# Diffusion Maps and Spectral Clustering

Uri Shaham

March 4, 2024

## 1 Spectral Clustering

### 1.1 Graph Laplacians

Spectral clustering is a clustering algorithm which is far more expressive than $k$-means, and is not limited to convex clusters.

**Definition 1.1** (Undirected weighted graph)**.** *A weighted graph is an undirected graph $G = (V, E)$ where $V = \{v_1, \ldots, v_n\}$ and each to vertices $v_i, v_j$ are connected with an edge with non-negative weight $w_{ij} \geq 0$. We denote the weight matrix as $W$.*

**Remark 1.2.** *When data lies in Euclidean space and the weights are not given, it is a common practice to use Gaussian kernel to compute the weights*

$$w_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2\right)}{2\sigma^2},$$

*for such bandwidth parameter $\sigma$.*

**Remark 1.3.** *A common practice is to connect each point only to its k nearest neighbors, which is known as knn graph. The graph can be easily made symmetric, for example by using $W \leftarrow \frac{1}{2}\left(W + W^T\right)$.*

**Definition 1.4** (Degree and Degree matrix)**.** *The degree $d_i$ of a vertex $v_i$ and the graph degree matrix $D$ are defined as*

- *$d_i$ is the sum of weights on the edges of $v_i$, i.e., $d_i = \sum_j w_{ij}$.*
- *$D$ is a diagonal matrix with elements $D_{ii} = d_i$.*

**Definition 1.5** (Unnormalized graph Laplacian)**.** *The unnormalized graph Laplacian is defined as $L_{un} = D - W$.*

Observe that $L_{\text{un}}$ is symmetric.

**Proposition 1.6.** *For every vector $f \in \mathbb{R}^n$, $f^T L_{un} f = \sum_{i,j} w_{ij} \left(f_i - f_j\right)^2$.*

*Proof.* Exercise. □

It follows that $L_{\text{un}}$ is positive semi-definite

**Proposition 1.7.** *The smallest eigenvalue of of $L_{un}$ is 0, and its corresponding eigenvector is the constant vector $\frac{1}{\sqrt{n}}\mathbb{1}$.*

*Proof.* Let $f = \frac{1}{\sqrt{n}}\mathbb{1}$. Then by proposition 1.6

$$f^T L_{\text{un}} f = 0,$$

hence $f, 0$ are an eigenpair. □

**Proposition 1.8.** *The multiplicity of the zero eigenvalue equals the number $k$ of connected components of $G$, and the corresponding eigenvectors are indicator vectors of the components.*

*Proof.* We know that 0 is an eigenvalue. Let $f$ be a corresponding eigenvector. Then since $f^T L_{\text{un}} f = 0$, we have that $f_i = f_j$ whenever $w_{ij} > 0$. For $k > 1$ connected components, $L_{\text{un}}$ is block diagonal, so its spectrum is the union of the spectra of all blocks. □

**Definition 1.9** (Normalized graph Laplacian)**.** *We define two versions of normalized Laplacians:*

- *The random walk graph Laplacian is $L_{rw} = D^{-1}L_{un} = I - D^{-1}W$.*
- *The Symmetric graph Laplacian is $L_{sym} = D^{-\frac{1}{2}}L_{un}D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.*

**Remark 1.10.** *Observe that $L_{rw}$ and $L_{sym}$ are similar matrices, and hence share the same spectrum. In particular, all eigenvalues of $L_{rw}$ are real.*

**Proposition 1.11.** *$L_{sym}$ is positive semi-definite (and hence also $L_{rw}$)*

*Proof.* Exercise. □

**Proposition 1.12.** *0 is an eigenvalue of $L_{rw}$ (and hence also of $L_{sym}$), associated with the constant eigenvector.*

*Proof.* Exercise. □

**Proposition 1.13.** *The multiplicity of the zero eigenvalue in $L_{rw}$ and $L_{sym}$ equals the number $k$ of connected components of $G$, and the corresponding eigenvectors are indicator vectors of the components.*

*Proof.* Analogous to the proof of proposition 1.8. □

## 1.2 Spectral Clustering

We have seen that eigenvectors corresponding to the zero eigenvalues (to whom we refer from now the first eigenvectors) of the Laplacian indicate the connected components. It therefore makes sense to use these eigenvectors to represent the data and identify the cluster structure.

Spectral clustering works by representing the data using the first $k$ eigenvectors, and running $k$-means on this representation. Unlike $k$-means it can handle non-convex data (see Figure 1. Spectral clustering can interestingly be motivated by a graph cut point of view (see section 5 in Von Luxburg's excellent tutorial).

**Remark 1.14.** *Since the first eigenvector is constant in $L_{un}$ and $L_{rw}$, it is often omitted.*

# Further Reading

A excellent spectral clustering tutorial is `https://link.springer.com/content/pdf/10.1007/s11222-007-9033-z.pdf`.

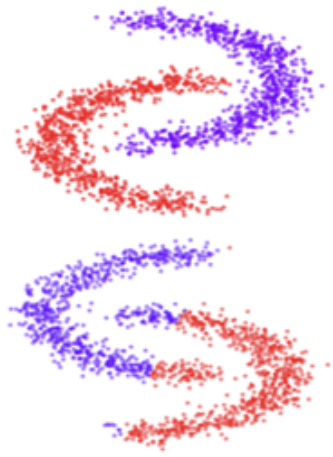

Figure 1: Spectral clustering (top) versus $k$-means on non-convex clusters.

## 2 Diffusion Maps

Diffusion map is a dimensionality reduction technique, that captures geometrical properties of data. In order to do so, it utilizes weighted graphs, which encode local similarity between pairs of points. This local interaction then allows to obtain global representations of the entire data.

Observe that $P := D^{-1}W$ is a Markov matrix, and hence encodes transition probabilities between the vertices. $P$ is a diffusion (averaging) operator, and defines directions of propagation. powers of $P$ corresponds to multi-step walks.

**Proposition 2.1.** *$P$ has stationary distribution $\pi$, with $\pi_i = \frac{d_i}{\sum_j d_j}$, i.e., $\pi^T P = \pi^T$.*

*Proof.* Exercise. □

**Corollary 2.2.** *a stationary distribution $\pi$ is a left eigenvector of $P$, with eigenvalue 1. Since left and right eigenvalues are the same, $P$ as a 1 eigenvalue.*

**Corollary 2.3.** *Let $u$ be an arbitrary vector of length $n$. Each entry of $Pu$ is a convex combination of the entries of $u$. and hence cannot be larger than the maximal one. Therefore any eigenvalue of $P$ can be at most 1.*

Let $1 = \lambda_0 \geq \lambda_1 \geq \ldots$ and $\psi_0, \psi_1, \ldots$ be the eigenvalues and eigenvectors of $P$ (exercise: why $\lambda_0 = 1$?)

**Definition 2.4** (Diffusion distance)**.** *The diffusion distance at time $t$ between vertices $v_i$ and $v_j$ is defined as:*

$$D_t(v_i, v_j) = \|P_i^t - P_j^t\|^2_{\ell^2/d} = \sum_{k=1}^{n} (P_{ik} - P_{jk})^2/d_k,$$

*where $P_i^t$ is the $i$'th row of $P^t$.*

**Remark 2.5.** *The notation $\|x\|^2_{\ell^2/d}$ means that the length is $\sqrt{\sum_i x_i^2/d_i}$.*

Intuitively, if $v_i$ and $v_j$ are connected by a large number of paths, their diffusion probabilities will be similar, and hence their diffusion distance will be small.

The following theorem shows that the diffusion distance squared is in fact the Euclidean distance in the eigenspace of $P$

**Theorem 2.6.**

$$D_t(v_i, v_j) = \left( \sum_{l=1}^{n} \lambda_l^{2t} \left( \psi_{l,i} - \psi_{l,j} \right)^2 \right)^{\frac{1}{2}},$$

*where $\psi_1, \dots \psi_n$ are the eigenvectors of $P$.*

**Proposition 2.7.** *Let $P = D^{-1}W$ and $A = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. Let $(\lambda, \phi)$ be an eigenpair of $A$. Then $(\lambda, D^{-\frac{1}{2}}\phi)$ are an eigenpair of $P$.*

*Proof.* Exercise. □

*Proof.* Consider $A = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. We already know that $P$ has real eigenvalues (since $P$ is similar to $A$, and $A$ is symmetric). Let $A = \sum \lambda_l \phi_l \phi_l^T$. Then

$$\begin{aligned} P &= D^{-\frac{1}{2}} A D^{\frac{1}{2}} \\ &= \sum_l \lambda_l D^{-\frac{1}{2}} \phi_l \phi_l^T D^{\frac{1}{2}} \end{aligned} \tag{1}$$

The vectors $D^{\frac{1}{2}}\phi_l$, $l = 1, \dots ...n$ are orthogonal in $\| \cdot \|_{\ell^2/d}$. This means that we can view the $i$'th row of $P$ as an expansion in that basis with coefficients $\lambda_l D^{-\frac{1}{2}} \phi_{li} := \lambda_l \psi_{li}$. Consequently,

$$\|P_i - P_j\|^2_{\ell^2/d} = \sum_l \lambda^2 \left( \psi_{li} - \psi_{lj} \right)^2.$$

Analogously, for $P^t$ we have

$$\|P_i^t - P_j^t\|^2_{\ell^2/d} = \sum_l \lambda^{2t} \left( \psi_{li} - \psi_{lj} \right)^2.$$

□

## 2.1 Representation

The above suggests that we can represent a node $v_i$ using a feature vector

$$\Psi_t(v_i) = \left( \lambda_1^t \psi_{1,i}^t, \dots, \lambda_1^t \psi_{L,i}^t \right)^T,$$

for some $L \leq n$.

Recall that by proposition 1.11, $L_{\text{sym}}$ is positive semi-definite. This implies that the entire spectrum of $L_{\text{rw}}$ lies between 1 and 0, and consequently also the spectrum of $P$. Since the spectrum decays, for each $t$ only few eigenvalues $\lambda^t$ have significant effect on the diffusion distances, which means that we can use dimensionality $L$ such that $\lambda_L^t$ is sufficiently small, and achieve dimensionality reduction.