Graph Spectral Clustering Quiz

Last Name	First Name	SID

Rules:

- **Unless otherwise stated, all your answers need to be justified and your work must be shown. Answers without sufficient justification will get no credit.**
- You have 30 minutes to complete the quiz.
- Collaboration with others is strictly prohibited.
- You may not reference your notes, the textbook, and any material that can be found through the course website. You may not use Google to search up general knowledge.
- For any clarifications you have, please create a private Piazza post. We will have a Google Doc that shows our official clarifications.

1 True/False [1 point each]

1) The largest eigenvectors of the similarity graph Laplacian matrix should be used to construct the eigenvector matrix that will eventually be used for clustering.

   (a) True

   (b) False

> False. The smallest eigenvectors should be used to construct the eigenvector matrix in the clustering algorithm.

2) When performing k-means in in the graph spectral clustering algorithm, k-means should be performed on the columns of the eigenvector matrix of the Laplacian matrix.

   (a) True

   (b) False

> False. K-means is performed on the rows of the eigenvector matrix.

3) Graph spectral clustering builds on a centroid-based clustering algorithm to achieve more complex decision boundaries between clusters.

   (a) True

   (b) False

> True. The GSC algorithm includes k-means, which is a centroid-based clustering technique.

4) All entries in the similarity matrix used in graph spectral clustering are nonnegative.

   (a) True

   (b) False

> True. The similarity matrix is a symmetric matrix with nonnegative entries because each kernel function outputs nonnegative numbers for all possible inputs.

5) The goal of spectral graph clustering is to separate points into clusters such that intra-cluster (within each cluster) similarity is minimized and inter-cluster (between different clusters) similarity is maximized.

   (a) True

   (b) False

False. Graph spectral clustering maximizes intra-cluster similarity and minimizes inter-cluster similarity.

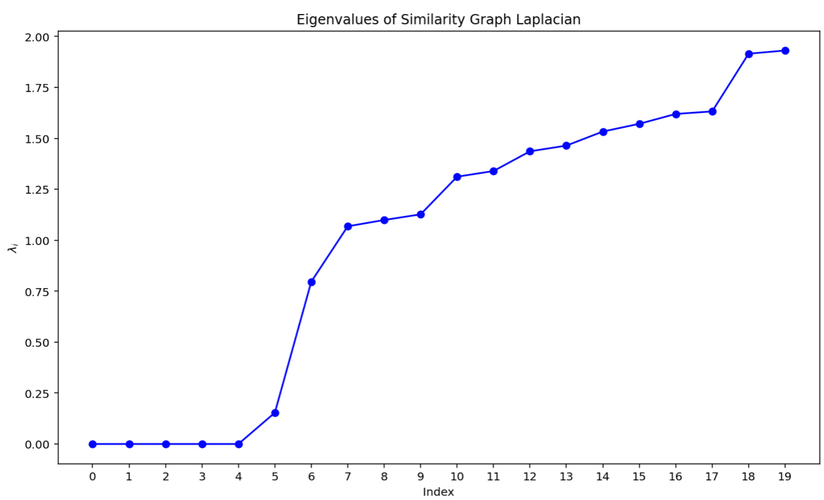
2 Multiple Choice [2 points each]

1) Which matrix has diagonal entries that correspond to the sum of the weights of edges adjacent to each vertex?

(a) Degree matrix
(b) Adjacency matrix
(c) Laplacian matrix
(d) None of the above

(a) Degree matrix. This is the definition of the degree matrix.

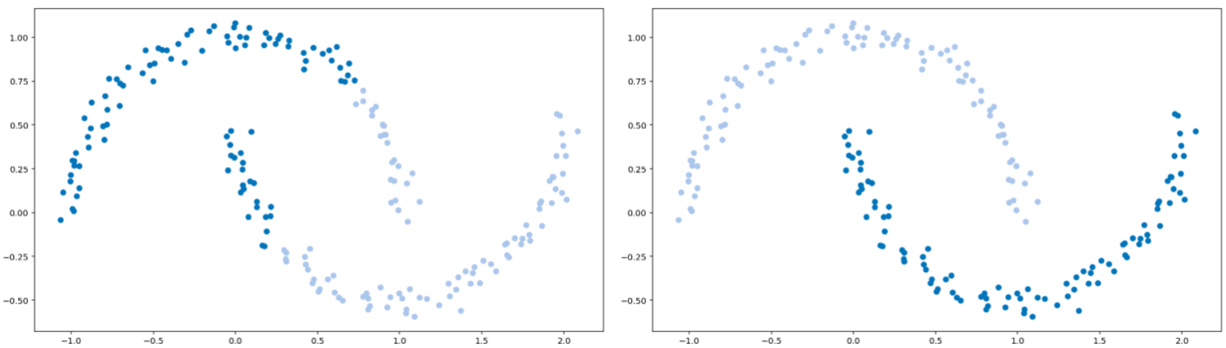
2) Looking at the plot below for eigenvalue growth rate, which of the following values of k would be best to choose?



(a) 3
(b) 5
(c) 7
(d) 9

(b) 5. From the plot, we can see that five of the eigenvalues are nearly zero, which is a good guess for how many clusters there are. Six would also be a good guess for k, if that were a possible answer.

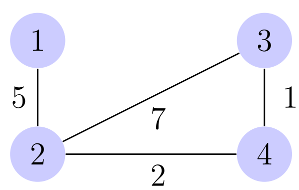
3) For the two plots shown below, one is the result of spectral clustering with a Gaussian similarity function with $\sigma = 0.1$, and another is the result of spectral clustering with a Gaussian similarity function with $\sigma = 1$. Label the plots with the corresponding σ value of the Gaussian similarity function used.



The left graph is $\sigma = 1$, the right graph is $\sigma = 0.1$. As you have seen in the iPython notebook, when σ is too large (such as $\sigma = 1$), spectral clustering resembles K-means. In the sweet spot, when σ is between 0.02 and 0.2, graph spectral clustering separates the two moons correctly.

3 Short Answers [3 points each]

1) Find the Laplacian matrix for this graph:



$$D = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 14 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad W = \begin{bmatrix} 0 & 5 & 0 & 0 \\ 5 & 0 & 7 & 2 \\ 0 & 7 & 0 & 1 \\ 0 & 2 & 1 & 0 \end{bmatrix}$$

$$L := D - W = \begin{bmatrix} 5 & -5 & 0 & 0 \\ -5 & 14 & -7 & -2 \\ 0 & -7 & 8 & -1 \\ 0 & -2 & -1 & 3 \end{bmatrix}$$

2) What is the role of a similarity function in graph spectral clustering? List 2 other examples of valid similarity functions.

The similarity function is used to construct a similarity graph from the data points. It allows us to take two points in Euclidean space and define a relationship between the two points. Two examples of similarity functions are the Gaussian similarity function and exponential similarity function.

3) Explain why it is always possible to compute the first k eigenvectors of the Laplacian matrix in spectral clustering.

The spectral theorem states that symmetric matrices can always be diagonalized/eigendecomposed. Since the Laplacian matrix L is symmetric by construction, any number k up to n (the number of rows of L) eigenvectors are obtainable.

4) What effect does decreasing σ have on the RBF kernel?

The RBF kernel is $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Decreasing σ causes $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ to approach zero for virtually all values.

5) Why do we prefer using k-means++ over k-means in general? Give two reasons.

k-means++ converges faster than k-means and guarantees that the cluster assignment ends up being an $O(\log k)$-approximation of the optimal cluster assignment unlike k-means where the cluster assignment found can be arbitrarily bad.

6) Fill in the blanks to complete the pseudocode of the normalized graph spectral clustering algorithm.

Return clusters $A_1, \cdots, A_k$
Given data $X \in \mathbb{R}^{n \times d}$, number of desired clusters k

(a) Construct a ____________ $S \in \mathbb{R}^{n \times n}$ according to some well defined metric, such as the Gaussian kernel or Euclidean distance.

(b) Compute $L_n =$ ____________, the symmetric normalized graph Laplacian of S.

(c) Compute the first k ____________ of L, where the first k ____________ correspond to ____________.

(d) Construct a matrix H for L with ____________, making sure to normalize the ____________ to have norm 1.

(e) Run ____________ on ____________ to cluster the data.

Return clusters $A_1, \cdots, A_k$
Given data $X \in \mathbb{R}^{n \times d}$, number of desired clusters k

(a) Construct a similarity graph $S \in \mathbb{R}^{n \times n}$ according to some well defined metric, such as the Gaussian kernel or Euclidean distance.

(b) Compute $L_n = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, the symmetric normalized graph Laplacian of S.

(c) Compute the first k eigenvectors of L, where the first k eigenvectors correspond to the k smallest eigenvalues.

(d) Construct a matrix H for L with the k eigenvectors as the columns, making sure to normalize the rows to have norm 1.

(e) Run k-means on the rows of H to cluster the vertices.