Machine Learning on Graphs
MDI343
Spectral Embedding

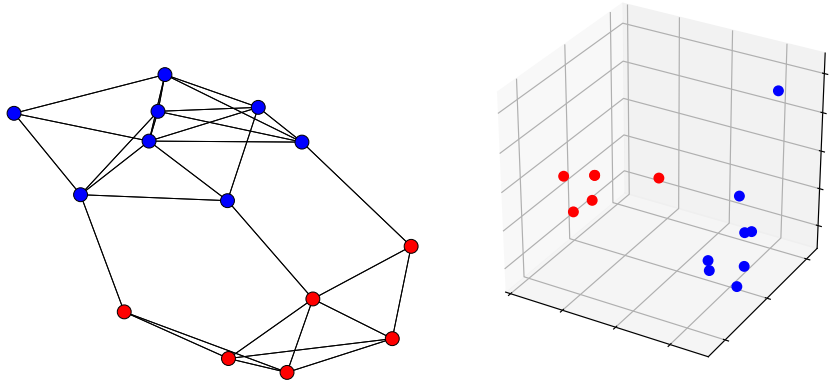
Thomas Bonald
Institut Polytechnique de Paris

2020 – 2021

These lecture notes introduce graph embedding, a technique consisting in transforming graph data into vector data. Specifically, the objective is to represent each node of the graph by a vector of low dimension, so that "close" nodes in the graph remain close in the vector space (for the Euclidean distance). This embedding can in turn be used to apply standard learning techniques, like classification or clustering. We here present a classical approach based on the spectral decomposition of the Laplacian matrix. We also describe an approach based on the (generalized) singular value decomposition (SVD) of the adjacency matrix, and show the link with the spectral approach. We refer the reader to [1, 2, 3] for advanced material on this topic.

1 Notion of embedding

Consider an undirected graph $G = (V, E)$ of n nodes, with adjacency matrix A. We denote by $D = \text{diag}(d)$ with $d = A1$ the diagonal matrix of node degrees. The graph is assumed to be connected. We aim at representing the graph in some vector space of low dimension, say $\mathbb{R}^K$ with K much lower than n. Specifically, each node $i \in V$ is represented by some vector $X_i \in \mathbb{R}^K$. We denote by X the matrix of dimension $n \times K$ whose i-th row X_i corresponds to the embedding of node i. The structure of the graph must be encoded in its representation X in the sense that two "close" nodes i, j in the graph should correspond to two "close" vectors X_i, X_j in the embedding space (see Figure 1).



(a) Graph. (b) Embedding in $\mathbb{R}^3$.

Figure 1: Graph embedding.

2 Spectral embedding

A natural approach to graph embedding is to minimize the expected square distance between nodes that are connected. Forcing the embedding to be centered, we get the optimization problem:

$$\min_{X:X^T1=0} \sum_{i,j\in V} A_{ij}||X_i - X_j||^2.$$

Of course, this optimization problem alone is not interesting as the solution is trivial, with all nodes located at the origin (i.e., $X = 0$). We need a constraint to force the embedding to occupy the vector space, e.g.,

$$\min_{X:X^T1=0, X^TX=I_K} \sum_{i,j\in V} A_{ij}||X_i - X_j||^2. \tag{1}$$

Observe that X^TX is, up to a normalization constant, the covariance matrix of the random vector $X_i \in \mathbb{R}^K$ with node i sampled uniformly at random. The constraint $X^TX = I_K$ forces the coordinates of the embedding to have the same (positive) variance and to be uncorrelated.

Denoting by $L = D - A$ the Laplacian matrix of the graph, we have the following key result:

Lemma 1 *We have:*

$$\mathrm{tr}(X^TLX) = \frac{1}{2}\sum_{i,j\in V} A_{ij}||X_i - X_j||^2.$$

Proof. First note that it is enough to prove the result for $K = 1$, as both sides of the equality are sums over the components $k = 1, \ldots, K$. Then X is a vector of dimension n and we get:

$$\begin{aligned} X^TLX &= X^T(D - A)X, \\ &= \sum_{i=1}^n d_iX_i^2 - \sum_{i,j=1}^n X_iA_{ij}X_j, \\ &= \sum_{i,j=1}^n A_{ij}X_i(X_i - X_j) = \frac{1}{2}\sum_{i,j=1}^n A_{ij}(X_i - X_j)^2. \end{aligned} \tag{2}$$

□

In view of (2), the Laplacian matrix is positive semi-definite. Since it is symmetric, there is an orthogonal matrix $V = (v_1, \ldots, v_n)$ of eigenvectors:

$$LV = V\Lambda, \quad V^TV = I, \tag{3}$$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\lambda_1 = 0 \leq \lambda_2 \leq \ldots \leq \lambda_n$. The multiplicity of the eigenvalue $\lambda_1 = 0$ is equal to the number of connected components of the graph. Since we assume the graph connected, the eigenvalue $\lambda_1 = 0$ is simple (i.e., $\lambda_2 > 0$) and $v_1 \propto 1$.

Theorem 1 *We have:*

$$\min_{X:X^T1=0, X^TX=I_K} \mathrm{tr}(X^TLX) = \sum_{k=2}^{K+1} \lambda_k. \tag{4}$$

The minimum is reached for X equal to the matrix of eigenvectors of the Laplacian matrix associated with the eigenvalues $\lambda_2, \ldots, \lambda_{K+1}$.

Proof. First consider the problem without the centering and orthogonality constraints, i.e.,

$$\min_{X:\mathrm{diag}(X^TX)=I_K} \mathrm{tr}(X^TLX),$$

where $\mathrm{diag}(M)$ refers to the diagonal matrix with the same diagonal as M. The Lagrangian of this optimization problem is:

$$\mathcal{L} = \mathrm{tr}(X^T L X - (X^T X - I_K)\Gamma),$$

where Γ is the diagonal matrix of the K Lagrange multipliers associated with the constraint $\mathrm{diag}(X^T X) = I_K$. Taking the gradient with respect to X gives:

$$LX = X\Gamma,$$

so that X is a matrix of eigenvectors with eigenvalues equal to the Lagrange multipliers. Taking the orthogonality constraint into account yields $\mathrm{tr}(X^T L X) = \mathrm{tr}(\Gamma)$, which is a sum of eigenvalues of L. In view of the centering constraint, the first eigenvalue must be skipped and $\mathrm{tr}(X^T L X)$ is minimized for $\Gamma = \mathrm{diag}(\lambda_2, \ldots, \lambda_{K+1})$. □

We refer to the spectral embedding of the graph as the matrix X of eigenvectors of the Laplacian matrix associated with the eigenvalues $\lambda_2, \ldots, \lambda_{K+1}$. In view of Lemma 1 and Theorem 1, it solves (1), i.e., it is optimal with respect to the expected square distance between nodes sampled from the edges, under the constraint that the coordinates are centered, with a unit covariance matrix.

3 A mechanical system

The spectrum of the Laplacian can be interpreted through the following mechanical system[1]. Consider n points of unit mass where points i and j are linked by a spring of unit stiffness following Hooke's law (i.e., attractive force proportional to the distance). Now if the points are located according to some vector $x \in \mathbb{R}^n$ along a line, the potential energy accumulated in the springs is:

$$\frac{1}{2}\sum_{i<j} A_{ij}(x_i - x_j)^2,$$

that is $\frac{1}{2}x^T L x$ in view of Lemma 1.

Energy. Assume that the moment of inertia of the system (for a rotation around the origin) is equal to 1, that is $x^T x = 1$. Clearly, the vector x that minimizes the potential energy is $x \propto 1$ (the corresponding potential energy is null). Now if we impose $x^T 1 = 0$, meaning that the centre of mass is at the origin, we obtain $x = v_2$ (the eigenvector known as the Fiedler vector) and $x^T L x = \lambda_2$, so that the eigenvalue λ_2 corresponds to twice the minimum value of potential energy. More generally, the spectrum of the Laplacian can be interpreted as levels of energy of the mechanical system, as shown by the following result.

Theorem 2 *For all $k = 1, \ldots, n$,*

$$\lambda_k = \min_{\substack{x : x^T x = 1 \\ x^T v_1 = 0, \ldots, x^T v_{k-1} = 0}} x^T L x, \tag{5}$$

the minimum being attained for $x = v_k$.

Proof. Let $x \in \mathbb{R}^n$ such that $x^T x = 1$. The vector $y = V^T x$, corresponding to the coordinates of x in the basis of eigenvectors, satisfies:

$$y^T \Lambda y = x^T V \Lambda V^T x = x^T L x \quad \text{and} \qquad y^T y = x^T V V^T x = 1,$$

so that the optimization problem (5) is equivalent to:

$$\min_{\substack{y : y^T y = 1 \\ y_1 = 0, \ldots, y_{k-1} = 0}} y^T \Lambda y.$$

[1] Another physical interpretation exists using an electrical system, see [4]

The result then follows from the equality:

$$y^T \Lambda y = \sum_{k=1}^{n} \lambda_k y_k^2,$$

which is minimized for $y_k = 1$. □

Harmonic oscillator. Now consider the dynamical system, with nodes of unit mass. Let $x(t) \in \mathbb{R}^n$ be the state of the system at time t (location of each node on the line). The force exerted on node i is:

$$\sum_{j \in V} A_{ij}(x_j - x_i) = -(Lx)_i.$$

By Newton's law, we get:

$$-Lx = \ddot{x}.$$

Letting $x(t) = xe^{j\omega t}$, we get:

$$Lx = \omega^2 x.$$

We deduce that the eigenvectors of L correspond to the eigenmodes of the dynamical system; the square roots of the eigenvalues give the corresponding eigenfrequencies.

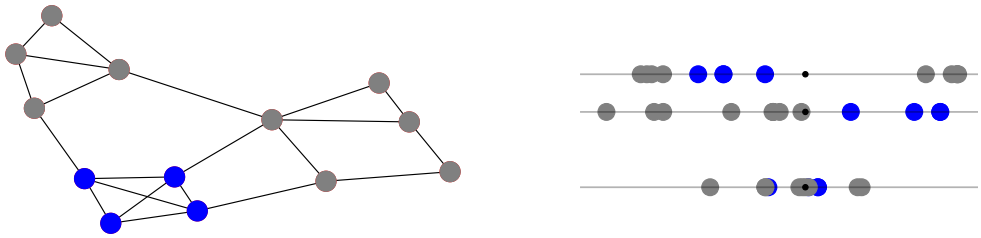


Figure 2: A graph and some eigenmodes of the corresponding mechanical system.

4 Generalized eigenvalues

The constraint of the optimization problem (1) involves the covariance matrix of the random vector $X_i \in \mathbb{R}^K$ with node i sampled *uniformly* at random. Another natural constraint follows from edge sampling. We get:

$$\min_{X : X^T d = 0, X^T D X = I_K} \sum_{i,j \in V} A_{ij} ||X_i - X_j||^2. \tag{6}$$

Now both the centering constraint $X^T d$ and the covariance constraint $X^T D X = I_K$ correspond to nodes sampled in proportion to their degrees.

Generalized eigenvalue problem. The solution involves the following *generalized* eigenvalue problem[2]:

$$LV = DV\Lambda, \quad V^T D V = I, \tag{7}$$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\lambda_1 = 0 \leq \lambda_2 \leq \ldots \leq \lambda_n$. These generalized eigenvectors are the eigenmodes of the mechanical system with vector of masses d. In particular, all eigenmodes (except the first) have their

[2]We use the same notation as for the regular spectral decomposition of the Laplacian matrix but both the eigenvectors and the eigenvalue are different.

center of inertia at the origin. We have $V = D^{-\frac{1}{2}}U$ where U is the orthogonal matrix of eigenvectors of the *normalized* Laplacian matrix:

$$\bar{L} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}.$$

Observe that, like the Laplacian matrix, the normalized Laplacian matrix is symmetric and positive semi-definite and thus has a spectral decompostion of the form:

$$\bar{L}U = U\Lambda, \quad U^TU = I,$$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\lambda_1 = 0 < \lambda_2 \leq \ldots \leq \lambda_n$.

We have the analogue of Theorem 1 (the proof is similar):

Theorem 3 *We have*

$$\min_{X : X^Td=0, X^TDX=I_K} \mathrm{tr}(X^TLX) = \sum_{k=2}^{K+1} \lambda_k. \tag{8}$$

The minimum is reached for X equal to the matrix of generalized eigenvectors of the Laplacian matrix associated with the eigenvalues $\lambda_2, \ldots, \lambda_{K+1}$.

Transition matrix. Let $P = D^{-1}A$ be the transition matrix of the random walk in the graph. In view of (7), we have:

$$PV = V(I - \Lambda), \quad V^TDV = I, \tag{9}$$

so that V is also a matrix of eigenvectors of P, with respective eigenvalues $\mu_1 = 1 > \mu_2 \geq \ldots \geq \mu_n \geq -1$ (the modulus of each eigenvalue cannot exceed 1 as the matrix P is stochastic), with $\mu_n > -1$ unless the graph is bipartite.

Spectral embedding. We refer to the (generalized) spectral embedding of the graph as the matrix X of generalized eigenvectors of the Laplacian matrix associated with the eigenvalues $\lambda_2, \ldots, \lambda_{K+1}$. This is also known as Laplacian eigenmaps [1]. In view of Lemma 1 and Theorem 3, it solves (6), i.e., it is optimal with respect to the expected square distance between nodes, under centering and covariance constraints, with nodes sampled from the edges for both the objective function and the constraints.

Barycenter. In view of (9), we have:

$$PX = XM, \quad M = \mathrm{diag}(\mu_2, \ldots, \mu_{K+1}), \tag{10}$$

that is,

$$\forall i \in V, \quad X_iM = \sum_{j \in V} P_{ij}X_j.$$

Thus the location of each node in the embedding space is, up to the scaling by the eigenvalues of the transition matrix P, that of the *barycenter* of its neighbors.

Scaling. Consider the following embedding:

$$Y = XM, \quad M = \mathrm{diag}(\mu_2, \ldots, \mu_{K+1}), \tag{11}$$

where X is the spectral embedding of the graph in dimension K. The embedding is still centered but the covariance matrix is now:

$$Y^TDY = M^2.$$

Assuming that K is such that $\mu_2, \ldots, \mu_{K+1} \geq 0$, this scaling gives more weight to the first eigenvectors, corresponding to eigenmodes of lower energy in the mechanical system. Since Y are also eigenvectors of the transition matrix, i.e., $PY = YM$, the barycenter property is preserved. In view of (10), $Y = PX$ so that the embedding Y may be viewed as the projection of the transition matrix P, a representation of the graph in dimension n, over X, a basis of K generalized eigenvectors of the Laplacian matrix.

5 Extensions

Weighted graphs. The spectral embedding can be extended to weighted graphs, with A the weighted adjacency matrix and D the diagonal matrix of node weights. In the mechanical system, the stiffness of the spring between two nodes is equal to the weight of the corresponding edge, if any. Observe that the spectral embedding still minimize the expected square distance between nodes sampled from the edges (in proportion to their weights).

Bipartite graphs. The spectral embedding also apply to bipartite graphs, seen as undirected graphs. Specifically, a bipartite graph $G = (V_1, V_2, E)$ with biadjacency matrix B is an undirected graph with adjacency matrix:

$$A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}.$$

The corresponding diagonal matrix of node degrees is:

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix},$$

with $D_1 = \operatorname{diag}(B1)$ and $D_2 = \operatorname{diag}(B^T 1)$ are the diagonal matrices of the degrees of each part of the graph. We get an embedding of the form:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

where X_1 and X_2 are the embeddings of each part of the graph, V_1 and V_2. The barycenter property can be written:

$$\begin{aligned} P_1 X_2 &= X_1 M \\ P_2 X_1 &= X_2 M \end{aligned} \tag{12}$$

with $P_1 = D_1^{-1} B$ the transition matrix from V_1 to V_2, $P_2 = D_2^{-1} B^T$ the transition matrix from V_2 to V_1, and $M = \operatorname{diag}(\mu_2, \ldots, \mu_{K+1})$ the diagonal matrix of the first K eigenvalues (skipping the first) of the transition matrix:

$$P = \begin{bmatrix} 0 & P_1 \\ P_2 & 0 \end{bmatrix}.$$

In view of (12), the embedding of each part is, up to some scaling by the eigenvalues, given by the barycenter of the embedding of the other part.

Observe that the spectrum of P is symmetric in the sense that if μ is an eigenvalue for P, then $-\mu$ is also an eigenvalue for P:

$$P \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \mu \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad \Longrightarrow \quad P \begin{bmatrix} X_1 \\ -X_2 \end{bmatrix} = -\mu \begin{bmatrix} X_1 \\ -X_2 \end{bmatrix}$$

In practice, this means that it is useless to account for negative eigenvalues. The spectral embedding should be restricted to eigenvectors associated with positive eigenvalues of the transition matrix, possibly after scaling as in (11).

Directed graphs. A simple way to get the embedding of a directed graph $G = (V, E)$ with adjacency matrix A is to view it as a bipartite graph with biadjacency matrix $B = A$. Each node of G is represented twice in the bipartite graph, once as a source of edges and the other as a destination of edges. We refer to this graph as the *mirror graph*. The embedding of the graph G is then taken as the spectral embedding of the first part of the mirror graph, say X_1. Again, the embedding may be scaled by the (positive) eigenvalues of the transition matrix, as in (11).

6 Singular value decomposition

Another standard way to reduce dimension is through the singular value decomposition (SVD), that provides the best low-rank approximation of any matrix. We here consider the *generalized* SVD of the biadjacency matrix B of some bipartite graph $G = (V_1, V_2, E)$, given by:

$$\begin{array}{l} BV = D_1 U \Sigma \\ B^T U = D_1 V \Sigma \end{array} \quad \text{with} \quad U^T D_1 U = V^T D_2 V = I, \tag{13}$$

with $D_1 = \text{diag}(B1)$, $D_2 = \text{diag}(B^T 1)$, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$ the diagonal matrix of generalized singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ and r the rank of B. Observe that B may be replaced by the adjacency matrix of any graph, which is equivalent to consider the associate mirror graph as above.

It is easy to check that the generalized SVD of the biadjacency matrix B is equivalent to the spectral decomposition of the transition matrix P of the graph G, with adjacency matrix:

$$A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}.$$

Specifically, we have:

$$P \begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} \Sigma.$$

Thus the generalized singular vectors of B give the eigenvectors of P associated with positive eigenvalues; the embeddings X_1 and X_2 provided by first K generalized singular vectors of B (skipping the first) correspond to the spectral embedding of the bipartite graph G.

Another interpretation of the embedding X_1 (resp. X_2) is through the coneighbor graph G_1 (resp. G_2), a graph with nodes V_1 (resp. V_2) and weighted adjacency matrix $A_1 = BD_2^{-1}B^T$ (resp. $A_2 = B^T D_1^{-1} B$). The weight between nodes $i, j \in V_1$ is given by:

$$\sum_{k \in V_2} \frac{B_{ik} B_{jk}}{d_k}.$$

Observe that this is positive if and only if i and j have common neighbors. Since $d_1 = A_1 1$ (i.e., the node weights are preserved), it follows from (13) that:

$$P_1 X_1 = X_1 \Sigma^2, \quad X_1^T D_1 X_1 = I_K,$$

where P_1 is the transition matrix of the random walk in the co-neighbor graph G_1. We deduce that X_1 is exactly the Laplacian eigenmap of the co-neighbor graph G_1. The corresponding eigenvalues are positive.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 2003.

[2] L. Lovász. Random walks on graphs. *Combinatorics, Paul Erdos is eighty*, 1993.

[3] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 2007.

[4] P. Snell and P. Doyle. Random walks and electric networks. *Free Software Foundation*, 2000.