communities C_1^* and C_2^*, the edge connection probability is defined as

$$P(A_{ij} = 1) = \begin{cases} p & \text{if } i \text{ and } j \text{ are from the same community} \\ q & \text{o.w.} \end{cases}$$

Where A_{ij} is the adjacency matrix and by convention $A_{ii} = 0$. The goal is to infer the true underlying partition C_1^* and C_2^*, given (A, p, q).

Last time we have introduced the spectral clustering. Recall that

$$E[A] - \frac{p+q}{2}\mathbf{J} + p\mathbf{I} = \frac{p-q}{2}Y^*,$$

where $\mathbf{J}$ is the all 1 matrix and Y^* is the partition matrix such that

$$Y^* = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \mathbf{J}_{n/2\times n/2} = \begin{bmatrix} \mathbf{J}_{n/2\times n/2} & -\mathbf{J}_{n/2\times n/2} \\ -\mathbf{J}_{n/2\times n/2} & \mathbf{J}_{n/2\times n/2,} \end{bmatrix} = xx^T$$

where $\otimes$ denotes the Kronecker product and x is the underlying true cluster label. It follows that the leading eigenvector of Y^* is given by $\bar{u} \triangleq x/\|x\|_2$. Hence, $\mathbb{E}[A]$ contains the underlying cluster information. However, we do not observe $E[A]$; instead we observe A, which motivates us to use

Algorithm 4 Spectral algorithm for binary symmetric SBM

1: Compute the leading eigenvector of $A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}$, denoted by u.
2: Output $\widehat{x} \triangleq \text{sign}(u)$ as the estimated cluster label vector.

spectral clustering.

In this lecture, we will introduce

1. Analysis of spectral clustering using Davis-Kahan theorem.
2. Concentration of random graph (through the adjacency matrix).

10.2 Analysis of spectral clustering using D-K theorem

We will follow exactly the same road map that we used to analyze the spectral clustering for Gaussian mixture model. We expect that u is close to the leading eigenvector of Y^* given that $\|A - \mathbb{E}[A]\|_2$ is small. Recall that $\bar{u} = \frac{x}{\sqrt{n}}$ is the leading eigenvector of Y^*.

Theorem 10.1 (Upper bound on mislcassifcation rate of spectral clustering)**.** *The fraction of nodes misclassified by the spectral clustering satisfies*

$$\frac{1}{n} d_H(\widehat{x}, x) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\widehat{x}_i \neq x_i\}} \leq \frac{2\|A - \mathbb{E}[A]\|_2^2}{(n(p-q)/2 - \|A - \mathbb{E}[A]\|_2)^2}.$$

In particular, if $\frac{(p-q)n}{\|A-\mathbb{E}[A]\|_2} \to \infty$*, then* $\frac{1}{n} d_H(\widehat{x}, x) \to \infty$*, i.e., the fraction of nodes misclassified by* $\widehat{x}$ *vanishes.*

Proof. Recall that in lecture 6, we have shown that

$$\frac{1}{n} d_H(\widehat{x}, x) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\widehat{x}_i \neq x_i\}} \leq d_s^2(u, \bar{u}).$$

Hence, to prove the theorem, it reduces to upper bound $d_s^2(u, \bar{u})$. By Davis-Kahan sin-θ theorem,

$$d_s(u, \bar{u}) \triangleq \min\{\| u - \bar{u} \|_2, \| u + \bar{u} \|_2\} \leq \frac{\sqrt{2} \| \Delta \|_2}{\delta},$$

where

$$\begin{aligned}
\Delta &= A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I} - \frac{p-q}{2}Y^* \\
&= A - E[A] + E[A] - \frac{p+q}{2}\mathbf{J} + p\mathbf{I} - \frac{p-q}{2}Y^* = A - E[A]
\end{aligned}$$

and

$$\begin{aligned}
\delta &= \lambda_1\left(\frac{p-q}{2}Y^*\right) - \lambda_2\left(A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}\right) \\
&= \frac{p-q}{2}n - \lambda_2\left(\Delta + \frac{p-q}{2}Y^*\right)
\end{aligned}$$

with λ_1 denotes the largest eigenvalue and λ_2 denotes the second largest eigenvalue. By Weyl's theorem,

$$\begin{aligned}\lambda_2\left(\Delta+\frac{p-q}{2}Y^*\right) &\leq \lambda_2\left(\frac{p-q}{2}Y^*\right)+\parallel\Delta\parallel_2\\ &=\parallel\Delta\parallel_2\end{aligned}$$

It follows that

$$\delta \geq \frac{p-q}{2}n- \parallel\Delta\parallel_2$$

Hence, we get that

$$d_s(u,\bar{u}) \leq \frac{\sqrt{2}\parallel\Delta\parallel_2}{\frac{p-q}{2}n- \parallel\Delta\parallel_2},$$

which completes the proof. □

10.3 Concentration of random graphs

As we shown in the last section, the fraction of nodes misclassified by the spectral clustering crucially depends on $\|A-\mathbb{E}[A]\|_2$. In this section, we study the concentration of A. Recall that $\{A_{ij}\}$ are independent conditional on Y^*, and

$$A_{ij}\sim\begin{cases}Ber(p) & \text{if } Y^*_{ij}=1, \text{ i.e., } i \text{ and } j \text{ are from the same cluster}\\ Ber(q) & \text{if } Y^*_{ij}=-1, \text{ i.e., } i \text{ and } j \text{ are from two different clusters}\end{cases}$$

It follows that $\mathbb{E}[A]$ is (To be precise, it should be $\mathbb{E}[A|Y^*]$; for ease of exposition, we omit the conditioning.)

$$\mathbb{E}[A]=\begin{bmatrix}p & q\\ q & p\end{bmatrix}\otimes\mathbf{J}_{n/2\times n/2}-p\mathbf{I}=\begin{bmatrix}p\mathbf{J}_{n/2\times n/2} & q\mathbf{J}_{n/2\times n/2}\\ q\mathbf{J}_{n/2\times n/2} & p\mathbf{J}_{n/2\times n/2}\end{bmatrix}-p\mathbf{I}.$$

Since $A_{ij}-\mathbb{E}[A_{ij}]$ is bounded by 1, and has zero mean, it is sub-Gaussian ($\mu=0,\sigma=2$). One can show that with high probability, $\|A-\mathbb{E}[A]\|_2\lesssim\sqrt{n}$. However, this bound is not tight when p is small. We will derive a tighter bound when $p,q\to 0$ based on *Matrix Bernstein's inequality.*

Theorem 10.2 (Matrix Bernstein's Inequality)**.** *Let $\{X_i\}_{i=1}^N$ denote $n\times n$ symmetric, independent random matrices with mean* $\mathbf{0}$ *and $\parallel X_i\parallel_2\leq L$. Let $S=\sum_{i=1}^N X_i$. Then*

$$\mathbb{P}[\parallel S\parallel_2\geq t]\leq 2n\exp\left\{-\frac{t^2}{2\sigma^2+\frac{2Lt}{3}}\right\},$$

where $\sigma^2\triangleq\parallel\sum_{i=1}^N E[X_i^2]\parallel_2$. It futher follows that

$$\mathbb{P}\left[\|S\|_2\geq\sqrt{2\sigma^2u}+\frac{2}{3}Lu\right]\leq 2ne^{-u}.$$

Remark 10.1. Matrix Bernstein's inequality is formalized very recently, and pioneered by [Ahlswede-Winter],..., and [J. Tropp]. If $N=1$, it reduces to the Bernstein's inequality in the scalar case.

The following lemma upper bounds $\parallel A-\mathbb{E}[A]\parallel_2$ via Matrix Bernstein Inequality.

Lemma 10.1 (Concentration inequality for A)**.** *Suppose that A is a symmetric matrix such that $A_{ii} = 0$, and A_{ij} are independently distributed over $[0, 1]$ for $i < j$. Let $\sigma^2 = \max_i \sum_{j \neq i} var(A_{ij})$. Then*

$$\mathbb{P}\left[\|A - \mathbb{E}[A]\|_2 \geq \sqrt{2\sigma^2 u} + \frac{2}{3}u\right] \leq 2ne^{-u}.$$

Proof. First, write $A - \mathbb{E}[A]$ as a summation of independent, symmetric random matrices:

$$A - \mathbb{E}[A] = \sum_{i<j} (A_{ij} - \bar{A}_{ij}) \left(e_i e_j^\top + e_j e_i^\top\right).$$

Let $\bar{A} \triangleq \mathbb{E}[A]$. Notice that $A_{ij} - \bar{A}_{ij}$'s are independent random variables for $i < j$. Let $X_{ij} \triangleq (A_{ij} - \bar{A}_{ij})(e_i e_j^\top + e_j e_i^\top)$. Then

$$\begin{aligned} \| X_{ij} \|_2 =& \| (A_{ij} - \bar{A}_{ij})(e_i e_j^T + e_j e_i^T) \|_2 \\ \leq& \| e_i e_j^T + e_j e_i^T \|_2 \\ \leq& 1 \end{aligned}$$

Furthermore,

$$\begin{aligned} E[X_{ij}^2] &= E[(A_{ij} - \bar{A}_{ij})^2 (e_i e_j^T + e_j e_i^T)^2] \\ &= Var(A_{ij})(e_i e_j^T + e_j e_i^T)^2 \\ &= Var(A_{ij})(e_i e_i^T + e_j e_j^T) \end{aligned}$$

It follows that

$$\begin{aligned} \sum_{i<j} E[X_{ij}^2] &= \sum_{i<j} Var(A_{ij})(e_i e_i^T + e_j e_j^T) \\ &= \frac{1}{2} \sum_{i \neq j} Var(A_{ij})(e_i e_i^T + e_j e_j^T) \\ &= \sum_{i \neq j} Var(A_{ij})(e_i e_i^T) \end{aligned}$$

Hence

$$\sigma^2 = \left\| \begin{pmatrix} \sum_{j \neq 1} var(A_{1j}) & & & \\ & \sum_{j \neq 2} var(A_{2j}) & & \\ & & \ddots & \\ & & & \sum_{j \neq n} var(A_{nj}) \end{pmatrix} \right\|_2 = \max_i \sum_{j \neq i} var(A_{ij}).$$

The lemma follows by invoking matrix Bernstein's inequality. □

Corollary 10.1. *Suppose that A is a symmetric matrix such that $A_{ii} = 0$, and A_{ij} are independently distributed as Bern(P_{ij}) for $i < j$. Let $\sigma^2 = \max_i \sum_{j \neq i} P_{ij}(1 - P_{ij}) \leq \max_i \sum j \neq i P_{ij} \triangleq \max_i \mathbb{E}[d_i]$. Then*

$$\mathbb{P}\left[\|A - \mathbb{E}[A]\|_2 \geq \sqrt{2\sigma^2 u} + \frac{2}{3}u\right] \leq 2ne^{-u}.$$

In particualr, it follows that

$$\mathbb{P}\left[\|A - \mathbb{E}[A]\|_2 \leq \sqrt{2\sigma^2(1+\epsilon)\log(n)} + \frac{2}{3}(1+\epsilon)\log(n)\right] \geq 1 - 2n^{-\epsilon}.$$

Example 10.1. Let's take a look at some special cases of P_{ij}.

Case 1: $(P_{ij} = p,\ \forall\ i \neq j)$. In this case, $A \sim \mathcal{G}(n,p)$. It follows that

$$\begin{aligned}\|A - E[A]\|_2 &\leq \sqrt{2\ np(1-p)(1+\epsilon)\log(n)} + \frac{2}{3}(1+\epsilon)\log(n) \\ &\lesssim \sqrt{np\log(n)} + \log(n) \qquad \text{(ignoring the constants)}\end{aligned}$$

Hence,

$$\|A - E[A]\|_2 \lesssim \begin{cases}\sqrt{np\log(n)} & np = \omega(\log(n)) \\ \log(n) & np = \mathcal{O}(\log(n)).\end{cases}$$

Case 2: $P = \begin{bmatrix} p & q \\ q & p \end{bmatrix} \otimes J_{n/2\times n/2}$ with $p \geq q$. In this case, A is the adjacency matrix of the binary symmetric SBM. We have that

$$\begin{aligned}\|A - E[A]\|_2 &\leq \sqrt{n(p(1-p)+q(1-q))(1+\epsilon)\log(n)}\ + \frac{2}{3}(1+\epsilon)\log(n) \\ &\lesssim \sqrt{np\log(n)} + \log(n)\end{aligned}$$

Question: Is the bound $\|A - \mathbb{E}\,[A]\,\|_2$ derived from matrix Bertein's inequality tight or not ?
Answer: Here we give an intuitive argument by considering $A \sim \mathcal{G}(n,p)$. Recall that if $B \in \mathcal{R}^{n\times d}$ is a Gaussian random matrix such that $B_{ij} = B_{ji} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$ for $i < j$ and $B_{ii} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,2\sigma^2)$. Then $\|B\|_2 \leq \sigma(\sqrt{n} + \sqrt{d} + \rho_n)$ with probability converging to 1, where $\rho_n \to \infty$ as $n \to \infty$ arbitrarily slowly. In the Bernoulli case, $\mathsf{var}(A_{ij}) \leq p$. If we expect $(A - E[A])$ to behave like Gaussian random matrix B, then we would have $\|A - E[A]\|_2 \lesssim \sqrt{np}$ with high probability. If this is true, this means that we could possibly get rid of the extra $\sqrt{\log n}$ factor in the upper bound derived from matrix Bertein's inequality when $np = \Omega(\log n)$. We will show later that this is indeed the case if $np = \omega(\log n)$. On the other hand, it turns out that we cannot significantly improve the bound derived from matrix Bernstein's inequality. To see this, observe that

$$\|A - E[A]\|_2 \geq \max_i \|A_{i\cdot} - E[A_{i\cdot}].\|_2 = \max_i \sqrt{\sum_{j\neq i}(A_{ij} - p)^2} = \sqrt{(1-2p)d_{\max}},$$

where $A_{i,\cdot}$ denotes the i-th row of A and $d_{\max}$ is the maximum node degree.

- In the regime $np = \Omega(\log n)$, we have shown that $\max_i d_i$ concentrates around np. It follows that with high probability, $\|A - E[A]\|_2 \geq \sqrt{np}$, which shows that the bound derived from matrix Bernstein's inequality: $\|A - E[A]\|_2 \lesssim \sqrt{np\log(n)}$ is suboptimal by at most an $\sqrt{\log n}$ factor.
- In the regime $np = o(\log n)$, we have shown that $\max_i d_i$ is at least on the order of $\frac{\log n}{\log(\log n/(np))}$, which is $\omega(np)$. It follows that $\|A - E[A]\|_2 \geq \sqrt{\frac{\log n}{\log(\log n/(np))}} = \omega(\sqrt{np})$. In this case, we see that $\|A - E[A]\|_2$ behaves significantly differently from Gaussian random matrix B. As we will show later, this leads to the failure of spectral clustering in the relatively sparse regime $np = o(\log n)$.

The following theorem gives the desired concentration inequality for Bernoulli random matrix A.

Theorem 10.3 (Concentration of Bernoulli random matrix A in the dense regime)**.** *Suppose A is a symmetirc matrix such that $A_{ii} = 0$ and $A_{ij} \overset{i.i.d.}{\sim} Bern(P_{ij})$ for $i < j$. Suppose $P_{ij}(1 - P_{ij}) \le r$ and $nr = \Omega(\log(n))$. Then for any $c > 0$, there exists an $c' > 0$ such that*

$$\|A - E[A]\|_2 \le c'\sqrt{nr},$$

with probability $1 - n^{-c}$, where c, c' are absolute constants.

Remark 10.2. The above bound improves the bound derived from the matrix Bertein's inequality by getting rid of the extra $\sqrt{\log(n)}$ factor.

The condition $nr = \Omega(\log(n))$ is crucial. As we argued for $A \sim \mathcal{G}(n, p)$, $\|A - \mathbb{E}[A]\|_2 = \omega(\sqrt{np})$ if $np = o(\log n)$.

Proof. The proof is left as a homework problem. □

Theorem 10.4 (Concentration of Bernoulli random matrix A in the sparse regime)**.** *Suppose $A \sim \mathcal{G}(n, p)$ and $np = o(\log n)$ and $p = n^{-1+o(1)}$. With probability converging to 1,*

$$\|A - E[A]\|_2 \gtrsim \sqrt{\frac{\log(n)}{\log(\frac{log(n)}{np})}}.$$

Remark 10.3. Notice that $\frac{\log(n)}{\log(\frac{log(n)}{np})} = \omega(\sqrt{np})$ and hence $\|A - E[A]\|_2 = \omega(\sqrt{np})$. This is due to the fact that $\|A - E[A]\|_2 \gtrsim \sqrt{d_{\max}}$ and in the sparse regime, $d_{\max}$ does not concentrate around $\mathbb{E}[d_i] \ \ np$; instead it is at least on the order of $\frac{\log(n)}{\log(\frac{log(n)}{np})}$ with high probability. The proof of the theorem is left as homework.

Here, we give a short summary on the concentration results of Bernoulli random matrix in the special case $A \sim \mathcal{G}(n, p)$:

$$\|A - E[A]\|_2 = \begin{cases} \mathcal{O}(\sqrt{np}) & \text{if } np = \Omega(\log(n)) \text{ (dense graph regime)} \\ \omega(\sqrt{np}) & \text{if } np = o(\log(n)) \text{ (sparse graph regime)} \end{cases}$$

Armed with the concentration results of Bernoulli random matrix, we are ready to state the sufficienit conditions for spectral clustering in misclassifying a vanishing fraction of nodes.

Theorem 10.5 (Sufficient condition for spectral clustering in the dense regime)**.** *Suppose $np = \Omega(\log(n))$. If $n(p - q)/\sqrt{np} \to \infty$, then $\frac{1}{n}d_H(x, \widehat{x}) \to 0$ as $n \to \infty$ with high probability.*

Proof. Recall that

$$\frac{1}{n}d_H(\widehat{X}, X) \le \frac{2\|\Delta\|_2}{\left(\frac{p-q}{2}n - \|\Delta\|_2\right)}$$

and in the dense regime $np = \Omega(\log n)$, $\|A - E[A]\|_2 \lesssim \sqrt{np}$ with high probability. Therefore, $\frac{1}{n}d_H(\widehat{X}, X) \to 0$ under the condition that $\frac{(p-q)n}{\sqrt{np}} \to \infty$. □

Remark 10.4. The sufficiet condition $n(p - q)/\sqrt{np} \to \infty$ is known as *spectral condition*. It is crucial to assume $np = \Omega(\log n)$ here. If instead $np = o(\log n)$, since $\|A - \mathbb{E}[A]\|_2 = \omega(\log n)$, the spectral condition is $n(p - q)/\sqrt{np} \to \infty$ not sufficient for the success of the spectral clustering. We could use the matrix Bernstein's inequality to show that $\|A - \mathbb{E}[A]\|_2 \lesssim \log n$ in the sparse regime $np = o(\log n)$ and then $\frac{1}{n}d_H(\widehat{x}, x) \to 0$, provided that $\frac{(p-q)n}{\log(n)} \to \infty$. However, since $np = o(\log n)$, such condition is never satisifed.

The above discussion immediately leads to the following two fundamental questions, which will be addressed in the next lecture:

1. Relatively dense regimes ($np = \Omega(\log(n))$: Can we do better than spectral clustering? Is it possible to succeed even when the spectral condition is unsatisifed?

2. In the relatively parse case ($np = o(\log(n))$): Can we hope to recover the communities when the spectral condition $\frac{(p-q)n}{\sqrt{np}} \to \infty$ is satisifed?

Chapter 11

SDP clustering with stochastic block models

11.1 Brief recap of concentration of random graphs and spectral clustering

In the last lecture, we have discussed the concentration of random graphs. We observe two different behavior depending on whether the graph is dense or sparse. Specifically, let $A \sim \mathcal{G}(n,p)$. Then

$$\|A - \mathbb{E}[A]\|_2 = \begin{cases} O(\sqrt{np}), & \text{if } \ np = \Omega(\log n) \\ w(\sqrt{np}), & \text{if } \ np = o(\log n). \end{cases}$$

It readily implies that for the binary symmetric stochastic block model, in the dense regime with $np = \Omega(\log n)$, the fraction of nodes misclassified by spectral clustering converges to 0, provided that the following *spectral condition* is satisfied as $n \to \infty$:

$$\frac{n(p-q)}{\sqrt{np}} \to \infty.$$

The above spectral condition can be intuitively understood as a measure of signal-to-noise ratio: the numerator $n(p-q)$ is roughly the second largest eigenvalue of $\mathbb{E}[A]$, or the largest eigenvalue of $\mathbb{E}[A] - (p+q)/2\mathbf{J} + p\mathbf{I}$, which characterizes the signal strength, because the leading eigenvector of $\mathbb{E}[A] - (p+q)/2\mathbf{J} + p\mathbf{I}$ is parallel to the true cluster label vector; the denominator is roughly $\|A - \mathbb{E}[A]\|_2$, which captures the noise magnitude. When the spectral condition is satisfied, the signal strength is much larger than the noise magnitude, and thus the leading eigenvector of centered adjacency matrix $A - (p+q)/2\mathbf{J} + p\mathbf{I}$ is approximately parallel to the true cluster label vector.

In contrast, in the sparse regime with $np = o(\log n)$, because $\|A - \mathbb{E}[A]\|_2 = \omega(\sqrt{np})$, the spectral condition is no longer sufficient for the spectral method to achieve a vanishing fraction of misclassified nodes. In this lecture, we will introduce the semidefinite programming relaxation of MLE, which is able to achieve a vanishing fraction of misclassified nodes when the spectral condition is satisfied even in the sparse regime.

Example 11.1. In this example, we generate a binary SBM with 1000 nodes partitioned into two equal-sized clusters uniformly at random. We first focus on the dense regime $p = \frac{\log n}{n}$ and $q = p/8$. Fig. 11.1 shows the adjacency matrix A. Fig 11.2 shows the leading eigenvector u of the centered adjacency matrix $A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}$. We can see that the sign vector of u is strongly correlated with

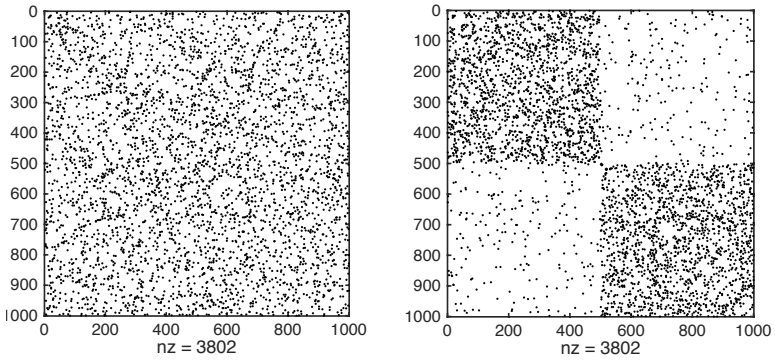


Figure 11.1: $n = 1000$, $p = \frac{\log n}{n}$, $q = \frac{p}{8}$. Left: adjacency matrix A; Right: A with rows and columns sorted according to the true cluster label.

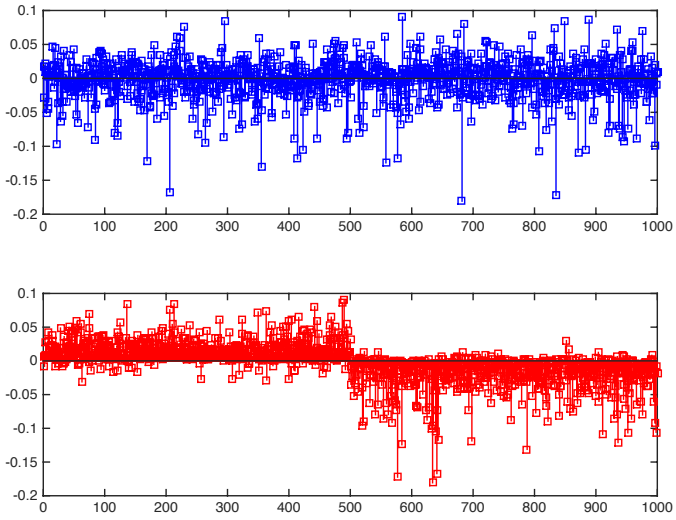


Figure 11.2: Top: the eigenvector u corresponding to the largest eigenvalue of $A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}$. Bottom: u sorted according to the true cluster label.

the true cluster label vector. In particular, if we sort u according to the true cluster label with the first 500 coordinates corresponding to one cluster, and the next 500 coordinates corresponding to the other cluster, then most of the first 500 coordinates of u are positive, while most of the next 500 coordinates of u are negative. Hence, the estimated cluster label vector $\widehat{x} = \text{sign}(u)$ is close to the true cluster label vector.

Next, we shift our focus to the sparse regime where $p = \frac{\sqrt{\log n}}{n}$ and $q = \frac{p}{8}$. Fig. 11.3 shows the adjacency matrix A. Fig. 11.4 shows the leading eigenvector u of the centered adjacency matrix $A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}$. We can see that the sign vector of u is no longer correlated with the true cluster label vector. Moreover, u turns out to be very spiky with few entries of large magnitudes. Fig. 11.5 shows the clustering result if we insist on using $\widehat{x} = \text{sign}(u)$ to estimate the clusters. The fraction of misclassified nodes is 45.30%, which is no much better than randomly guessing. Moreover, we see that one cluster estimated by $\widehat{x}$ is in fact a small dense subgraph. It is this small dense subgraph that induces the spiky eigenvector u.

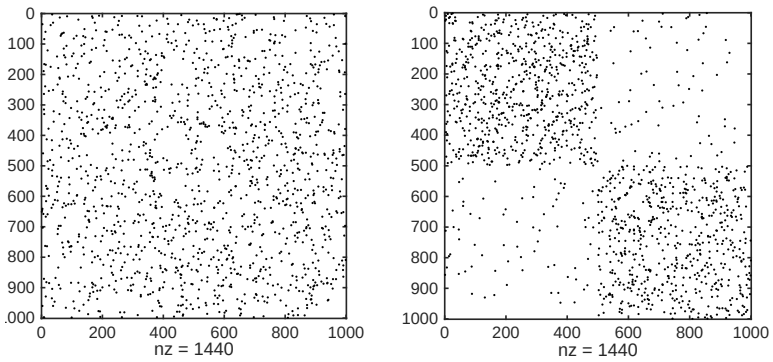


Figure 11.3: $n = 1000$, $p = \frac{\sqrt{\log n}}{n}$, $q = \frac{p}{8}$. Left: adjacency matrix A; Right: A with rows and columns sorted according to the true cluster label

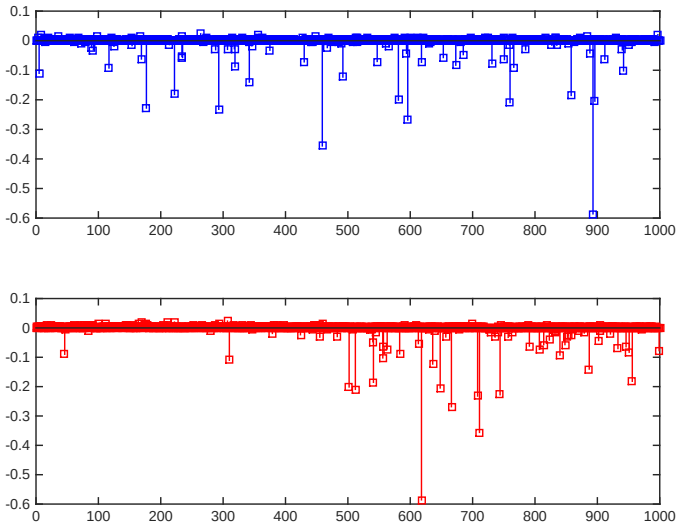


Figure 11.4: Top: the eigenvector u corresponding to the largest eigenvalue of $A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}$. Bottom: u sorted according to the true cluster label.

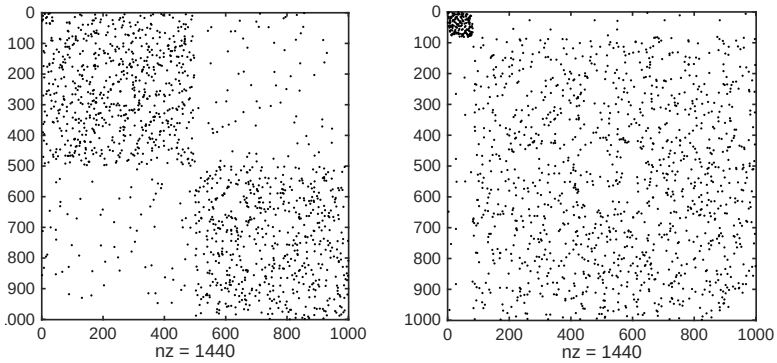


Figure 11.5: Left: A with rows and columns sorted according to the true cluster label. Right: A with rows and columns sorted according to the estimated cluster label vector.

11.2 Semi-definite relaxation of MLE

For ease of expose, we will focus on binary symmetric SBM with two equal-sized clusters chosen uniformly at random. The general case will be left as homework. Let us first consider the maximum likelihood estimation. In particular, we treat the cluster label vector $x \in \{\pm 1\}^n$ as the parameter to learn from the observation of A. Hence, the ML estimator can be formulated as

$$\begin{aligned} \max_x \ & \mathbb{P}\left[A|x\right] \\ \text{s.t.} \ & x_i \in \{\pm 1\}, \quad i \in [n] \\ & x^\top \mathbf{1} = 0, \end{aligned}$$

where we impose the constraint $x^\top \mathbf{1} = 0$ because we assume the two clusters have equal sizes. Since the edges are independently generated conditionally on the cluster label vector x, the likelihood function has the following simple product form:

$$\begin{aligned} \mathbb{P}\left[A|x\right] &= \prod_{i<j} \mathbb{P}\left[A_{ij}|x_i, x_j\right] \\ &= \prod_{\substack{i<j \\ x_i = x_j}} p^{A_{ij}}(1-p)^{1-A_{ij}} \prod_{\substack{i<j \\ x_i \neq x_j}} q^{A_{ij}}(1-q)^{1-A_{ij}}. \end{aligned}$$

It follows that the log likelihood function is

$$\begin{aligned} \log P(A|x) &= \sum_{\substack{i<j \\ x_i = x_j}} \left(A_{ij}\log p + (1-A_{ij})\log(1-p)\right) + \sum_{\substack{i<j \\ x_i \neq x_j}} \left(A_{ij}\log q + (1-A_{ij})\log(1-q)\right) \\ &= \sum_{i<j} \frac{x_i x_j + 1}{2}\left(A_{ij}\log p + (1-A_{ij})\log(1-p)\right) + \sum_{i<j} \frac{1 - x_i x_j}{2}\left(A_{ij}\log q + (1-A_{ij})\log(1-q)\right) \\ &= \frac{1}{2}\left(\sum_{i<j} A_{ij} x_i x_j\right)\log\frac{p(1-q)}{q(1-p)} + \frac{1}{2}\left(\sum_{i<j} x_i x_j\right)\log\frac{(1-p)}{1-q} + \text{ terms independent of } x, \end{aligned}$$

where the second equality holds because $(x_i x_j + 1)/2 = \mathbf{1}_{\{x_i = x_j\}}$ and $(1 - x_i x_j)/2 = \mathbf{1}_{\{x_i \neq x_j\}}$, and in the third equality, the remaining terms are independent of x. It is simple to check that $\sum_{i<j} x_i x_j = -n/2$ and thus does not depend on x because $x^\top \mathbf{1} = 0$ and $x_i \in \{\pm 1\}$. Hence, in the case of $p > q$, the ML estimation is equivalent to solving

$$\begin{aligned} \max_x \ & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t.} \ & x \in \{\pm\}^n \\ & \sum x_i = 0, \end{aligned} \tag{11.1}$$

which is also equivalent to the following MIN BISECTION problem:

$$\begin{aligned} \min_x \ & \sum_{i,j} A_{ij} \frac{1 - x_i x_j}{2} \\ \text{s.t.} \ & x \in \{\pm\}^n \\ & \sum x_i = 0. \end{aligned}$$

Notice that the objective function $\sum_{i,j} A_{ij} \frac{1-x_i x_j}{2}$ is equal to twice the number of cross edges from cluster $+1$ to cluster -1. Hence, MLE is equivalent to minimum bisection problem, which minimizes the number of crossing edges among all equal-sized bi-partitions. Notice that MIN BISECTION is NP-hard in the worst case. However, since in our problem, the adjacency matrix A is randomly generated according to the stochastic block model. It is still possible to efficiently solve the problem with high probability despite of the NP-hardness in the worst case. This leads to the following interesting question: how to derive a polynomial time procedure from MLE (11.1)?

11.2.1 First idea: spectral relaxation

Recall that we have discussed spectral relaxation of minimizing k-means objective in Lecture 3. Similar idea can be also used to obtain spectral relaxation of MLE (11.1). In particular, we replace the integer constraint $x \in \{\pm 1\}^n$ in (11.1) by the L_2 norm constraint $\|x\|_2 = \sqrt{n}$. This leads to the following constrained eigenvalue maximization problem:

$$\begin{aligned} &\max_x x^\top A x \\ &\text{s.t. } \sum_i x_i = 0 \\ &\|x\|_2 = \sqrt{n} \end{aligned}$$

We can further relax the problem by putting the hard constraint $\sum x_i = 0$ as a Lagrangian regularizer in the objective function:

$$\begin{aligned} &\max_x x^\top \left(A - \frac{p+q}{2}\mathbf{J} + n\mathbf{I} \right) x \\ &\text{s.t.} \quad \|x\|_2 \in \sqrt{n}. \end{aligned} \tag{11.2}$$

The optimal solution of (11.2) is nothing but the leading eigenvector of $A - \frac{p+q}{2}\mathbf{J} + n\mathbf{I}$, and thus can be efficiently computed. Hence, spectral relaxation of MLE recovers the spectral clustering algorithm that we discussed earlier. Notice that the objective function in (11.2) is non-convex in x, and hence spectral relaxation is a non-convex relaxation of MLE.

11.2.2 Second idea: SDP relaxation

In this section, we introduce a convex relaxation of MLE based on semidefinite programming (SDP). First, recall that $x^\top A x = \mathsf{Tr}(x^\top A x) = \mathsf{Tr}(A x x^\top)$. Let $Y = xx^T$. Then $x^\top A x = \langle A, Y \rangle$. Moreover, $Y_{ii} = 1$ is equivalent to $x_i \in \{\pm 1\}$ and $\langle Y, \mathbf{J} \rangle = 0$ is equivalent to $x^\top \mathbf{1} = 0$. Therefore, MLE (11.1) can be equivalently recast as

$$\begin{aligned} &\max_{Y,x} \langle A, Y \rangle \\ &\text{s.t.} \quad Y = xx^\top \\ &\qquad Y_{ii} = 1, \quad i \in [n] \\ &\qquad \langle \mathbf{J}, Y \rangle = 0. \end{aligned} \tag{11.3}$$

Notice that the matrix $Y = xx^\top$ is a rank-one positive semidefinite matrix. If we relax this condition by dropping the rank-one restriction, we obtain the following convex relaxation of MLE, which is a

semidefinite program:

$$\begin{aligned} \widehat{Y} = \arg\max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle \mathbf{J}, Y \rangle = 0. \end{aligned} \tag{11.4}$$

We remark that (11.4) does not rely on any knowledge of the model parameters except that $p > q$; for the case $p < q$, we replace arg max in (11.4) by arg min. The SDP program (11.4) can be efficiently solved in polynomial-time algorithm.

Remark 11.1. Curious reader may wonder why SDP relaxations could succeed in the sparse regime while spectral methods fail. To understand the distinction between SDP relaxation (11.4) and spectral relaxation (11.2), let us further relax the SDP program (11.4) by replacing the diagonal constraint $Y_{ii} = 1, \forall i \in [n]$ by $\mathsf{Tr}(Y) = \sum_i Y_{ii} = n$:

$$\begin{aligned} \max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & \mathsf{Tr}(Y) = n \\ & \langle \mathbf{J}, Y \rangle = 0. \end{aligned}$$

We can further relax the problem by putting the hard constraint $\langle \mathbf{J}, Y \rangle = 0$ into the objective function as a Lagrangian regularizer:

$$\begin{aligned} \max_{Y} \ & \langle A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & \mathsf{Tr}(Y) = n. \end{aligned} \tag{11.5}$$

It is easy to check that (11.5) is equivalent to the spectral relaxation (11.2), More precisely, the optimal solution of (11.5) is given by $Y = nuu^\top$, where u is the leading eigenvector of $A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}$. Therefore, the crucial difference between SDP relaxation (11.4) and spectral relaxation (11.2) is that SDP relaxation constraints $Y_{ii} = 1$ for each i, while the spectral relaxation only constraints $\sum_i Y_{ii} = n$. In the sparse graph regime, the leading eigenvector u of $A - \frac{p+q}{2}\mathbf{J} + p\mathbf{I}$ turns out to be very spiky with few coordinates with extremely large magnitudes $|u_i|$. By letting $Y = nuu^\top$, Y certainly satisfies the constraint $\sum_i Y_{ii} = n$, but violates the constraint $Y_{ii} = 1$ for all $i \in [n]$. Hence, the SDP relaxation has the regularization effect by preventing the spiky eigenvectors from maximizing the objective function.

Next, we introduce the analysis of SDP relaxation (11.4). Let $Y^* = (x^*)(x^*)^\top$ denote the true partition matrix, where x^* is the true cluster label vector. We are interested in the following two questions:

1. When is $\widehat{Y}$ close to Y^*?
2. When is $\widehat{Y}$ exactly equal to Y^*?

In the next section, we address the first question by deriving a high-probability bound on $\|\widehat{Y} - Y^*\|_F^2$. The second question will be addressed in the next lecture.

11.3 Analysis of SDP relaxations for weak recovery

In this section, we present the analysis of SDP relaxations, which is adapted from [GV15]. The key in the analysis is the application of Grothendieck's inequality [Gro53, LP68] that we will introduce later.

Lemma 11.1. *For any positive semidefinite matrix $Y \succeq 0$ such that $Y_{ii} \le 1$ for all $i \in [n]$, we have $|Y_{ij}| \le 1$.*

Proof. Since $Y \succeq 0$, we can write $Y = UU^\top$, where $U \in \mathbb{R}^{n\times n}$ with i-th row of U given by $u_i^\top$. Then $Y_{ij} = \langle u_i, u_j\rangle$. Since $Y_{ii} \le 1$ for all $i \in [n]$, we have $\|u_i\|_2 \le 1$ for all $i \in [n]$. By Cauchy-Schwartz inequality, $|Y_{ij}| \le \|u_i\|_2\|u_j\|_2 \le 1$. □

Lemma 11.2. *Assume $p > q$. Let $\widehat{Y}$ denote the optimal solution to SDP relaxation (11.4). Then*

$$\|\widehat{Y} - Y^*\|_F^2 \le \frac{8}{p-q} \sup_{Y\succeq 0, Y_{ii}=1} \big|\langle A - \bar{A}, Y\rangle\big|.$$

Proof. By the optimality of $\widehat{Y}$, we have

$$\begin{aligned} 0 &\le \langle A, \widehat{Y} - Y^*\rangle \\ &= \langle \bar{A}, \widehat{Y} - Y^*\rangle + \langle A - \bar{A}, \widehat{Y} - Y^*\rangle. \end{aligned} \tag{11.6}$$

For the first term in (11.6), since $\bar{A} = E(A) = \frac{p-q}{2}Y^* + \frac{p+q}{2}\mathbf{J} - p\mathbf{I}$, and $\widehat{Y}$ and Y^* are feasible to SDP relaxation (11.4), it follows that

$$\langle \bar{A}, \widehat{Y} - Y^*\rangle = \frac{p-q}{2}\langle Y^*, \widehat{Y} - Y^*\rangle.$$

where we used the identities $\langle \mathbf{J}, \widehat{Y} - Y^*\rangle = 0$ and $\langle \mathbf{I}, \widehat{Y} - Y^*\rangle = 0$. It follows from the last two displayed equations that

$$\begin{aligned} \frac{p-q}{2}\langle Y^*, \widehat{Y} - Y^*\rangle &\le \langle A - \bar{A}, \widehat{Y} - Y^*\rangle \\ &\le |\langle A - \bar{A}, \widehat{Y}\rangle| + |\langle A - \bar{A}, Y^*\rangle| \\ &\le 2 \sup_{Y\succeq 0, Y_{ii}=1} |\langle A - \bar{A}, Y\rangle|, \end{aligned}$$

where the last inequality follows because both $\widehat{Y}$ and Y^* are feasible solutions. Finally, notice that

$$\|Y^* - \widehat{Y}\|_F^2 = \|Y^*\|_F^2 + \|\widehat{Y}\|_F^2 - 2\langle Y^*, \widehat{Y}\rangle \le 2n^2 - 2\langle Y^*, \widehat{Y}\rangle \le 2\langle Y^*, Y^* - \widehat{Y}\rangle,$$

where we used the fact that $|\widehat{Y}_{ij}| < 1$ which follows from Lemma 11.1. The conclusion readily follows by combining the last two displayed equations. □

Remark 11.2. The proof steps in Lemma 11.2 are used very often in controlling the estimation error of an estimator given by the optimal solution of a certain optimization problem. It starts with the optimality condition, and followed by breaking the objective function difference into two terms: one is associated with the signal part, and the other is associated with the noise part.

In view of Lemma 11.2, to get a high-probability upper bound to $\|\widehat{Y} - Y^*\|_F^2$, we are left to derive a high-probability upper bound to the quantity $\sup_{Y\succeq 0, Y_{ii}=1} \big|\langle A - \bar{A}, Y\rangle\big|$. This quantity is not easy to control, because of the supreme taken over the positive semidefinite cone. The key is to use the Grothendieck's inequality.

Theorem 11.1 (Grothendieck's Inequality)**.** *Suppose $B \in \mathbb{R}^{n\times n}$ satisfies*

$$\sup_{s,t\in\{\pm 1\}^n} |s^\top Bt| \le 1,$$

then there exists a constant K_G such that

$$\sup_{S,T\in\mathbb{R}^{n\times n}} \left|\mathsf{Tr}(S^\top BT)\right| \le K_G,$$

where $S = \begin{pmatrix} S_1^\top \\ \vdots \\ S_n^\top \end{pmatrix}$ *and* $T = \begin{pmatrix} T_1^\top \\ \vdots \\ T_n^\top \end{pmatrix}$ *with* $\|S_i\|_2 = 1$ *and* $\|T_i\|_2 = 1$.

Remark 11.3. Notice that K_G is known as Gothendieck's constant. It has been proved that $K_G \le \frac{\pi}{2ln(1+\sqrt{n})} \le 1.783$. Determining the exact value of K_G is still an open problem. It is instructive to write

$$S^\top BT = (S_1, S_2, ..., S_n)B \begin{pmatrix} T_1^\top \\ \vdots \\ T_n^\top \end{pmatrix} = \sum_{i,j} B_{i,j}\langle S_i, T_j\rangle$$

Comparing to $s^\top Bt = \sum_{i,j} B_{i,j}s_it_j$, we can see that $S^\top BT$ is a multi-dimensional extension of $s^\top Bt$, where for each (i,j) pair, the scalar s_i is replaced by $S_i \in \mathbb{R}^n$ and scalar t_j is replaced by $T_j \in \mathbb{R}^n$. Because of this, by definition $K_G \ge 1$.

It turns out that $\sup_{s,t\in\{\pm1\}^n} |s^\top Bt|$ is equal to $\|B\|_{\infty\to 1}$, where the $\|\cdot\|_{\infty\to 1}$ is defined as follows.

Definition 11.1 (Infinity to one norm)**.** For a matrix $B \in \mathbb{R}^{n\times n}$, $\|B\|_{\infty\to 1}$ is defined as

$$\|B\|_{\infty\to 1} \triangleq \sup_{y\in\mathbb{R}^n, \|y\|_\infty \le 1} \|By\|_1$$

It is simple to check that

$$\|B\|_{\infty\to 1} = \sup_{x,y\in\{\pm1\}^n} |x^\top By|.$$

Armed with the powerful Grothendieck's inequality, it becomes straightforward to bound $\sup_{Y\succeq 0, Y_{ii}=1} \left|\langle B, Y\rangle\right|$ for any symmetric matrix B.

Lemma 11.3. *For any symmetric matrix $B \in \mathbb{R}^{n\times n}$,*

$$\sup_{Y\succeq 0, Y_{ii}=1} \left|\langle B, Y\rangle\right| \le K_G \sup_{x,y\in\{\pm1\}^n} |x^\top By| = K_G\|B\|_{\infty\to 1}.$$

Proof. Since $Y \succeq 0$, we can write $Y = UU^T$ such that

$$U = \begin{pmatrix} U_1^T \\ \vdots \\ U_n^T \end{pmatrix}$$

with $\|U_i\|_2 = 1$. It follows that

$$\begin{aligned}
\sup_{Y\succeq 0, Y_{ii}=1} \big|\langle B, Y\rangle\big| &= \sup_{\|U_i\|_2=1} |\langle B, UU^T\rangle| \\
&\le \sup_{\|U_i\|_2=1} \sum_{i,j} B_{i,j}\langle U_i, U_j\rangle \\
&\le \sup_{\|U_i\|_2=1, \|V_j\|_2=1} \sum_{i,j} B_{i,j}\langle U_i, V_j\rangle \\
&\le K_G \sup_{x,y\in\{\pm 1\}} \sum_{i,j} B_{i,j} x_i y_j,
\end{aligned}$$

where the last inequality follows from Grothendieck's inequality. □

By plugging $B = A - \bar{A}$ into Lemma 11.3, we get the following corollary.

Corollary 11.1.

$$\sup_{Y\succeq 0, Y_{ii}=1} |\langle A - \bar{A}, Y\rangle| \le K_G \sup_{x,y\in\{\pm 1\}^n} |x^T(A-\bar{A})y| = K_G \|A - \bar{A}\|_{\infty\to 1}$$

Remark 11.4. It is instructive to compare $\|A - \bar{A}\|_{\infty\to 1}$ with $\|A - \bar{A}\|_2$ in the sparse regime. Specifically, in the sparse regime with $np = \Omega(1)$ and $np = o(\log n)$, we have argued that $\|A - \bar{A}\|_2 = \omega(\sqrt{np})$. In contrast, $\|A - \bar{A}\|_{\infty\to 1} = O(\sqrt{np})$. In fact, by definition $\|A - \bar{A}\|_{\infty\to 1}$ is the supreme of $x^\top(A-\bar{A})y$ over all $x, y \in \{\pm\}^n$; hence spiky vectors do not affect $\|A - \bar{A}\|_{\infty\to 1}$. On the contrary, as we commented before, those spike vectors indeed induce large values of $\|A - \bar{A}\|_2$. This insensitivity of $\|A - \bar{A}\|_{\infty\to 1}$ with respect to spiky vectors results in the robustness of SDP in the sparse regime.

In view of Corollary 11.1, we are left to bound the quantity $\sup_{x,y\in\{\pm 1\}^n} |x^\top(A - \bar{A})y|$. This quantity can be viewed as the maximum of 2^{2n} dependent random variables, which can be bounded using the union bound together with concentration inequality for each random variable. Notice that for fixed x and y, $x^\top(A - \bar{A})y$ can be written as a sum of independent, bounded random variables; hence we can use the Bernstein's inequality to bound the tail of $x^\top(A - \bar{A})y$. In particular, we have the following lemma.

Lemma 11.4. *With probability at least* $1 - 2(\frac{2}{e})^{2n}$,

$$\sup_{x,y\in\{\pm 1\}^n} |x^T(A - \bar{A})y| \le \left(2\sqrt{2(p+q)n} + \frac{8}{3}\right) n.$$

Proof. As discussed above, the proof is based on Bernstein's inequality and the union bound, and will be left as homework. □

By Combining Lemma 11.2, Corollary 11.1, and Lemma 11.4, we get the following theorem.

Theorem 11.2. *Assume $p > q$. Let $\widehat{Y}$ denote the optimal solution to SDP relaxation (11.4). Then with probability at least* $1 - 2(\frac{2}{e})^{2n}$,

$$\|\widehat{Y} - Y^*\|_F^2 \le \frac{8K_G}{p-q}\left(2\sqrt{2(p+q)n} + \frac{8}{3}\right) n.$$

The above theorem gives a high-probability upper bound on $\|\widehat{Y} - Y^*\|_F^2$. However, $\widehat{Y}$ does not directly yield clusters. Since $\widehat{Y}$ is close to Y^* and the leading eigenvector of Y^* is parallel to the true cluster label vector, we can use the leading eigenvector of $\widehat{Y}$ to estimate clusters, as we did for the spectral clustering. More specifically, we have the following lemma.

Lemma 11.5. *Let u be the eigenvector of $\widehat{Y}$ corresponding to the eigenvalue with the largest magnitude and $\widehat{x} = sign(u)$. Then*

$$\frac{1}{n}\min\{d_H(x^*, \widehat{x}), d_H(-x^*, \widehat{x})\} \leq \frac{8||\widehat{Y} - Y^*||_F^2}{n^2}.$$

We are considering the minimum Hamming distance with respect to x^* and $-x^*$ in Lemma 11.5, because by symmetry, there is no way to correctly tell whether the true cluster label is x^* or $-x^*$. The proof of Lemma 11.5 relies on the Davis-Kahan $\sin\theta$ theorem and will be left as homework.

By combining Theorem 11.2 with Lemma 11.5, we have the following corollary.

Corollary 11.2. *Suppose $\frac{n(p-q)}{\sqrt{np}} \to 0$ and $n(p-q) \to \infty$ as $n \to \infty$. Then*

$$\frac{1}{n}\min\{d_H(x^*, \widehat{x}), d_H(-x^*, \widehat{x})\} \to 0.$$

11.4 Two recovery goals

In the previous lectures, we have been focusing on achieving a vanishing fraction of misclassified nodes. In this section, we formally introduce two recovery goals. Notice that it is impossible to tell whether the true cluster label is x^* or $-x^*$.

Definition 11.2 (Weak Recovery)**.** Let d_H denote the Hamming distance. In the binary stochastic block model with true cluster label vector denoted by x^*, we say an estimated cluster label vector $\widehat{x}$ achieves *weak recovery*, if as $n \to \infty$

$$\frac{1}{n}\min\{d_H(x^*, \widehat{x}), d_H(-x^*, \widehat{x})\} \to 0, \text{ in probability.} \tag{11.7}$$

Since $0 \leq \frac{1}{n}\min\{d_H(x^*, \widehat{x}), d_H(-x^*, \widehat{x})\} \leq 1$, convergence in probability is equivalent to convergence in L_1. In other words, (11.7) is equivalent to

$$\frac{1}{n}\mathbb{E}\left[\min\{d_H(x^*, \widehat{x}), d_H(-x^*, \widehat{x})\}\right] \to 0.$$

Definition 11.3 (Exact Recovery)**.** In the binary stochastic block model with true cluster label vector denoted by x^*, we say an estimated cluster label vector $\widehat{x}(A)$ achieves *exact recovery*, if as $n \to \infty$,

$$\mathbb{P}\left[\widehat{x} \in \{x^*, -x^*\}\right] \to 1.$$

For weak recovery, we have shown that:

- Spectral clustering achieves weak recovery in the dense regime $np = \Omega(\log n)$ if $\frac{n(p-q)}{\sqrt{np}} \to \infty$.
- SDP achieves weak recovery if $\frac{n(p-q)}{\sqrt{np}} \to \infty$ and $n(p-q) \to \infty$.

In the next lecture, we will discuss the information-theoretic limits of weak recovery and exact recovery, as well as how to achieve exact recovery efficiently using SDP.

Chapter 12

Exact recovery via SDP clustering

This lecture:

- Weak and exact Recovery
- When does SDP achieve exact recovery

12.1 Review of weak and exact recovery

Consider the binary symmetric SBM with n nodes and two equal-sized clusters, where $x^* \in \{\pm 1\}^n$ denotes the true cluster label vector. Recall the following two notations of recovery of x^* based on the observed adjacency matrix A.

Definition 12.1. (Weak Recovery) Let $\widehat{x}(A)$ be the estimated cluster label and $x^*(A)$ be the true cluster label. We say $\widehat{x}(A)$ achieves weak recovery of $x^*(A)$ if as $n \to \infty$,

$$\frac{1}{n}\min\{d_H(\widehat{x}, x^*), d_H(\widehat{x}, -x^*)\} \longrightarrow 0 \ ; \ \text{converge in probability}$$

Note : Since $\frac{1}{n}\min\{d_H(\widehat{x}, x^*), d_H(\widehat{x}, -x^*)\}$ is bounded, convergence in probability implies L_1 convergence, i.e.

$$\frac{1}{n}\mathbb{E}[\min\{d_H(\widehat{x}, x^*), d_H(\widehat{x}, -x^*)\}] \longrightarrow 0.$$

Definition 12.2. (Exact Recovery) We say $\widehat{X}(A)$ achieves exact recovery of X^* if

$$\mathbb{P}[\widehat{x} \in \{x^*, -x^*\}] \longrightarrow 1; \ \text{ as n} \longrightarrow \infty$$

Clearly, exact recovery is stronger than weak recovery.

Recall : It was shown in the last lecture that SDP achieves weak recovery if

$$n(p-q) \longrightarrow \infty \text{ and } \frac{n(p-q)}{\sqrt{np}} \longrightarrow \infty$$

In this lecture, we will study when SDP achieves exact recovery.

12.2 Exact Recovery via SDP

Recall the SDP relaxations of MLE derived in the last lecture:

$$\begin{aligned} \max_{Y} \quad & \langle A, Y \rangle \\ \text{s.t.} \quad & Y \succeq 0 \\ & Y_{ii} = 1 \\ & \langle Y, \mathbf{J} \rangle = 0 \end{aligned} \tag{12.1}$$

Let $Y^* = x^*(x^*)^\top$ be a true cluster partition matrix, and denote $\widehat{Y}$ as the optimal solution of (12.1). Notice that one can recover x^* up to a global sign flip from Y^*. Hence, The exact recovery of x^* is equivalent to the exact recovery of Y^*. Therefore, we would like to understand when $\mathbb{P}[\widehat{Y} = Y^*] \to 1$ as $n \to \infty$ (*exact recovery*).

We assume $p = \frac{a\log(n)}{n}$ and $q = \frac{b\log(n)}{n}$, where $a > b > 0$ are two fixed constants. Notice that if the observed graph has isolated nodes (nodes with zero degree), then there is no way to tell which cluster those isolated nodes are from and hence exact recovery is fundamentally impossible. It turns out that p has to scale as $\frac{\log(n)}{n}$ so that with high probability, the observed graph does not contain any isolated node. More pecisely, we have the following claim.

Lemma 12.1. *Consider the binary symmetric SBM with two equal-sized clusters. If $\frac{a+b}{2} < 1$, then w.h.p., there are about $n^{1-\frac{a+b}{2}}$ isolated nodes.*

The proof will be left as homework excercise. In fact, it is easy to compute the expected number of isolated nodes. Specifically,

$$\begin{aligned} \mathbb{P}(d_i = 0) &= (1-p)^{\frac{n}{2}-1}(1-q)^{\frac{n}{2}}; \\ &= \left(1 - \frac{a\log(n)}{n}\right)^{\frac{n}{2}-1} \left(1 - \frac{b\log(n)}{n}\right)^{\frac{n}{2}}; \quad (1-x) \sim e^{-x} \text{for small } x \\ &= \exp\left(-\left(\frac{n}{2} - 1\right)\frac{a\log(n)}{n} - \frac{n}{2}\frac{b\log(n)}{n}\right) \\ &= \exp\left(-\frac{a+b}{2}\log(n)\right) \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{E}[\text{number of isolated nodes}] &= \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}_{d_i=0}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}_{d_i=0}\right] \\ &= \sum_{i=1}^{n} \mathbb{P}[d_i = 0] \\ &= n \times \exp\left(\left(-\frac{a+b}{2}\right)\log(n)\right) \\ &= n^{1-\frac{a+b}{2}}. \end{aligned} \tag{12.2}$$

Remark 12.1. We have argued that in expectation, there are about $n^{1-\frac{a+b}{2}}$ isolated nodes. Such an argument is known as *first moment method*. Notice that we cannot immediately conclude that with high probability, there are about $n^{1-\frac{a+b}{2}}$ isolated nodes. It is because there might be some rare event of small probability with overwhelmingly large number of isolated nodes, and on all the other events, the number of isolated nodes is small. To exclude such an rare event exists, we need to use the *second moment method*.

Lemma 12.1 implies that $\frac{a+b}{2} < 1$ is necessary for exact recovery. Is this condition sufficient? It turns out that exact recovery requires a stronger condition.

Theorem 12.1. *If $\sqrt{a} - \sqrt{b} > \sqrt{2}$, then $\mathbb{P}[\widehat{Y} = Y^*] = 1 - n^{-\Omega(1)}$ as $n \longrightarrow \infty$.*

Theorem 12.2. *If $\sqrt{a} - \sqrt{b} < \sqrt{2}$ then for any possible estimator $\widehat{Y}(A)$, $\mathbb{P}[\widehat{Y} = Y^*] \to 0$ as $n \to \infty$.*

Theorem 12.1 and Theorem 12.2 together imply that SDP attains the optimal exact recovery threshold. Next, we will first present the proof of Theorem 12.1. The high-level proof idea is to construct a set of dual variables and show that they together with Y^* satisfy the KKT conditions of the SDP program with high probability.

12.2.1 Dual certificate lemma and its proof

The following lemma provides a set of determinstic conditions on the dual variables to guarantee that Y^* is the unique solution to the SDP.

Lemma 12.2. ***(Dual Certificate Lemma)*** *Suppose there exists $D^* = diag(d_i^*)$ and $\lambda^* \in \mathbb{R}$ such that $S^* = D^* - A + \lambda^* \mathbf{J}$ satisfies*

$$S^* \succeq 0 \tag{12.3}$$
$$\lambda_2(S^*) > 0 \tag{12.4}$$
$$S^* x^* = 0, \tag{12.5}$$

where $\lambda_2(S^*)$ denotes the second smallest eigenvalue of S^*. Then Y^* is the unique optimal solution to the SDP.

The conditions in the dual cerficate lemma in fact correspond to the so-called *KKT conditions* of the SDP program. More specifically, define the dual variables corresponding to each constraint of the SDP as follows.

$$\begin{aligned} \max_{Y} \quad & \langle A, Y \rangle & \\ \text{s.t.} \quad & Y \succeq 0 & \longrightarrow S \succeq 0 \\ & Y_{ii} = 1 & \longrightarrow D = diag(d_i) \\ & \langle Y, \mathbf{J} \rangle = 0 & \longrightarrow \lambda \in \mathbb{R} \end{aligned}$$

Then the so-called *Lagrangian function* is defined as:

$$L(Y, S, D, \lambda) \triangleq \langle A, Y \rangle + \langle S, Y \rangle - \langle D, Y - \mathbf{I} \rangle - \lambda \langle Y, \mathbf{J} \rangle \tag{12.6}$$

The KKT conditions are given by

1. First order condition: $\frac{\partial L(Y)}{\partial Y}|_{Y=Y^*} = A + S - D - \lambda \mathbf{J} = 0 \iff S = D - A + \lambda \mathbf{J}$
2. Complementary Slackness: $\langle S, Y^* \rangle = 0$. Note that $Y^* = x^*(x^*)^\top$ and $S \succeq 0$. It follows that $\langle S, Y^* \rangle = 0 \iff Sx^* = 0$.

The extra condition $\lambda_2(S^*) > 0$ in the dual certficate lemma is used to guarantee that the optimal solution of SDP is unique.

Proof of Dual Certificate Lemma. For any feasible solution Y of SDP,

$$L(Y, S^*, D^*, \lambda^*) = \langle A, Y\rangle + \langle S^*, Y\rangle - \langle D^*, Y - \mathbf{I}\rangle - \lambda^*\langle Y, \mathbf{J}\rangle.$$

Since Y is feasible, it follows that

$$\langle D^*, Y - I\rangle = \sum_i d_i^*(Y_{ii} - 1) = 0,$$

and $\langle Y, \mathbf{J}\rangle = 0$. Moreover, since $Y \succeq 0$, its eigenvalue decomposition can be written as

$$Y = \sum_{i=1}^{n} \lambda_i u_i u_i^T \ \ ; 0 \le \lambda_1 \le \lambda_2 \cdots \le \lambda_n,$$

Hence,

$$\begin{aligned}\langle S^*, Y\rangle &= \sum_{i=1}^{n} \lambda_i \langle u_i u_i^T, S\rangle \\ &= \sum_{i}^{n} \lambda_i u_i^T S u_i \ge 0,\end{aligned}$$

where the last equality holds by trace operation, which states $\mathrm{Tr}(u_i u_i^T S) = \mathrm{Tr}(u_i^T S u_i) = u_i^T S u_i$. Using these results, we have

$$\begin{aligned}\langle A, Y\rangle &\le L(Y, S^*, D^*, \lambda^*) \\ &= \langle S^* - D^* + A - \lambda^*\mathbf{J}, Y\rangle + \langle D^*, I\rangle \\ &= \langle D^*, I\rangle \\ &= \langle D^*, Y^*\rangle \\ &= \langle S^* + A - \lambda^*\mathbf{J}, Y^*\rangle \ \ ; \ \ \text{since} \langle S^*, Y^*\rangle = \langle \mathbf{J}, Y^*\rangle = 0 \\ &= \langle A, Y^*\rangle\end{aligned}$$

As a result, $\langle A, Y\rangle \le \langle A, Y^*\rangle$ and hence Y^* is the optimal.

Next we show the uniqueness of Y^*. Let $\widetilde{Y}$ be an optimal solution to the SDP, one wants to show that $\widetilde{Y} = Y^*$. Notice that

$$\begin{aligned}\langle S^*, \widetilde{Y}\rangle &= \langle D^* - A + \lambda^*\mathbf{J}, \widetilde{Y}\rangle \\ &= \langle D^* - A, \widetilde{Y}\rangle \\ &= \langle D^* - A, Y^*\rangle \ \ ; \ \ \text{since} \langle A, \widetilde{Y}\rangle = \langle A, Y^*\rangle \ \text{ and } \ D^* \ \text{ is diagonal.} \\ &= \langle S^*, Y^*\rangle \\ &= 0.\end{aligned}$$

Since $S^* \succeq 0$ and $\lambda_2(S^*) > 0$, $\widetilde{Y}$ must be a multiple of $x^*(x^*)^\top$. Moreover, since $\widetilde{Y}_{ii} = 1$, $\widetilde{Y} = x^*(x^*)^\top = Y^*$. □

12.2.2 Proof of Theorem 12.1

With the dual cerficiate lemma, we are ready to prove Theorem 12.1.

Proof of Theorem 12.1. We use the same notation as dual certificate lemma. Note that $S^* = D^* - A + \lambda^* \mathbf{J}$. Then $S^* x^* = 0$ is equivalent to

$$0 = D^* x^* - A x^* + \lambda^* \mathbf{J} x^*.$$

Since x^* is assumed to represent equal-sized partition, $\mathbf{J}x^* = 0$, thus, $D^* x^* = A x^*$. It follows that

$$d_i^* x_i^* = \sum_{j \neq i} A_{ij} x_j^*.$$

Since $x_i^*, x_j^* \in \{\pm 1\}$, we have

$$\begin{aligned} d_i^* &= \sum_{j \neq i} A_{ij} x_j^* x_i^* \\ &= \left[\sum_{j \neq i, x_j^* = x_i^*} A_{ij} - \sum_{j \neq i, x_j^* \neq x_i^*} A_{ij} \right]. \end{aligned}$$

Note that

$$\sum_{j \neq i, x_j^* = x_i^*} A_{ij} \sim \mathrm{Bin}(\frac{n}{2} - 1, p),$$

and

$$\sum_{j \neq i, x_j^* \neq x_i^*} A_{ij} \sim \mathrm{Bin}(\frac{n}{2}, q).$$

So we are left to check that with high probability, $\lambda_2(S^*) > 0$, or more specifically,

$$\mathbb{P}\left[\inf_{x \perp x^*, \|x\|_2 = 1} x^T S^* x > 0 \right] \geq 1 - n^{-\Omega(1)}. \tag{12.7}$$

Fix any $x \perp x^*, \|x\|_2 = 1$. By the definition of S^*, we have

$$\begin{aligned} x^T S^* x &= x^T (D^* - A + \lambda^* \mathbf{J}) x \\ &= x^T (D^* - \bar{A} + \lambda^* \mathbf{J}) x + x^T (\bar{A} - x) x, \end{aligned} \tag{12.8}$$

where

$$\bar{A} \triangleq \mathbb{E}[A] = \frac{p-q}{2} Y^* + \frac{p+q}{2} \mathbf{J} - p\mathbf{I}.$$

Therefore, it follows from (12.8) that

$$\begin{aligned} x^T S^* x &= x^T \left(D^* - \frac{p-q}{2} Y^* - \frac{p+q}{2} \mathbf{J} + p\mathbf{I} + \lambda^* \mathbf{J} \right) x + x^T (\bar{A} - x) x \\ &= x^T D^* x + \left(\lambda^* - \frac{p+q}{2} \right) x^T \mathbf{J} x + p + x^T (\bar{A} - A) x. \end{aligned} \tag{12.9}$$

Since $x^T\mathbf{J}x$ is always nonnegative, by choosing any λ^* such that $\lambda^* \geq \frac{p+q}{2}$, it follows from (12.9) that

$$\begin{aligned} x^T S^* x &\geq x^T D^* x + x^T(\bar{A} - A)x + p. \\ &\geq x^T D^* x + p - \|A - \bar{A}\|_2 \\ &\geq \min_i d_i^* + p - \|A - \bar{A}\|_2, \end{aligned}$$

where the second last inequality is due to $x^T(\bar{A} - A)x \geq \lambda_{\min}(\bar{A} - A) \geq -\|\bar{A} - A\|_2$. Next, there are two tasks remaininig. In particular, we need to argue that with high probability,

1. $\min_i d_i^*$ is "large".
2. $\|A - \bar{A}\|_2$ is "small".

First Task

Let's first compute the expectation of d_i^*:

$$\begin{aligned} \mathbb{E}[d_i^*] &= \mathbb{E}\left[\sum_{j\neq i, x_j \neq x_i} A_{ij} - \sum_{j\neq i, x_j \neq x_i} A_{ij}\right] \\ &= \left(\frac{n}{2} - 1\right) p - \frac{n}{2} q \\ &= \left(\frac{n}{2} - 1\right) \frac{a \log n}{n} - \frac{n}{2} \frac{b \log n}{n} \sim \frac{1}{2}(a - b) \log n. \end{aligned}$$

Therefore, in expectation, d_i^* is on the order of $\log n$. The next lemma bounds the tail probability of $d_i^* \leq \frac{\log n}{\log \log n}$ for each fixed i.

Lemma 12.3. *Let $X \sim Bin\left(\frac{n}{2} - 1, \frac{a \log n}{n}\right)$, $R \sim Bin\left(\frac{n}{2}, \frac{b \log n}{n}\right)$, and assume that these two random variables are independent (i.e., $X \perp R$). Then it follows that*

$$\mathbb{P}\left\{X - R \leq \frac{\log n}{\log \log n}\right\} \leq n^{-\frac{(\sqrt{a} - \sqrt{b})^2}{2} + o(1)},$$

The proof of the lemma is based on Chernoff's bound and will be left as homework. By Lemma 12.3, we have

$$\begin{aligned} \mathbb{P}\left\{\min_i d_i \leq \frac{\log n}{\log \log n}\right\} &\leq \mathbb{P}\left\{\exists i \text{ s.t. } d_i^* \leq \frac{\log n}{\log \log n}\right\} \\ &\leq \sum_{i=1}^n \mathbb{P}\left\{d_i^* \leq \frac{\log n}{\log \log n}\right\} \\ &= n \cdot n^{-\frac{(\sqrt{a} - \sqrt{b})^2}{2} + o(1)} \\ &= n^{1 - \frac{(\sqrt{a} - \sqrt{b})^2}{2} + o(1)} = n^{-\Omega(1)}. \end{aligned}$$

Since $\sqrt{a} - \sqrt{b} > \sqrt{2}$ by the assumption, with probability at least $1 - n^{-\Omega(1)}$, $\min_i d_i^* \geq \frac{\log n}{\log \log n}$.

Second Task

In homework 4, we have proved a high-probability upper bound on $\|A - \bar{A}\|_2$. In particular,

Lemma 12.4. *For any $c > 0$, there exists $c' > 0$ such that*

$$\mathbb{P}\left\{\|A - \bar{A}\|_2 \le c'\sqrt{np}\right\} \ge 1 - n^{-c}.$$

Proof of positiveness of $\lambda_2(S)$

Now, we are ready to finish the proof of (12.7). Define event $\mathcal{E}$ as

$$\mathcal{E} \triangleq \{\min_i d_i^* \ge \frac{\log n}{\log\log n}\} \cap \{\|A - \bar{A}\|_2 \le c'\sqrt{np}\}.$$

Since

$$\mathbb{P}\left\{\min_i d_i^* \le \frac{\log n}{\log\log n}\right\} \le n^{-\Omega(1)}$$

and

$$\mathbb{P}\left\{\|A - \bar{A}\|_2 \ge c'\sqrt{np}\right\} \le n^{-c},$$

it follows from union bound that $\mathbb{P}[\mathcal{E}] \ge 1 - n^{-\Omega(1)} - n^{-c}$. On event $\mathcal{E}$, for all sufficiently large n,

$$\inf_{x \perp x^*, \|x\|_2 = 1} x^T S^* x \ge \frac{\log n}{\log\log n} + p - c'\sqrt{\log n} \ge p, \tag{12.10}$$

Hence on event $\mathcal{E}$, all the conditions of dual certificate lemma are satisifed and thus $\widehat{Y} = Y^*$. Therefore,

$$\mathbb{P}\left[\widehat{Y} = Y^*\right] \ge \mathbb{P}[\mathcal{E}] \ge 1 - n^{-\Omega(1)} - n^{-c} = 1 - n^{-\Omega(1)}.$$

□

Note: Let us summarize what we have ve discussed so far.

- Dual certificate lemma provides a set of deterministic conditions to cerfity Y^* is the unique optimal solution to the SDP program.
- We construct D^*, λ^*, and S^*, which satisfy the conditions in the dual certificate lemma. In particular, $S^*x^* = (D^* - A + \lambda^* J)x^* = 0$ is equivalent to $d_i^* = \sum_{j \ne i} A_{ij} x_i^* x_j^*$. Notice that d_i^* corresponds to the number of node i's neighors in its own cluster minus the number of node i's neigbhors in the other cluster. As we will show later, d_i^* is closely related to the information-theoretic lower bounds of exact recovery. In particular, we will prove that if there exists a node i from cluster $+$ and node j from cluster $-$ such that $d_i^* < -1$ and $d_i^* < -1$, then the maximum likelihood estimator of x^* will not coincide with the true cluster label vector x^*.

12.3 Information-theoretic lower bounds for exact recovery

In this section, we aim to prove Theorem 12.2. We start by proving that the maximum likelihood estimator is always optimal in minizing the probability of error when the underlying parameter is uniformly distributed.

Let us consider a general parameter estimaiton setting. Let θ^* denote the true parameter which is drawn from a parameter space Θ. Let Y denote the observation and $p(y|\theta)$ denote a conditional probability kernal such that $Y \sim p(\cdot|\theta^* = \theta)$. The goal is to estimate θ^* based on Y. More formally, we aim to develop an estimator $\widehat{\theta}(Y)$ as a funcdtion of Y so that $\widehat{\theta}(Y)$ is close to θ^*.

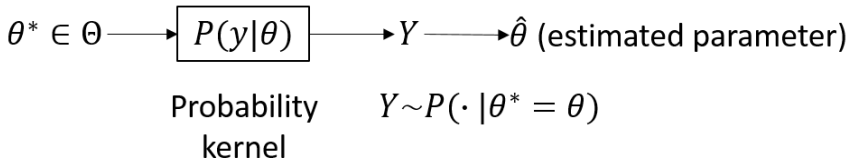


Figure 12.1: General parameter estimation setting

Example 12.1. (Binary symmetric SBM)
Let's see how binary symmetric SBM is adapted to the general parameter estimation procedure. Note that Y in the Figure 12.1 is different from our partition matrix.

- θ^*: Y^* (partition matrix for underlying two clusters)
- Θ: $\mathcal{Y} = \left\{Y : Y = xx^T, x \in \{\pm 1\}^n, \sum_{i=1}^n x_i = 0\right\}$ (the set of all the possible partition matrices)
- $P(y \mid \theta)$: $P(A \| Y^*)$ (The probailistic way of genearating graph according to SBM)
- Y: adjacency matrix A
- $\widehat{\theta}$: estimated partition matrix $\widehat{Y}$

As we introduced before, the maximum likelihood estimator is defined as $\widehat{\theta}_{ML}(y) = \arg\max_{\theta \in \Theta} P(y \mid \theta)$. The following theorem implies that the maximum likelihood estimator minimizes the probability of error when θ^* is uniformly generated from Θ.

Theorem 12.3. *Suppose θ^* is uniformly generated from Θ. Then for any estimator $\widehat{\theta}(y)$,*

$$\mathbb{P}\left\{\widehat{\theta}(Y) = \theta^*\right\} \leq \mathbb{P}\left\{\widehat{\theta}_{ML}(Y) = \theta^*\right\}.$$

In the next lecture, we will prove this theorem and use it to prove Theorem 12.2.

Chapter 13

Information-theoretic lower bounds for exact and weak recovery

Outline

- Information-theoretic lower bounds for exact recovery
- Information-theoretic lower bounds for weak recovery

13.1 Information-theoretic lower bounds for exact recovery

Let us first consider a general setting. Let θ^* denote the true parameter drawn from a parameter space Θ. Given $\theta^* = \theta$, observation Z is generated according to a conditional probability distribution $P(z|\theta)$. The goal is to estimate θ^* based on observation Z. In particular, we would like to come up an estimator $\widehat{\theta}(Z)$, which is as close to the ground truth θ^* as possible.

$$\theta^* \in \Theta \longrightarrow \boxed{p(z|\theta)} \longrightarrow Z \longrightarrow \widehat{\theta}(Z)$$

As we have discussed in the previous lectures, a classical estimator is the *maximum likelihood estimator* which is given by

$$\widehat{\theta}_{\mathrm{ML}}(Z) \in \arg\max_{\theta \in \Theta} \mathbb{P}\left[Z|\theta^* = \theta\right].$$

The following theorem shows that if θ^* is uniformly generated from Θ, then ML estimator minimizes the probability error among all possible estimators.

Theorem 13.1. *If θ^* is uniformly generated from Θ, then for any estimator $\widehat{\theta}(Z)$,*

$$\mathbb{P}\left[\widehat{\theta}(Z) = \theta^*\right] \leq \mathbb{P}\left[\widehat{\theta}_{\mathrm{ML}}(Z) = \theta^*\right].$$

Proof. In the following proof, we assume observation Z is discrete for ease of exposition; the same

proof holds for continuous observation as well. Note that

$$\begin{aligned}
\mathbb{P}\left[\widehat{\theta}(Z)=\theta^*\right] &\overset{(a)}{=} \sum_{\theta\in\Theta}\mathbb{P}\left[\widehat{\theta}(Z)=\theta|\theta^*=\theta\right]\mathbb{P}\left[\theta^*=\theta\right] \\
&\overset{(b)}{=} \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\mathbb{P}\left[\widehat{\theta}(Z)=\theta|\theta^*=\theta\right] \\
&\overset{(c)}{=} \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\sum_{z}\mathbb{P}\left[Z=z|\theta^*=\theta\right]\mathbf{1}_{\left\{\widehat{\theta}(z)=\theta\right\}} \\
&= \frac{1}{|\Theta|}\sum_{z}\sum_{\theta\in\Theta}\mathbb{P}\left[Z=z|\theta^*=\theta\right]\mathbf{1}_{\left\{\widehat{\theta}(z)=\theta\right\}}
\end{aligned}$$

where (a) holds dues to total probability formula; (b) follows because θ^* is uniformly distribution by assumption. For any fixed z, since $\sum_{\theta\in\Theta}\mathbf{1}_{\left\{\widehat{\theta}(z)=\theta\right\}}=1$, it follows that

$$\sum_{\theta\in\Theta}\mathbb{P}\left[Z=z|\theta^*=\theta\right]\mathbf{1}_{\left\{\widehat{\theta}(z)=\theta\right\}} \leq \max_{\theta\in\Theta}\mathbb{P}\left[Z=z|\theta^*=\theta\right],$$

where the equality is achieved when

$$\widehat{\theta}(z)\in\arg\max_{\theta\in\Theta}\mathbb{P}\left[Z=z|\theta^*=\theta\right].$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left[\widehat{\theta}(Z)=\theta^*\right] &\leq \frac{1}{|\Theta|}\sum_{z}\max_{\theta\in\Theta}\mathbb{P}\left[Z=z|\theta^*=\theta\right] \\
&= \frac{1}{|\Theta|}\sum_{z}\sum_{\theta\in\Theta}\mathbb{P}\left[Z=z|\theta^*=\theta\right]\mathbf{1}_{\left\{\widehat{\theta}_{\mathrm{ML}}(z)=\theta\right\}} \\
&= \mathbb{P}\left[\widehat{\theta}_{\mathrm{ML}}(Z)=\theta^*\right].
\end{aligned}$$

□

13.1.1 A common strategy for proving lower bounds for exact recovery

Recall that exact recovery requires that the probability of error vanishes. Since we have shown that if θ^* is uniformly generated from Θ, then ML estimator minimizes the probability error among all possible estimators. Hence, to prove a lower bound for exact recovery, one can assume a uniform prior of θ^* and prove ML estimator fails under a certain condition. More specifically, we have the following routine to prove a lower bound for exact recovery.

1. Suppose θ^* is uniformly generated form Θ.
2. Identify an event $\mathcal{F}$ which is defined on the joint probability space of (θ^*, Z), such that for any $(\theta, z)\in\mathcal{F}$, there exists $\theta'\in\Theta, \theta'\neq\theta$ such that

$$\mathbb{P}\left[Z=z|\theta^*=\theta\right] < \mathbb{P}\left[Z=z|\theta^*=\theta'\right].$$

Note: If we can find such an event $\mathcal{F}$, then for any realization $(\theta, z) \in \mathcal{F}$,

$$\begin{aligned}
&\Longrightarrow \widehat{\theta}_{\mathrm{ML}}(z) \neq \theta \\
&\Longrightarrow \mathcal{F} \subseteq \{\widehat{\theta}_{\mathrm{ML}}(Z) \neq \theta^*\} \\
&\Longrightarrow \mathbb{P}[\mathcal{F}] \leq \mathbb{P}\left[\widehat{\theta}_{\mathrm{ML}}(Z) \neq \theta^*\right].
\end{aligned}$$

3. Show that $\mathbb{P}[\mathcal{F}] \geq \delta$ for some constant $\delta > 0$.

Note: If we can successfully complete all above three steps, then we have shown that the probability of error of maximum likelihood estimation is lower bounded by some positive constant δ. This immediately implies that it is impossible for maximum likelihood estimator to achieve a vanishing probability of error and hence the impossibility of exact recovery. The key in the above three steps is to find such an event $\mathcal{F}$ with $\mathbb{P}[\mathcal{F}] \geq \delta$, i.e., to find a *typical* failing event of maximum likelihood estimation.

13.2 Information-theoretic lowers bound for exact recovery under binary symmetric SBM

In this section, we illustrate how to carry out the common strategy for proving exact recovery lower bounds by considering the binary symmetric SBM.

Theorem 13.2. *Consider binary symmetric SBM $\mathcal{G}(n, p, q)$ with two equal-sized clusters. Let Y^* denote the partition matrix corresponding to the true cluster partition. Assume $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$ with $a > b \geq 0$, where a and b are two constants. If $\sqrt{a} - \sqrt{b} < \sqrt{2}$, then for any estimator $\widehat{Y}$, as $n \to \infty$,*

$$\mathbb{P}\left[\widehat{Y} = Y^*\right] \to 0.$$

Note: Recall that we have shown in the previous lectures that SDP achieves exact recovery if $\sqrt{a} - \sqrt{b} > \sqrt{2}$. Hence, the achievability result of SDP and impossibility result in Theorem 13.2 together imply that the information-theoretic limit for exact recovery under binary symmetric SBM is $\sqrt{a} - \sqrt{b} = \sqrt{2}$, and it can be achieved efficiently in polynomial-time.

Proof. Let $\mathcal{Y} = \{Y : Y = xx^\top, x \in \{\pm 1\}^n, \sum_i x_i = 0\}$ denote the set of all possibly partition matrices Assume Y^* is uniformly generated from $\mathcal{Y}$.

We would like to identify an event $\mathcal{F}$ such that on this event $\mathcal{F}$, the ML estimator does not coincide with Y^*. Recall that the ML estimation is to maximize the number of edges inside clusters. If in the observed graph, one can find a node i in cluster $+1$ with more neighbors in cluster -1 than cluster $+1$, and similarly find a node j in cluster -1 with more neighbors in cluster $+1$ than cluster -1, then one can swap the cluster memberships of those two nodes and increase the number of edges inside clusters. Therefore, the existence of such two nodes i and j corresponds to a failing event of ML estimator. More formally, we define event $\mathcal{F}$ as follows.

Let $d_i \triangleq \sum_{j \neq i} A_{ij} x_i^* x_j^*$, x^* denotes the true cluster label vector, and $Y^* = x^*(x^*)^\top$. Then d_i equals to the number of neighbors in its own cluster minus the numbers of neighbors in the other cluster. Define

$$\begin{aligned}
\mathcal{F}_1 &= \{\exists i : x_i^* = +1, d_i \leq -2\}, \\
\mathcal{F}_2 &= \{\exists j : x_i^* = -1, d_j \leq -2\}.
\end{aligned}$$

and let $\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2$. Next, we verify that on this event $\mathcal{F}$, ML estimator will not coincide with the true partition Y^*. Indeed, on event $\mathcal{F}$, there must a node $\exists i_0$ such that $x^*_{i_0} = +1, d_{i_0} \le -2$ and a node j_0 such that $x^*_{j_0} = -1, d_{j_0} \le -2$. Define

$$x'_i = \begin{cases} -x^*_i & \text{if } i \in \{i_0, j_0\} \\ x^*_i & \text{o.w} \end{cases}$$

and $Y' = (x')(x')^\top$. Notice that x' is a new cluster label vector by swapping the cluster memberships of nodes i_0 and j_0. Hence $Y' \in \mathcal{Y}$ corresponds to a new partition matrix. We are left to check that Y' indeed has a higher likelihood than Y^*:

$$\begin{aligned} \log \mathbb{P}\left[A|Y^*\right] - \log \mathbb{P}\left[A|Y'\right] &= \frac{1}{2} \log \frac{p(1-q)}{q(1-p)} \left(\sum_{i<j} A_{ij}(x^*_i x^*_j - x'_i x'_j) \right) \\ &= \log \frac{p(1-q)}{q(1-p)} \left(d_{i_0} + d_{j_0} + 2A_{i_0 j_0}\right) < 0 \end{aligned}$$

where the last inequality follows because $d_{i_0} \le -2$ and $d_{j_0} \le -2$.

Finally, we show that event $\mathcal{F}$ is typical, i.e., if $\sqrt{a} - \sqrt{b} < \sqrt{2}$, then $\mathbb{P}[\mathcal{F}] \to 1$ as n $\to \infty$. Note that d_i corresponds to the difference of two independent binomial random variables:

$$d_i \sim \text{Binom}\left(\frac{n}{2}, \frac{a \log n}{n}\right) - \text{Binom}\left(\frac{n}{2}, \frac{b \log n}{n}\right).$$

Then $\mathbb{E}[d_i] \sim \frac{a-b}{2} \log n$. The following proposition shows that there is a still a small chance that $d_i \le -\frac{\log n}{\log \log n}$.

Proposition 13.1.

$$\mathbb{P}\left[d_i \le -\frac{\log n}{\log \log n}\right] \ge n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{2} + o(1)}.$$

The following argument shows that with high probability, there exists a node i such that $d_i \le -\frac{\log n}{\log \log n}$, by assuming d_i's are independent (notice that d_i's are in fact dependent; we will see how to relax this assumption later):

$$\begin{aligned} &\mathbb{P}\left[\exists i : x^*_i = +1, d_i \le -2\right] \\ &= 1 - \mathbb{P}\left[\forall i \text{ with } x^*_i = +1 : d_i > -2\right] \\ &= 1 - \prod_{i: x^*_i = +1} \mathbb{P}\left[d_i > -2\right] \\ &= 1 - \prod_{i: x^*_i = +1} \left(1 - \mathbb{P}\left[d_i \le -2\right]\right) \\ &\overset{(a)}{\ge} 1 - \prod_{i: x^*_i = +1} \exp\left(-\mathbb{P}\left[d_i \le -2\right]\right) \\ &\overset{(b)}{\ge} 1 - \prod_{i: x^*_i = 1} \exp\left(-n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{2} + o(1)}\right) \\ &= 1 - \exp\left(-\frac{1}{2} n^{1 - \frac{(\sqrt{a}-\sqrt{b})^2}{2} + o(1)}\right) \to 1, \end{aligned} \tag{13.1}$$

where (a) holds due to $1-x \le e^{-x}$; (b) holds due to

$$\mathbb{P}\left[d_i \le -2\right] \ge \mathbb{P}\left[d_i \le -\frac{\log n}{\log\log n}\right] \ge n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{2}+o(1)}.$$

Hence, $\mathbb{P}\left[\mathcal{F}_1\right] \to 1$. Similarly, one can show that $\mathbb{P}\left[\mathcal{F}_2\right] \to 1$. Since $\mathcal{F} = \mathcal{F}_1 \cap \mathcal{F}_2$, it follows that $\mathbb{P}\left[\mathcal{F}\right] \to 1$.

□

Remark 13.1. Proposition 13.1 implies that

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}_{\left\{d_i \le -\frac{\log n}{\log\log n}\right\}}\right] \ge n^{1-\frac{(\sqrt{a}-\sqrt{b})^2}{2}+o(1)} \to \infty,$$

where the last inequality holds due to $\sqrt{a}-\sqrt{b} < \sqrt{2}$. In other words, the expected number of i's such that $d_i \le -\frac{\log n}{\log\log n}$ is large. However, as we discussed in the previous lectures on the number of isolated nodes, the expected number of i's such that $d_i \le -\frac{\log n}{\log\log n}$ is large does not directly imply that with high probability the number of i's such that $d_i \le -\frac{\log n}{\log\log n}$ is large. One way to show this is indeed the case is the celebrated *second moment method*.

The argument presented in (13.1) is another widely used method to show that indeed with high probability, there exists an i such that $d_i \le -\frac{\log n}{\log\log n}$, by assuming d_i's are independent. However, in SBM, d_i's are in fact dependent because A_{ij} appears in both d_i and d_j. Please refer to [ABH16, Section 5] for a technique to deal with the dependency.

Remark 13.2. Recall that in the proof of the achievability of SDP, we need $\min_i d_i$ to be large with high probability. In fact, we used the following concentration inequality:

$$\mathbb{P}\left[d_i \le \frac{\log n}{\log\log n}\right] \le n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{2}+o(1)},$$

The above inequality together with Proposition 13.1, yield that

$$n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{2}+o(1)} \le \mathbb{P}\left[d_i \le -\frac{\log n}{\log\log n}\right] \le \mathbb{P}\left[d_i \le \frac{\log n}{\log\log n}\right] \le n^{-\frac{(\sqrt{a}-\sqrt{b})^2}{2}+o(1)}.$$

13.3 Information-theoretic lower bounds for weak recovery

In this section, we will derive the information-theoretic lower bounds for weak recovery under binary SBM. Notice that weak recovery requires a vanishing fraction of misclassified nodes, i.e.,

$$\frac{1}{n}\min\{d_H(\widehat{x}, x^*), d_H(\widehat{x}, -x^*)\} \to 0 \quad \text{in probability.}$$

For exact recovery, we reduce the task of proving information-theoretic lower bounds to proving ML estimator fails. However, for weak recovery, ML estimator is not necessarily optimal in minimizing the fraction of misclassified nodes. Hence, we cannot directly reduces the task of proving information-theoretic lower bounds for weak recovery to proving ML estimator fails. Instead, we will use genie-aided argument. More specifically, we assume that there is a genie which provides some extra side information and try to reduce the original more complicated estimation problem to a simpler one.

Theorem 13.3. *Assume $p > q$ in the binary symmetric SBM $\mathcal{G}(n, p, q)$. If weak recovery is possible, then $\frac{n(p-q)}{\sqrt{np}} \to \infty$.*

Note: Recall that we have shown that SDP achieves weak recovery if $\frac{n(p-q)}{\sqrt{np}} \to \infty$. The achievability result of SDP together with impossibility result in Theorem 13.3 imply that the information-theoretic limit for weak recovery is given by $\frac{n(p-q)}{\sqrt{np}} \to \infty$ and it can be achieved efficiently in polynomial-time.

Proof. Fix any estimator $\widehat{x}(A)$ which attains weak recovery. Then as $n \to \infty$, in probability,

$$\frac{1}{n} \min\{d_H(\widehat{x}, x^*), d_H(\widehat{x}, -x^*)\} \to 0.$$

Since $\frac{1}{n} \min\{d_H(\widehat{x}, x^*), d_H(\widehat{x}, -x^*)\}$ is a bounded random variable, it follows that as $n \to \infty$

$$\mathbb{E}\left[\frac{1}{n^2}\langle \widehat{x}, x^*\rangle^2\right] = \frac{1}{n^2}\mathbb{E}\big[\sum_{ij} \widehat{x}_i \widehat{x}_j x_i^* x_j^*\big] \to 1.$$

Intuitively, $\frac{1}{n^2}\langle \widehat{x}, x^*\rangle^2$ measures the overlap between $\widehat{x}$ and x^*. Since

$$\frac{1}{n^2}\mathbb{E}\big[\sum_{ij} \widehat{x}_i \widehat{x}_j x_i^* x_j^*\big] = \frac{n}{n^2} + \frac{n(n-1)}{n^2}\mathbb{E}[\widehat{x}_1 \widehat{x}_2 x_1^* x_2^*],$$

it further follows that $\mathbb{E}[\widehat{x}_1 \widehat{x}_2 x_1^* x_2^*] \to 1$, i.e., $\mathbb{P}\left[\widehat{x}_1 \widehat{x}_2 = x_1^* x_2^*\right] \to 1$. In other words, $\widehat{x}$ tells whether node 1 and node 2 are in the same cluster correctly with probability converging to 1.

Next, we apply the genie-aided argument. For ease of exposition, we will assume that $x_i^* \overset{\text{i.i.d.}}{\sim} \text{Unif}\{+1, -1\}$ (in this case, two clusters may not have equal sizes). Let $x_{\setminus 1}^*$ denote the true cluster label vector all nodes except node 1. Suppose there is a genie who reveals $x_{\setminus 1}^*$ Let $\mathcal{X}(A, x_{\setminus 1}^*)$ denote the set of all possible estimators of x^* which have access to both adjacency matrix and side information $x_{\setminus 1}^*$. Let $\mathcal{X}_1(A, x_{\setminus 1}^*)$ denote the set of all possible estimators of x_1^* which have access to both adjacency matrix and side information $x_{\setminus 1}^*$. Then we have

$$\begin{aligned}
\mathbb{E}[\widehat{x}_1 \widehat{x}_2 x_1^* x_2^*] &\overset{(a)}{\leq} \max_{\widetilde{x} \in \mathcal{X}(A, x_{\setminus 1}^*)} \mathbb{E}[\widetilde{x}_1 \widetilde{x}_2 x_1^* x_2^*] \\
&\overset{(b)}{=} \max_{z \in \mathcal{X}_1(A, x_{\setminus 1}^*)} \mathbb{E}[z x_1^*] \\
&= \max_{z \in \mathcal{X}_1(A, x_{\setminus 1}^*)} \mathbb{P}\left[z = x_1^*\right] - \mathbb{P}\left[z \neq x_1^*\right] \\
&= 2\mathbb{P}\left[z_{\text{ML}} = x_1^*\right] - 1,
\end{aligned}$$

where (a) holds because $\widehat{x}(A) \in \mathcal{X}(A, x_{\setminus 1}^*)$; (b) holds because for any $\widetilde{x} \in \mathcal{X}(A, x_{\setminus 1}^*)$, one can define $z \in \mathcal{X}_1(A, x_{\setminus 1}^*)$ such that $z = \widetilde{x}_1$ if $\widetilde{x}_2 = x_2^*$ and $z = -\widetilde{x}_1$ if $\widetilde{x}_2 = -x_2^*$. Then $\widetilde{x}_1 \widetilde{x}_2 x_1^* x_2^* = z x_1^*$; the last equality holds by defining

$$z_{\text{ML}} \in \arg\max_{x \in \{\pm 1\}} \mathbb{P}\left[A, x_{\setminus 1}^* \mid x_1^* = x\right].$$

Combining the last two displayed equations yield that $\mathbb{P}\left[z_{\text{ML}} = x_1^*\right] \to 1$. Notice that

$$\mathbb{P}\left[A, x_{\setminus 1}^* \mid x_1^* = x\right] = \mathbb{P}\left[x_{\setminus 1}^* \mid x_1^* = x\right] \mathbb{P}\left[A \mid x_{\setminus 1}^*, x_1^* = x\right] = 2^{-(n-1)} \mathbb{P}\left[A \mid x_{\setminus 1}^*, x_1^* = x\right],$$

where the last equality holds because $x_i^* \overset{\text{i.i.d.}}{\sim} \text{Unif}\{+1,-1\}$. It follows that

$$z_{\text{ML}} \in \arg\max_{x\in\{1,-1\}} \mathbb{P}\left[A \mid x^*_{\setminus 1}, x_1^* = x\right],$$

i.e.,

$$z_{\text{ML}} = \begin{cases} +1 & \mathbb{P}\left[A \mid x^*_{\setminus 1}, x_1^* = 1\right] \geq \mathbb{P}\left[A \mid x^*_{\setminus 1}, x_1^* = -1\right] \\ -1 & \text{o.w..} \end{cases}$$

Moreover,

$$\log \mathbb{P}\left[A \mid x^*_{\setminus 1}, x_1^* = +1\right] - \log \mathbb{P}\left[A \mid x^*_{\setminus 1}, x_1^* = -1\right] = \log \frac{p(1-q)}{q(1-p)} \left(\sum_{j\neq 1} A_{1j} x_j^*\right).$$

Hence,

$$z_{\text{ML}} = \begin{cases} +1 & \sum_{j\neq 1} A_{1j} x_j^* \geq 0 \\ -1 & \text{o.w..} \end{cases}$$

Therefore,

$$\mathbb{P}\left[z_{\text{ML}} = x_1^*\right] = \mathbb{P}\left[\sum_{j\neq 1} A_{ij} x_j^* x_1^* \geq 0\right] = \mathbb{P}\left[d_1 \geq 0\right],$$

where in the last equality we used $d_1 = \sum_{j\neq 1} A_{1j} x_j^* x_1^*$. Since we have shown that $\mathbb{P}\left[z_{\text{ML}} = x_1^*\right] \to 1$, it follows that $\mathbb{P}\left[d_1 \geq 0\right] \to 1$.

Note that $\{A_{1j} x_1^* x_j^*\}_{j\neq 1} \overset{\text{i.i.d.}}{\sim} \frac{p}{2}\delta_1 + \frac{q}{2}\delta_{-1} + \left(1 - \frac{p+q}{2}\right)\delta_0$. By central limit theorem, $\frac{1}{\sqrt{n-1}} d_1$ converges in distribution to $\mathcal{N}(\mu, \sigma^2)$, where $\mu = \sqrt{n-1}(p-q)/2$ and $\sigma^2 = (p+q)/2 - (p-q)^2/4$. Hence, $\mathbb{P}\left[d_1 \geq 0\right] \to 1$ implies that $\mathbb{P}\left[\mathcal{N}(\mu, \sigma^2) \geq 0\right] \to 1$, which further implies that $\frac{\mu}{\sigma} \to \infty$, i.e., $\frac{n(p-q)}{\sqrt{np}} \to \infty$. In Homework 5, we will rigorously justify the Gaussian approximation by applying Berry-Esseen's theroem. □

Chapter 14

Information Theoretic Methods

Outline

- f-divergence
- Data processing inequality [DPI] for f-divergence
- Mutual information bound (Fano's inequality)
- Examples

14.1 f-divergence

Recall the stochastic block model with true cluster label vector X^* and observed adjacency matrix A. To recover cluster structure X^* from A, we are essentially interested in distinguishing between the conditional distribution $P_{A|X^*=x}$ and conditional distribution $P_{A|X^*=x'}$ for different possible cluster partitions $x \neq x'$. To this end, we need to characterize the distance between two different probability distributions.

Definition 14.1 (f-divergence)**.** Let P and Q be two probability distributions defined on a common space. Then for any convex function $f : (0, \infty) \to \mathcal{R}$ such that f is strictly convex at 1 and $f(1) = 0$, the f-divergence of Q from P with $P \ll Q$ (P absolutely continuous with respect to Q) is defined as:

$$D_f(P||Q) \triangleq E_Q\left[f\left(\frac{dP}{dQ}\right)\right] \tag{14.1}$$

Note:

- We say that $P \ll Q$ if for any measurable set E, $Q(E) = 0 \Rightarrow P(E) = 0$. If $P \ll Q$ then we can define the Radon-Nikodym derivative $\frac{dP}{dQ}$: $P(E) = \int_E \frac{dP}{dQ} dQ$.
- We say f is strictly convex at 1, if $\forall x, y \in (0,1) \cup (1,\infty)$ and $\forall \lambda \in (0,1)$ such that $\lambda x + (1-\lambda)y = 1$,
$$\lambda f(x) + (1-\lambda) f(y) > f(1)$$
- An important property that we will use later is the so-called **change of measure property**:
$$E_{X\sim P}[g(x)] = E_{X\sim Q}\left[g(x)\frac{dP}{dQ}\right] = \int g(x)\frac{dP}{dQ} dQ.$$
- We will often write $\frac{dP}{dQ}$ simply as $\frac{P}{Q}$.

14.1.1 Examples of f-divergence: THE BIG FOUR

We will look at four most frequently used f-divergences.

1. **KL-divergence**: $f(x) = x \cdot \log(x)$
$$\begin{aligned} D(P||Q) &\triangleq E_P\left[\log\left(\frac{dP}{dQ}\right)\right] \\ &= E_Q\left[\left(\frac{dP}{dQ}\log\frac{dP}{dQ}\right)\right]. \end{aligned}$$
Note that KL-divergence is not symmetric, i.e., $D(P\|Q) \neq D(Q\|P)$.

2. **TV-divergence:** $f(x) = \frac{1}{2}|x-1|$
$$\begin{aligned} \mathrm{TV}(P,Q) &\triangleq \frac{1}{2}\int |P - Q| \\ &= \frac{1}{2} E_Q\left[\left|\frac{P}{Q} - 1\right|\right] \\ &= \sup_E \left(P(E) - Q(E)\right). \end{aligned}$$
Total variation divergence is symmetric, i.e., $\mathrm{TV}(P,Q) = \mathrm{TV}(Q,P)$.

3. χ^2**-divergence** $f(x) = (x-1)^2$
$$\begin{aligned} \chi^2(P||Q) &\triangleq E_Q\left[\left(\frac{P}{Q} - 1\right)^2\right] \\ &= \int \left(\frac{P}{Q} - 1\right)^2 Q \\ &= \int \left(\frac{P^2}{Q^2} - 2\frac{P}{Q} + 1\right) Q \\ &= \int \frac{P^2}{Q} - 1. \end{aligned}$$
χ^2-divergence is not symmetric.

4. **Squared Hellinger divergence** $f(x) = (\sqrt{x} - 1)^2$

$$H^2(P, Q) \triangleq E_Q\left[\left(\sqrt{\frac{P}{Q}} - 1\right)^2\right]$$
$$= \int \left(\sqrt{P} - \sqrt{Q}\right)^2.$$

Squared Hellinger divergence is symmetric.

The above four divergences appear frequently in the derivation of decision boundary for Hypothesis testing problem and information limits for estimation problem. In fact, the four divergences can be related to each other in various ways. See [Wu16, Chapter 5] for more details.

Theorem 14.1 (Properties of f-divergence)**.** *The following holds in general for f-divergence.*

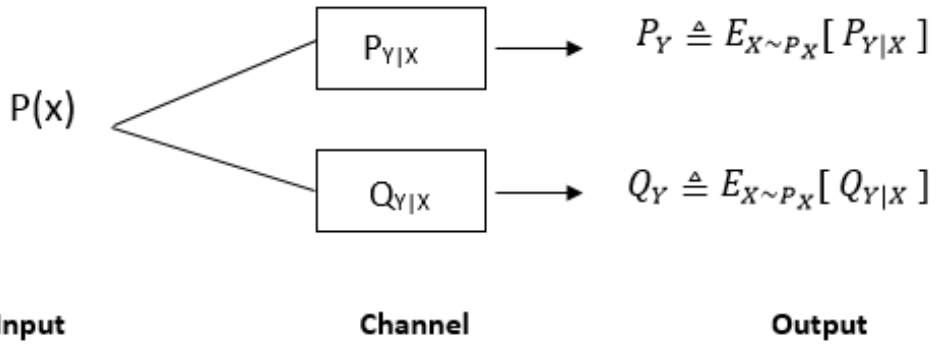
1. *Non-Negativity:* $D_f(P \parallel Q) \geq 0$ *with equality iff* $P = Q$.

   Proof. Use the convexity of f and Jensen's inequality. □

2. *Joint Convexity:* $(P, Q) \to D_f(P \parallel Q)$ *is jointly convex.*

   Proof. Check using the definition of convexity that the mapping $(p, q) \longleftrightarrow qf(\frac{p}{q})$ is jointly convex. Then $(P, Q) \to D_f(P \parallel Q)$ is a linear combination of convex functions and therefore convex. □

3. *Conditioning increases f-divergence : Define*



$$D_f(P_{Y|X} \parallel Q_{Y|X}|P_X) \triangleq E_{X\sim P_X}\left[D_f(P_{Y|X} \parallel Q_{Y|X})\right].$$

Then we have

$$D_f(P_Y \parallel Q_Y) \leq D_f(P_{Y|X} \parallel Q_{Y|X}|P_X) \tag{14.2}$$

Proof. Use Jensen inequality and the fact that $D_f(\cdot,\cdot)$ is convex to get that

$$\begin{aligned} D_f(P_Y \parallel Q_Y) &= D_f(E_{X\sim P_X}[P_{Y|X}] \parallel E_{X\sim P_X}[Q_{Y|X}]) \\ &\leq E_{X\sim P_X}\left[D_f(P_{Y|X} \parallel Q_{Y|X})\right] \\ &= D_f(P_{Y|X} \parallel Q_{Y|X}|P_X) \end{aligned}$$

□

14.2 Data Processing Inequality

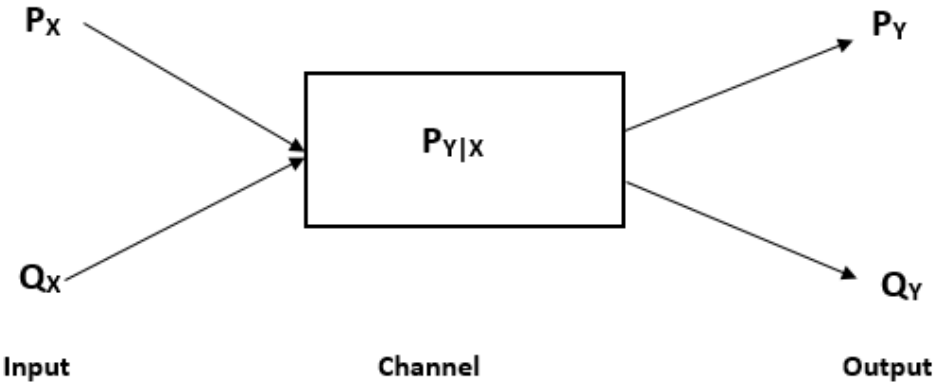


Figure 14.1: Data Processing

Theorem 14.2 (Data processing inequality)**.** *Suppose* $P_Y = \mathbb{E}_{X\sim P_X}[P_{Y|X}]$ *and* $Q_Y = \mathbb{E}_{X\sim Q_X}[P_{Y|X}]$. *Then*

$$D_f(P_X \parallel Q_X) \geq D_f(P_Y \parallel Q_Y) \tag{14.3}$$

Note: To get the intuition behind the data processing inequality, consider the following hypothesis testing problem, where given the observation X, we would like to determine whether the data is generated by Q_X or P_X.

$$\begin{aligned} H_0 &: X \sim Q_X \\ H_1 &: X \sim P_X. \end{aligned}$$

The data processing inequality in (14.3) implies that processing X makes it harder to distinguish between the two hypotheses.

Proof of Theorem 14.2. Starting from the left hand of inequality (14.3), we have

$$
\begin{aligned}
D_f(P_X \parallel Q_X) &= E_{X\sim Q_X}\left[f\left(\frac{P_X}{Q_X}\right)\right]. \\
&\overset{(a)}{=} E_{Q_{XY}}\left[f\left(\frac{P_X \cdot P_{Y|X}}{Q_X \cdot P_{Y|X}}\right)\right] \\
&= E_{Q_Y}\left[E_{Q_{X|Y}}\left[f\left(\frac{P_{XY}}{Q_{XY}}\right)\right]\right] \\
&\overset{(b)}{\geq} E_{Q_Y}\left[f\left(E_{Q_{X|Y}}\left[\frac{P_{XY}}{Q_{XY}}\right]\right)\right] \\
&\overset{(c)}{=} E_{Q_Y}\left[f\left(\frac{P_Y}{Q_Y}\right)\right] \\
&= D_f(P_Y \parallel Q_Y),
\end{aligned}
$$

where (a) follows because $\frac{P_X \cdot P_{Y|X}}{Q_X \cdot P_{Y|X}} = \frac{P_X}{Q_X}$; (b) holds due to the convexity of f and Jensen's inequality; (c) holds because

$$
\begin{aligned}
E_{Q_{X|Y}}\left[\frac{P_{XY}}{Q_{XY}}\right] &= E_{Q_{X|Y}}\left[\frac{P_{XY}}{Q_Y \cdot Q_{X|Y}}\right]. \\
&= \int_X Q_{X|Y}\frac{P_{XY}}{Q_Y \cdot Q_{X|Y}} \\
&= \frac{P_Y}{Q_Y}.
\end{aligned}
$$

□

Example 14.1. $P_{Y|X}$ is deterministic with $Y = h(X)$ and $h(X) = 1_{\{X\in E\}}$

$$
D_f(P_X \parallel Q_X) \geq D_f\left(\mathrm{Bern}(P_X(E)) \parallel \mathrm{Bern}(Q_X(E))\right).
$$

Example 14.2. $X = (X_1, X_2)$ and $P_{Y|X}$ is deterministic with $Y = h(X)$ with $h(X) = X_1$

$$
D_f(P_{X_1,X_2} \parallel Q_{X_1,X_2}) \geq D_f(P_{X_1} \parallel Q_{X_1}). \tag{14.4}
$$

14.3 Mutual Information Bound

An important quantity to measure the dependency between two random variables is the mutual information.

Definition 14.2 (Mutual Information)**.** Mutual Information is defined as the KL-divergence between the joint distribution P_{XY} and product of marginal distributions $P_X P_Y$. More formally,

$$
I(X;Y) \triangleq D(P_{XY} \| P_X P_Y) \tag{14.5}
$$

Note: By convention, we use semicolon ; to separate two random variables X and Y in $I(X;Y)$, emphasizing that $I(X;Y)$ is not a function of random variables X and Y themselves.

14.3.1 Properties of Mutual Information

- $I(X;Y) = D(P_{Y|X}\|P_Y|P_X) = \mathbb{E}_{X\sim P_X}[D(P_{Y|X}\|P_Y)]$
- Symmetry: $I(X;Y) = I(Y;X)$
- Measure of dependency: $I(X;Y) \geq 0$ with equality iff $X \perp Y$
- If X, Y are discrete random variables. Then,

$$I(X;Y) = H(Y) - H(Y|X) \tag{14.6}$$

where $H(Y)$ is called entropy and $H(Y|X)$ is called conditional entropy defined as

$$H(Y) \triangleq \sum_y P_Y(y) \log \frac{1}{P_Y(y)} \tag{14.7}$$

$$H(Y|X) \triangleq \sum_x P_X(x) \sum_y P_{Y|X=x}(y) \log \frac{1}{P_{Y|X=x}(y)} \tag{14.8}$$

Example 14.3 (Entropy for Bernoulli random variable)**.** Let $X \sim \text{Bern}(p)$, then

$$\begin{aligned} H(X) &= p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \\ &\triangleq h(p) \end{aligned} \tag{14.9}$$

where $h(p)$ is known as "binary entropy function". Note that, $h(p)$ attains maximum of $\log 2$ at $p = \frac{1}{2}$. Intuitively speaking, entropy is a measure of randomness. X contains no randomness for $p = 0$ or $p = 1$ as it is completely determined and it contains maximum randomness for $p = \frac{1}{2}$.

14.3.2 Properties of Entropy

- $H(X) \geq 0$
- $H(X) \geq H(Y|X)$ since $I(X,Y) \geq 0$

Example 14.4 (Binary symmetric channel)**.** Let $Y = X \oplus Z$ where $X \sim Bern(\delta)$ and $Z \sim Bern(\epsilon)$ such that $X \perp Z$. We are interested in computing $I(X,Y)$. As a simple case, $I(X,Y) = 0$ when $\epsilon = \frac{1}{2}$ because in this case X and Y become independent. For general case,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(X \oplus Z|X) \\ &= H(Y) - H(X \oplus Z \oplus X|X) \\ &= H(Y) - H(Z|X) \\ &= H(Y) - H(Z) \\ &= h(\delta(1-\epsilon) + (1-\delta)\epsilon) - h(\epsilon), \end{aligned}$$

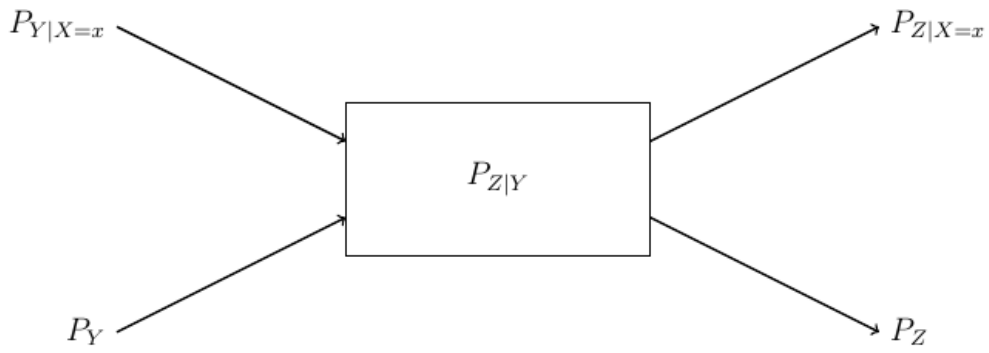
where the third equality holds because conditional on X, there is a one-to-one mapping from $X \oplus Z$ to $X \oplus Z \oplus X$; the fifth equality holds because $X \perp Z$; the last equality holds by using (14.9).

Theorem 14.3 (Data processing inequality for mutual information)**.** *Let* $X \xrightarrow{P_{Y|X}} Y \xrightarrow{P_{Z|Y}} Z$ *denote a Markov chain, i.e.* $Z \perp X|Y$*. Then,* $I(X;Z) \leq I(X;Y)$*.*

Proof. Note that,

$$I(X;Y) = \mathbb{E}_{X\sim P_X}[D(P_{Y|X}\|P_Y)] \tag{14.10}$$
$$I(X;Z) = \mathbb{E}_{X\sim P_X}[D(P_{Z|X}\|P_Z)] \tag{14.11}$$

Consider a channel such as:



Using data processing inequality, we know that

$$\begin{aligned}
&D(P_{Y|X=x}\|P_Y) \geq D(P_{Z|X=x}\|P_Z) \\
&\Rightarrow \mathbb{E}_{X\sim P_X}[D(P_{Y|X}\|P_Y)] \geq \mathbb{E}_{X\sim P_X}[D(P_{Z|X}\|P_Z)] \\
&\Rightarrow I(X;Y) \geq I(X;Z)
\end{aligned}$$

□

14.4 Mutual Information Bound

Consider a general statistical estimation problem, where

$$\theta \in \Theta \xrightarrow{P_{X|\theta}} X \to \widehat{\theta}:$$

θ is the underlying true parameter generated from the parameter space Θ with a prior distribution π, X is the data generated according to $P_{X|\theta}$, and $\widehat{\theta}$ is estimated parameter based on the observation X. Let $l(\theta, \widehat{\theta})$ be a loss function between θ and $\widehat{\theta}$. We are interested in designing optimal estimator which achieves the minimum expected loss $\mathbb{E}\left[\ell(\theta, \widehat{\theta})\right]$.

The data processing inequality for mutual information yields a lower bound on the mutual information between θ and X to attain a certain expected loss: letting $R^* = \mathbb{E}[l(\theta, \widehat{\theta})]$,

$$\inf_{P_{\widetilde{\theta}|\theta}:\mathbb{E}[l(\theta,\widetilde{\theta})]\leq R^*} I(\theta;\widetilde{\theta}) \leq I(\theta;\widehat{\theta}) \leq I(\theta;X) \tag{14.12}$$

Note that $\inf_{P_{\widetilde{\theta}|\theta}:\mathbb{E}[l(\theta,\widetilde{\theta})]\leq R^*} I(\theta;\widetilde{\theta})$ is minimum amount of mutual information needed to attain an expected loss of R^* and $I(\theta;X)$ is amount of information contained in data X about θ.

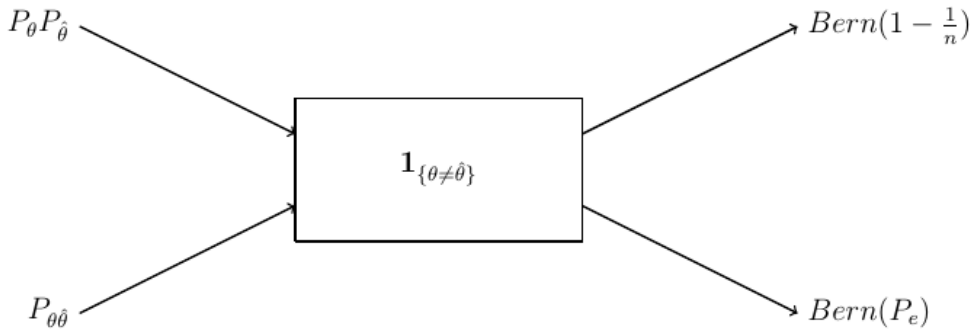
A well-known special case is the so-called Fano's inequality.

14.4.1 Fano's inequality

Theorem 14.4 (Fano's inequality)**.** *Assume $\theta \in \text{Unif}\{\theta_1, \theta_2 \ldots, \theta_n\}$. For any estimator $\widehat{\theta}(X)$, let $P_e \triangleq Pr\{\widehat{\theta} \neq \theta\}$, then*

$$P_e \geq 1 - \frac{\log 2 + I(\theta; X)}{\log n} \tag{14.13}$$

Proof. Consider a channel as shown below,



Note that both output distributions will follow Bernoulli distribution with parameter $Pr\{\theta \neq \widehat{\theta}\}$. However, for first distribution θ and $\widehat{\theta}$ are independent and θ is drawn uniformly from $\{\theta_1, \theta_2, \ldots, \theta_n\}$. Hence, $Pr\{\theta \neq \widehat{\theta}\} = 1 - \frac{1}{n}$. For second distribution, θ and $\widehat{\theta}$ follow the joint distribution $P_{\theta\widehat{\theta}}$ and hence $Pr\{\theta \neq \widehat{\theta}\} = P_e$. Using mutual information bound,

$$\begin{aligned} I(\theta, X) \geq I(\theta, \widehat{\theta}) &= D(P_{\theta\widehat{\theta}} \| P_\theta . P_{\widehat{\theta}}) \\ &\geq D(Bern(1 - \frac{1}{n}) \| Bern(P_e)) \\ &\geq -\log 2 + \log n - P_e \log n. \end{aligned}$$

After rearranging the terms,

$$P_e \geq 1 - \frac{\log 2 + I(\theta, X)}{\log n}.$$

□

To use the mutual information bound, a key challenge is to upper bound the mutual information $I(\theta; X)$, especially when θ or X are high-dimensional.

Theorem 14.5 (Tensorization of MI)**.** *The mutual information can be decomposed as*

$$I(X; Y) = I(X_1, \ldots, X_k; Y) = I(X_1; Y) + I(X_2; Y|X_1) + \cdots + I(X_k; Y|X_1, \ldots, X_{k-1}). \tag{14.14}$$

Moreover, when $P_{Y|X} = \prod_{i=1}^k P_{Y_i|X_i}$, then

$$I(X; Y) \leq \sum_{i=1}^n I(X_i; Y_i) \tag{14.15}$$

with equality if and only if X_i's are independent.

Proof. The proof of (14.14) is an easy application of the telescoping sum:

$$\log \frac{P_{Y|X_1,\dots,X_k}}{P_Y} = \log \frac{P_{Y|X_1}}{P_Y} + \log \frac{P_{Y|X_1,X_2}}{P_{Y|X_1}} + \cdots + \log \frac{P_{Y|X_1,\dots,X_k}}{P_{Y|X_1,\dots,X_{k-1}}}.$$

To show (14.15), note that if $P_{Y|X} = \prod_{i=1}^k P_{Y_i|X_i}$, then

$$\log \frac{P_{Y|X}}{P_Y} = \sum_{i=1}^k \log \frac{P_{Y_i|X_i}}{P_{Y_i}} + \log \frac{\prod_{i=1}^k P_{Y_i}}{P_Y}.$$

Taking expectation over both hand sides of the last displayed equations yields that

$$I(X;Y) = \mathbb{E}_{X,Y}\left[\log \frac{P_{Y|X}}{P_Y}\right] = \sum_{i=1}^k I(X_i;Y_i) - D(P_Y \| \prod_{i=1}^k P_{Y_i}) \le \sum_{i=1}^k I(X_i;Y_i),$$

where the equality holds if and only Y_i's are independent, or equivalently, X_i's are independent. □

Note: When $P_{Y|X} = \prod_{i=1}^n P_{Y_i|X_i}$, in view of (14.15), we can derive an upper bound on $I(X;Y)$ by computing marginal mutual information $I(X_i;Y_i)$.

Another way to upper bound $I(X;Y)$ is through the so-called *Golden formula*.

Theorem 14.6 (Golden formula of MI)**.**

$$I(X;Y) = \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X), \tag{14.16}$$

where the minimum is achieved when $Q_Y = P_Y$.

Proof. Note that

$$I(X;Y) = D(P_{Y|X} \| P_Y | P_X) = \mathbb{E}_{P_{X,Y}}\left\{\log\left[\frac{P_{Y|X}}{Q_Y}\frac{Q_Y}{P_Y}\right]\right\} = D(P_{Y|X} \| Q_Y | P_X) - D(P_Y \| Q_Y).$$

□

Note: In view of (14.16), by properly choosing Q_Y, we can always get an upper bound $I(X;Y) \le D(P_{Y|X} \| Q_Y | P_X)$.

14.5 Two examples

In this section, we illustrate the power of mutual information bound through two simple examples.

14.5.1 Exact recovery under SBM with multiple communities

Consider SBM with n nodes with r clusters, where two nodes in the same clusters are connected independently with edge probability p, and two nodes in two different clusters are connected independently with edge probability q. Let $x_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\{1,2,\dots,r\}$ be the cluster label of node i. Let $\bar{p} = \frac{1}{r}p + \frac{r-1}{r}q$ denote the average edge probability. Let A denote the observed adjacency matrix. Let $d(p\|q) \triangleq D(\text{Bern}(p)\|\text{Bern}(q))$ denote the binary KL-divergence.

Theorem 14.7. *If there exists an estimator $\widehat{x}(A)$ such that $\mathbb{P}[\widehat{x} \neq x] \geq \epsilon$ for some $\epsilon > 0$, then*

$$n\left(\frac{1}{r}d(p\|\bar{p}) + (1-\frac{1}{r})d(q\|\bar{p})\right) \geq 2(1-\epsilon)\log r - \frac{1}{n}2\log 2, \tag{14.17}$$

and

$$n\frac{1}{r}d(p\|q) \geq 2(1-\epsilon)\log r - \frac{1}{n}2\log 2, \tag{14.18}$$

then for any we have

Note:

- The necessary conditions in the theorem above are not sufficient when $r = \Theta(1)$. In particular, when $r = 2$ and $p = a\log n/n$ and $q = b\log n/n$,

  $$d(p\|\bar{p}) \asymp d(q\|\bar{p}) \asymp d(p\|q) \asymp \frac{\log n}{n}.$$

  Hence both the necessary conditions (14.17) and (14.18) are satisfied. However, we know that in this case the sharp information limit for exact recovery is given by $\sqrt{a} - \sqrt{b} > \sqrt{2}$.

- If $\frac{p}{q} = \Theta(1)$ and p is bounded away from 1, then $d(p\|q) \asymp \frac{(p-q)^2}{q}$. Hence the necessary condition (14.18) implies that $K(p-q)^2 \gtrsim q\log\frac{n}{K}$. This necessary condition turns out to be tight up to a constant factor if $\log\frac{n}{K} \asymp \log n$. See [CX14] for more details.

Proof. The proof is an easy application of Fano's inequality. Notice that x is uniformly chosen from all r^n possible cluster label vector. Hence,

$$\epsilon \geq P_e \geq 1 - \frac{\log 2 + I(x;A)}{\log(r^n)}.$$

It follows that

$$I(x;A) \geq (1-\epsilon)n\log r - \log 2.$$

To finish the proof, we need to upper bound $I(x;A)$. Notice that $P_{A|x} = \prod_{i<j} P_{A_{ij}|x_i,x_j}$. Hence, in view of (14.15), we get that

$$I(x;A) = \sum_{i<j} I(x_i,x_j;A_{ij}) = \binom{n}{2}\left(\frac{1}{r}d(p\|\bar{p}) + (1-\frac{1}{r})d(q\|\bar{p})\right)$$

which completes the proof of (14.17). Moreover, in view of (14.16),

$$\begin{aligned} I(x;A) &= \min_{Q_A} D(P_{A|x}\|Q_A|P_x) \\ &\leq D(P_{A|x}\|\mathrm{Bern}(q)^{\otimes\binom{n}{2}}|P_x) \\ &\overset{(a)}{=} \sum_{i<j} D(P_{A_{ij}|x_i,x_j}\|\mathrm{Bern}(q)|P_{x_i,x_j}) \\ &= \binom{n}{2}\frac{1}{r}d(p\|q), \end{aligned}$$

where $\mathrm{Bern}(q)^{\otimes\binom{n}{2}}$ denote the $\binom{n}{2}$ product distribution of $\mathrm{Bern}(q)$; and (a) follows because $P_{A|x} = \prod_{i<j} P_{A_{ij}|x_i,x_j}$ is also a product distribution. The proof of (14.18) is complete. □

14.5.2 Weak recovery under SBM with a single community

The mutual information bound can be also used to derive necessary conditions for weak recovery.

Consider the SBM with a single community, where K out of n nodes are in the community; two nodes are connected independently with edge probability p if both of them are in the community, and with edge probability q otherwise. Let $x_i = 1$ if node i is in the community, and $x_i = 0$ otherwise. Let A denote the observed adjacency matrix.

We are interested in weak recovery of x from A. In particular, for an estimator $\widehat{x}(A)$, we say $\widehat{x}$ achieves weak recovery, if $\mathbb{E}\left[d_H(x, \widehat{x})\right] = o(K)$, i.e., the expected number of misclassified nodes is $o(K)$. Notice that if $\widehat{x}$ is the all-zero vector, then $d_H(x, \widehat{x}) = K$.

Theorem 14.8. *If weak recovery is possible, then*

$$\liminf_{n\to\infty} \frac{(K-1)d(p\|q)}{\log \frac{n}{K}} \geq 2.$$

By assumption, there exists an estimator $\widehat{x}$ such that $\mathbb{E}\left[d_H(x, \widehat{x})\right] \leq \epsilon_n K$ with $\epsilon_n = o(1)$. Note that

Proof.

$$I(x; A) \geq I(x; \widehat{x}) \geq \min_{\widetilde{x}: \mathbb{E}[d_H(\widetilde{x}; x)] \leq \epsilon_n K} I(x; \widetilde{x}) \geq (1 + o(1)) \log \frac{n}{K},$$

where the detailed justification for the last inequality can be found in [HWX15]. Moreover, in view of (14.16),

$$I(x; A) \leq D(P_{A|x} \| \mathrm{Bern}(q)^{\otimes \binom{n}{2}} | P_x) = \binom{K}{2} d(p\|q).$$

The conclusion follows by combining the last two displayed equations. □

Bibliography

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016. arXiv 1405.3267.

[CX14] Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. In *Proceedings of ICML 2014 (Also arXiv:1402.1267)*, Feb 2014.

[Gro53] Alexander Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Resenhas do Instituto de Matemática e Estatistica da Universidade de São Paulo*, 2(4):401–481, 1953.

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

[HWX15] B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. arXiv 1509.07859, September 2015.

[LP68] Joram Lindenstrauss and A Pełczyński. Absolutely summing operators in L_p-spaces and their applications. *Studia Mathematica*, 3(29):275–326, 1968.

[Mac03] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[Mon15] Andrea Montanari. Ee 378b: Inference, estimation, and information processing lecture 3, 2015.

[Wu16] Yihong Wu. Lecture notes for ECE598YW: Information-theoretic methods for high-dimensional statistics, 2016.