The Laplacian Matrix and Spectral Graph Drawing. Courant-Fischer.

Zhiping (Patricia) Xiao

University of California, Los Angeles

October 8, 2020

Outline

Introduction
- Resources
- Basic Problems Examples
- Background

Eigenvalues and Optimization: The Courant-Fischer Theorem

The Laplacian and Graph Drawing

Intuitive Understanding of Graph Laplacian

Introduction

Course: Spectral Graph Theory from Yale.
Textbooks include:

- Spectral and Algebraic Graph Theory (Daniel A. Spielman)
- Scalable Algorithms for Data and Network Analysis (Shang-Hua Teng)

Objective of the course:

- To explore what eigenvalues and eigenvectors of graphs can tell us about their structure.

Prerequisites:

- Linear algebra, graphs, etc.

Textbook chapters:

- Spectral and Algebraic Graph Theory (Daniel A. Spielman) Chap $1 \sim 3$
- Scalable Algorithms for Data and Network Analysis (Shang-Hua Teng) Chap 2.4

Supplementary Materials:

- Prof. Cho's additional explanations on the matrices;
- The points Prof. Sun brought up on the random walk matrix $\mathbf{W}_G$ and the Courant-Fischer Theorem;
- Yewen's note related to Courant-Fischer Theorem `https://www.overleaf.com/read/bsbwwbckptpk`.

Problems listed in Prof. Teng's book Chap 2.4

- Significant Nodes: Ranking and Centrality
- Coherent Groups: Clustering and Communities
- Interplay between Networks and Dynamic Processes
- Multiple Networks: Composition and Similarity

Significant Nodes: Ranking and Centrality

Identifying nodes of relevance and significance. e.g.:

> *Which nodes are the most* ***significant*** *nodes in a network or a sub-network? How quickly can we identify them?*

Significance could be measured either *numerically*, or by *ranking* the nodes.

> *Network centrality is a form of "dimensionality reduction" from "high dimensional" network data to "low dimensional" centrality measures or rankings.*

e.g. PageRank

Identifying groups with significant structural properties.
Fundamental questions include:

- What are the significant clusters in a data set?
- How fast can we identify one, uniformly sample one, or enumerate all significant groups?
- How should we evaluate the consistency of a clustering or community-identification scheme?
- What desirable properties should clustering or community identification schemes satisfy?

Understanding the interplay between dynamic processes and their underlying networks.

A given social network can be part of different dynamic processes (e.g. epidemic spreading, viral marketing), which can potentially affect the relations between nodes. Fundamental questions include:

- How should we model the interaction between network nodes in a given dynamic process?
- How should we characterize node significance and group coherence with respect to a dynamic process?
- How fast can we identify influential nodes and significant communities?

Multiple Networks: Composition and Similarity

To understand multiple networks instead of individual networks.

- network composition, e.g. multi-layer social network, multi-view graphs
- network similarity
  - similarity between two different networks
  - construct a sparser network that approximates a known one

$G = (V, E)$ (Friendship graphs, Network graphs, Circuit graphs, Protein-Protein Interaction graphs, etc.)

- G: a graph/network
- V: its vertex/node set
- E: its edge set (pair of vertices); edges have weight 1 by default, could assign other weights optionally.

By default (unless otherwise specified), a graph to be discussed will be:

- undirected (unordered vertices pairs in E)
- simple (having no loops or multiple edges)
- finite (V and E being finite sets)

Why we care about matrices?
Given a vector $\mathbf{x} \in \mathbb{R}^n$ and a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$

- $\mathbf{M}$ could be an operator: $\mathbf{Mx} \in \mathbb{R}^n$
- $\mathbf{M}$ could be used to define a quadratic form: $\mathbf{x}^T\mathbf{Mx} \in \mathbb{R}$ (here it has to be $m == n$)

Adjacency matrix $\mathbf{M}_G$ of $G = (V, E)$:

$$\mathbf{M}_G(a, b) = \begin{cases} 1 & \text{if } (a, b) \in E \\ 0 & \text{otherwise} \end{cases}$$

- most natural matrix to associate with a graph
- least "useful" (means directly useful, but useful in terms of generating other matrices)

This statement is made because it is only a ***spread-sheet****, neither a natural* ***operator*** *or a natural* ***quadratic form****.*

Diffusion operator $\mathbf{D}_G$ of $G = (V, E)$ is a diagonal matrix, probably the most natural operator associated with G:

$$\mathbf{D}_G(a, a) = \mathbf{d}(a)$$

where $\mathbf{d}(a)$ is the degree of vertex a.

- unweighted case: number of edges attached to it
- weighted case: weighted degree

$$\mathbf{d} \stackrel{\text{def}}{=} \mathbf{M}_G \mathbb{1}$$

There is a linear operator $\mathbf{W}_G$ defined as:

$$\mathbf{W}_G = \mathbf{M}_G \mathbf{D}_G^{-1}$$

regarded as an operator denoting the *changes* of the graph between time steps.
Recall that diffusion operator $\mathbf{D}_G$ is a diagonal matrix, $\mathbf{W}_G$ is merely a rescaling of $\mathbf{M}_G$ if the graph is *regular* [1].

With vector $\mathbf{p} \in \mathbb{R}^n$ denoting the values of n vertices (called "*distribution of how much stuff*" in the textbook), the distribution of stuff at each vertex will be $\mathbf{W}_G \mathbf{p}$.

[1] Regular graph's vertices have the same degree.

This matrix is called a random-walk Markov matrix: [2]

$$\mathbf{W}_G = \mathbf{M}_G \mathbf{D}_G^{-1}$$

The next time step is:

$$\mathbf{W}_G \mathbf{p} = \mathbf{M}_G \mathbf{D}_G^{-1} \mathbf{p}$$

Think about the case where $\mathbf{p}$ is a one-hot vector δ_a where only $\delta_a(a) = 1$ and all other elements are 0.

$$\mathbf{W}_G \delta_a = \mathbf{M}_G \mathbf{D}_G^{-1} \delta_a = \mathbf{M}_G (\mathbf{D}_G^{-1} \delta_a)$$

We find the vector $\mathbf{D}_G^{-1}\delta_a$ has value $1/\mathbf{d}(a)$ at vertex a and 0 everywhere else; $\mathbf{M}_G \mathbf{D}_G^{-1}\delta_a$ has value $1/\mathbf{d}(a)$ at all a's **neighbors** and 0 otherwise.

[2] Reference from www.cmm.ki.si.

A commonly-seen form of $\mathbf{W}_G$ is sometimes more convenient:

$$\widetilde{\mathbf{W}}_G = \mathbf{I}/2 + \mathbf{W}_G/2$$

describing a *lazy random walk* (1/2 chance stay, 1/2 chance go).

> *One of the purposes of spectral theory is to understand what happens when a linear operator like* $\mathbf{W}_G$ *is repeatedly applied.*

That is why it is called a random walk Markov matrix.

$$\mathbf{W}_G = \mathbf{M}_G \mathbf{D}_G^{-1}$$

has each column summing up to 1. $\mathbf{W}_G(a, b)$, the value on the a^{th} row b^{th} column, is $\mathbf{d}(b)$ if $(a, b) \in E$ else 0.

In fact, what $\mathbf{W}_G\mathbf{p}$ resulting in is a "random walk" based on the neighbors' degree.

$\mathbf{W}_G^T\mathbf{p}$ will be the random walk based on the degree of each node itself. (An example in the upcoming page.) It could be computed as:

$$\mathbf{W}_G^T = \mathbf{D}_G^{-1} \mathbf{M}_G$$

An example:

$$\mathbf{M}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{D}_G = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{D}_G^{-1} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{W}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 0 & 0 \end{bmatrix}$$

$$\mathbf{W}_G\mathbf{p} = \begin{bmatrix} p_2 + p_3 \\ p_1/2 \\ p_1/2 \end{bmatrix} \qquad \mathbf{W}_G^T\mathbf{p} = \begin{bmatrix} (p_2 + p_3)/2 \\ p_1 \\ p_1 \end{bmatrix}$$

Laplacian matrix $\mathbf{L}_G$, the most natural quadratic form associated with the graph G:

$$\mathbf{L}_G \stackrel{\text{def}}{=} \mathbf{D}_G - \mathbf{M}_G$$

Given a vector $\mathbf{x} \in \mathbb{R}^n$, who could also be viewed as a *function* over the vertices, we have: [3]

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \sum_{(a,b)\in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

representing the Laplacian quadratic form of a weighted graph ($w_{a,b}$ is the weight of edge (a, b)), could be used to measures the smoothness of $\mathbf{x}$ (it would be small if $\mathbf{x}$ is not changing drastically over any edge).

[3] Note that G has to be undirected

An example ($w_{a,b} = 1$):

$$\mathbf{M}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{D}_G = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{L}_G = \mathbf{D}_G - \mathbf{M}_G = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} \mathbf{x}^T \mathbf{L}_G \mathbf{x} &= x_1(2x_1 - x_2 - x_3) + x_2(-x_1 + x_2) + x_3(-x_1 + x_3) \\ &= 2x_1^2 + x_2^2 + x_3^2 - 2x_1x_2 - 2x_1x_3 = (x_1 - x_2)^2 + (x_1 - x_3)^2 \end{aligned}$$

*Intuitively, $\mathbf{L}_G$, $\mathbf{D}_G$ and $\mathbf{M}_G$ could be viewed as the sum of many subgraphs, each containing **one** edge.*

Matrices for Graphs (*) [4]

Incidence Matrix: $\mathbf{I}_G$, where each row corresponds to an edge, and columns to vertices indexes.

A row, corresponding to $(a, b) \in E$, sums up to 0, with only 2 non-zero elements: the a^{th} column being 1 and b^{th} being -1, or could be the opposite (a^{th} column -1 and b^{th} column 1).
Following the previous example:

$$\mathbf{M}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{I}_G = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

In the case of weighted graph, ± 1 should be $\pm w_{a,b}$ instead.
There's very interesting relation:

$$\mathbf{L}_G = \mathbf{I}_G^T \mathbf{I}_G$$

[4] This part comes from Prof. Cho's explanations.

Explanation on the reason why:

$$\mathbf{L}_G = \mathbf{I}_G^T \mathbf{I}_G$$

could be from the perspective that, $\mathbf{L}_G$ is associated with Hessian and $\mathbf{I}_G$ be associated with Jacobian.

Also note that the introduction of the Incidence Matrix immediately makes this proof obvious:

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \sum_{(a,b)\in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \mathbf{x}^T \mathbf{I}_G^T \mathbf{I}_G \mathbf{x} = \|\mathbf{I}_G \mathbf{x}\|^2 = \sum_{(a,b)\in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

In practice we always use **normalized** Laplacian matrices. Intuitively, we want all diagonal entries to be 1. In a way, that is somewhat "regularize" of the matrix.

There are many ways of normalizing a Laplacian matrix. Two of them are:

- (symmetric)

$$\mathbf{L}_s = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$$

- (random walk)

$$\mathbf{L}_{rw} = \mathbf{L}\mathbf{D}^{-1} = (\mathbf{D} - \mathbf{M})\mathbf{D}^{-1} = \mathbf{I} - \mathbf{M}\mathbf{D}^{-1}$$

$\mathbf{L}_s$ preserves every property of $\mathbf{L}$. Such as being positive semidefinite:

$$\mathbf{x}^T\mathbf{L}_s\mathbf{x} = \sum_{(a,b)\in E} w_{a,b}\left(\frac{\mathbf{x}(a)}{\sqrt{\mathbf{d}(a)}} - \frac{\mathbf{x}(b)}{\sqrt{\mathbf{d}(b)}}\right)^2$$

Recall that $\mathbf{MD}^{-1}$ is the random walk Markov matrix $\mathbf{W}$. $\mathbf{L}_{rw} = \mathbf{I} - \mathbf{W}$. Therefore, $\mathbf{W}$ and $\mathbf{L}_{rw}$ have the same eigenvectors, while the corresponding eigenvalues sum up to 1:

$$\mathbf{A}\mathbf{x} = \mu\mathbf{x} \iff (\mathbf{A} - k\mathbf{I})\mathbf{x} = (\mu - k)\mathbf{x}$$

$$\mathbf{W}\boldsymbol{\psi} = \lambda\boldsymbol{\psi} \iff (\mathbf{I} - \mathbf{W})\boldsymbol{\psi} = (1 - \lambda)\boldsymbol{\psi}$$

Additional comments on λ and $1 - \lambda$:

Sometimes, for $0 \leq \lambda \leq 1$, after some operations, such as multiplying the matrix (say, $\mathbf{A}$) for multiple times, small eigenvalues will become close to zero.

However, if we consider a trick:

$$\mathbf{I} - \mathbf{A}$$

the corresponding eigenvalue will be $0 \leq 1 - \lambda \leq 1$. After power iteration, the smallest eigenvalue becomes the largest.

Review: the spectral theory for symmetric matrices (or those similar to symmetric matrices).

> $\mathbf{A}$ *is similar to* $\mathbf{B}$ *if there exists non-singular* $\mathbf{X}$ *such that* $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{B}$.

A vector $\boldsymbol{\psi}$ is an eigenvector of a matrix $\mathbf{M}$ with eigenvalue λ if:

$$\mathbf{M}\boldsymbol{\psi} = \lambda\boldsymbol{\psi}$$

λ is an eigenvalue if and only if $\lambda\mathbf{I} - \mathbf{M}$ is a singular matrix ($\therefore \det(\lambda\mathbf{I} - \mathbf{M}) = 0$). The eigenvalues are the **roots** of the *characteristic polynomial* of $\mathbf{M}$:

$$\det(x\mathbf{I} - \mathbf{M})$$

in other words, being a solution to the *characteristic equation*:

$$\det(x\mathbf{I} - \mathbf{M}) = 0$$

Additional explanation on why "λ is an eigenvalue if and only if $\lambda\mathbf{I} - \mathbf{M}$ is a singular matrix": [5]

$$\mathbf{M}\boldsymbol{\psi} = \lambda\boldsymbol{\psi}$$

$$(\lambda\mathbf{I} - \mathbf{M})\boldsymbol{\psi} = 0$$

is a homogeneous linear system for $\boldsymbol{\psi}$, with a trivial zero solution ($\boldsymbol{\psi} = 0$).
A homogeneous linear system has a nonzero solution $\boldsymbol{\psi} \neq 0$ iff its coefficient matrix (in this case, $\lambda\mathbf{I} - \mathbf{M}$), is singular.

[5] https://www-users.math.umn.edu/~olver/num_/lnv.pdf

Theorem (1.3.1 The Spectral Theorem)

If $\mathbf{M}$ *is an n-by-n, real, symmetric matrix,* ***then*** *there exist real numbers* $\lambda_1, \ldots, \lambda_n$ *and n mutually orthogonal unit vectors* $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ *and such that* $\boldsymbol{\psi}_i$ *is an eigenvector of* $\mathbf{M}$ *of eigenvalue* λ_i*, for each i.*

If the matrix $\mathbf{M}$ is not symmetric, it might not have n eigenvalues. And, even if it has n eigenvalues, their eigenvectors will not be orthogonal (linearly independent). Many studies will no longer apply to it when the matrix is not symmetric.

Review: solving the eigenvalues and eigenvectors. [6]

$$\mathbf{M} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$$

$$\mathbf{M}\boldsymbol{\psi} = \lambda\boldsymbol{\psi}$$

$$(\mathbf{M} - \lambda\mathbf{I})\boldsymbol{\psi} = 0$$

The determinant value of $\mathbf{M} - \lambda\mathbf{I}$ is 0 (by definition of the singular matrix, etc.).

$$\det(\mathbf{M} - \lambda\mathbf{I}) = 0$$

$$\det\left(\begin{bmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix}\right) = \lambda^2 + 3\lambda + 2 = (\lambda+1)(\lambda+2) = 0$$

The eigenvalues are:

$$\lambda_1 = -1, \lambda_2 = -2$$

Next we want to find the corresponding eigenvectors $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, by solving:

$$(\mathbf{M} - \lambda \mathbf{I})\boldsymbol{\psi} = 0$$

which means,

$$\begin{bmatrix} -\lambda_i & 1 \\ -2 & -3 - \lambda_i \end{bmatrix} \begin{bmatrix} \psi_{i,1} \\ \psi_{i,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\psi_{i,2} - \lambda_i \psi_{i,1} = 0$$
$$2\psi_{i,1} + (3 + \lambda_i)\psi_{i,2} = 0$$

With $\lambda_1 = -1$, we have:

$$\psi_{1,2} + \psi_{1,1} = 0$$
$$2\psi_{1,1} + 2\psi_{1,2} = 0$$

so the only constraint is that $\psi_{1,2} = -\psi_{1,1}$. We can choose any arbitrary constant k_1 and make it:

$$\boldsymbol{\psi}_1 = k_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

With $\lambda_2 = -2$, we have:

$$\psi_{2,2} + 2\psi_{2,1} = 0$$
$$2\psi_{2,1} + \psi_{2,2} = 0$$

again, we need an arbitrary constant k_2 and we have:

$$\boldsymbol{\psi}_2 = k_2 \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

We can also come up with an example where $\lambda_1 = \lambda_2$. For example:

$$\mathbf{M} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix}$$

$$\det(\mathbf{M} - \lambda \mathbf{I}) = 0$$

$$\det\left(\begin{bmatrix} -\lambda & 1 \\ -1 & 2-\lambda \end{bmatrix} \right) = \lambda^2 - 2\lambda + 1 = (\lambda - 1)^2 = 0$$

Then we have $\lambda_1 = \lambda_2 = 1$.

[6] lpsa.swarthmore.edu/MtrxVibe/EigMat/MatrixEigen.html

Eigenvalues are **uniquely** determined (but the values can be repeated), while eigenvectors are **NOT**.

- Specifically, if $\boldsymbol{\psi}$ is an eigenvector, then $k\boldsymbol{\psi}$ is as well, for any arbitrary constant real number k.
- If $\lambda_i = \lambda_{i+1}$, then $\boldsymbol{\psi}_i + \boldsymbol{\psi}_{i+1}$ will also be an eigenvector of eigenvalue λ_i. The eigenvectors of a given eigenvalue are only determined up to an orthogonal transformation.

$$\because (\lambda_i \mathbf{I} - \mathbf{M})\boldsymbol{\psi}_i = (\lambda_i \mathbf{I} - \mathbf{M})\boldsymbol{\psi}_{i+1} = 0$$

$$\therefore (\lambda_i \mathbf{I} - \mathbf{M})(\boldsymbol{\psi}_i + \boldsymbol{\psi}_{i+1}) = 0$$

Definition (1.3.2)

A matrix is positive definite if it is symmetric and all of its eigenvalues are positive. It is positive semidefinite if it is symmetric and all of its eigenvalues are nonnegative.

When a ***real*** $n \times n$ *matrix* $\mathbf{X}$ *being positive definite:* [a]

$$\forall y \in \mathbb{R}^n, y^T \mathbf{X} y > 0$$

[a] `https://mathworld.wolfram.com/PositiveDefiniteMatrix.html`

Fact (1.3.3)

The Laplacian matrix of a graph is positive semidefinite.

Proof (Fact 1.3.3)

Recall that previously we have that, for the Laplacian $\mathbf{L}_G$ *of (undirected) graph* G*, given a vector* $\mathbf{x} \in \mathbb{R}^n$*:*

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \sum_{(a,b) \in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

when the weights $w_{a,b}$ *are all non-negative, the value is non-negative as well.*

In practice, we always number the eigenvalues of the Laplacian from **smallest** to **largest**.

$$0 = \lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$$

We refer to $\lambda_2, \ldots \lambda_k$ (k is **small**) as low-frequency eigenvalues. λ_n is a high-frequency eigenvalue.

> *High and low frequency eigenmodes can be thought of as analogous to high and low frequency parts of the Fourier transform.* [7]

The second-smallest eigenvalue of the Laplacian matrix of a graph is zero ($\lambda_2 = 0$) iff the graph is disconnected. λ_2 is a measure of how well-connected the graph is. (See Chap 1.5.4 **The Fiedler Value**.)

[7] From a discussion on stackexchange.

In this textbook, eigenvalues are sometimes denoted as λ and sometimes denoted as μ.

To my observation, they tend to use λ when the eigenvalues are ordered from the smallest to the largest, and μ when ordered from the largest to the smallest.

e.g., in the later chapters we'll see: eigenvalues of the adjacency matrix is denoted as μ (recall that we use λ for Laplacian's eigenvalues) and $\mu_1 \geq \mu_2 \cdots \geq \mu_n$. This is to make μ_i corresponds to λ_i.

Eigenvalues and eigenvectors are very useful to solving vibrating system problems.

In practice, eigenvalues are often associated with frequency.

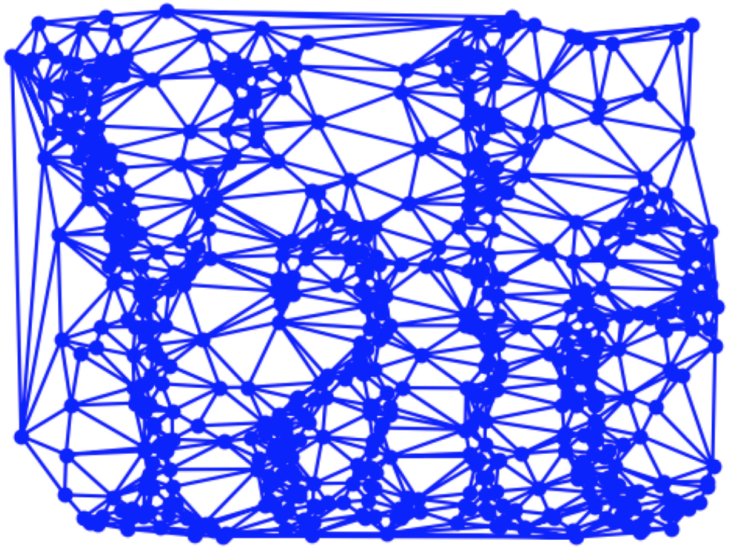
An example [8] have shown that, in *A Two-Mass Vibrating System*, they defined $\lambda = -\omega^2$.

ω values are then used to express the general solution:

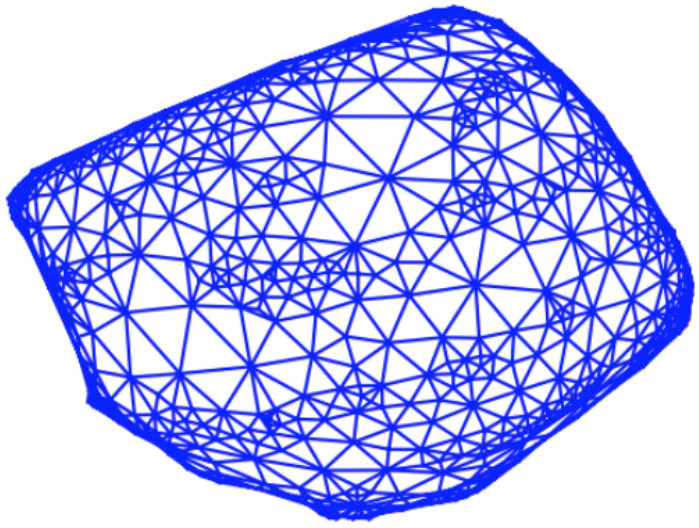
$$\mathbf{x}(t) = \sum_i c_{i,1}\mathbf{v}_i \cos(\omega_i t) + c_{i,2}\mathbf{v}_i \sin(\omega_i t)$$

where $\mathbf{v}_i$ are the corresponding eigenvector of ω_i.

[8] lpsa.swarthmore.edu/MtrxVibe/EigApp/EigVib.html



(a) The original points sampled from Yale logo, with coordinates omitted and transformed into graph.



(b) Plot of vertices at $(\boldsymbol{\psi}_2(a), \boldsymbol{\psi}_3(a))$ coordinate.

Figure: An example showing the use of eigenvectors. More examples are listed in the textbook, Chap 1.

Example: Why Eigenvectors as Coordinates (*) [9]

Intuitively, using eigenvalues and eigenvectors could be regarded as mapping the nodes onto sine and cosine function curves.

The sine and cosine functions generally preserve the distances between a pair of nodes, but for some disturbance brought by the periods (can have the same value again at another point). However, the use of multiple sets of eigenvalue-eigenvectors, could be viewed as having multiple frequencies to measure.

Therefore, a pair of nodes that is far away might seem to be close measured by sine or cosine value on a certain frequency, but won't be always close to each other under different frequencies.

[9] A summary of Prof. Cho's comments.

Example: Why Eigenvectors as Coordinates (*)

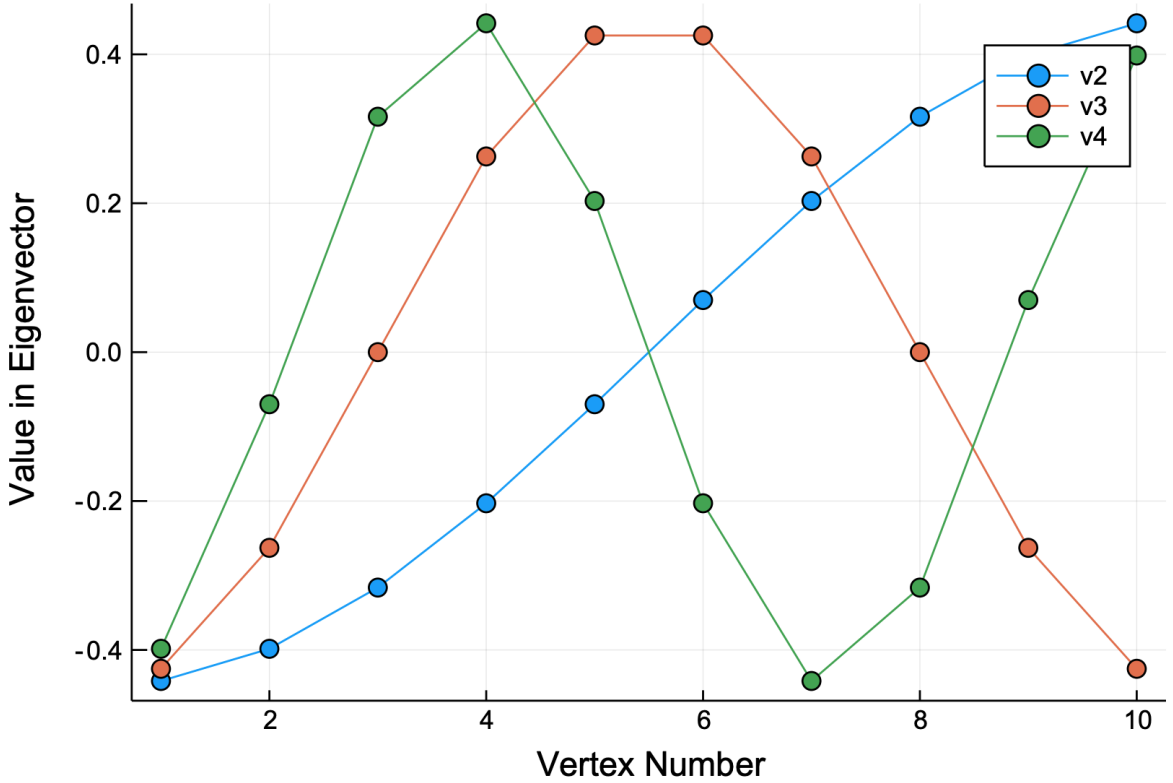


Figure: Plot of a length-4 path graph's (i.e. only $(i, i+1)$ are edges) Laplacian's eigenvectors $\mathbf{v}_2$, $\mathbf{v}_3$, $\mathbf{v}_4$, where $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \lambda_4$.

Eigenvalues and Optimization: The Courant-Fischer Theorem

One reason why we are interested in **eigenvalues** of matrices is that, they arise as the solution to natural **optimization** problems.

The formal statement of this is given by the **Courant-Fischer Theorem**. And this Theorem could be proved by the **Spectral Theorem**.

It has various other names: the min-max theorem, variational theorem, Courant–Fischer–Weyl min-max principle.

> *It gives a variational characterization of eigenvalues of* ***compact Hermitian operators*** *on* ***Hilbert spaces.***

- In the real-number field, a Hermitian matrix means a symmetric matrix.
- The real numbers $\mathbb{R}^n$ with $\langle \mathbf{u}, \mathbf{v} \rangle$ defined as the vector dot product of $\mathbf{u}$ and $\mathbf{v}$ is a typical finite-dimensional Hilbert space. [10]

[10] `https://mathworld.wolfram.com/HilbertSpace.html`

Theorem (2.0.1 Courant-Fischer Theorem)

Let $\mathbf{M}$ *be a symmetric matrix with eigenvalues* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. *Then,*

$$\mu_k = \max_{\substack{\mathcal{S} \subseteq \mathbb{R}^n \\ \dim(\mathcal{S}) = k}} \min_{\substack{\mathbf{x} \in \mathcal{S} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \min_{\substack{\mathcal{T} \subseteq \mathbb{R}^n \\ \dim(\mathcal{T}) = n-k+1}} \max_{\substack{\mathbf{x} \in \mathcal{T} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

where the maximization and minimization are over subspaces $\mathcal{S}$ *and* $\mathcal{T}$ *of* $\mathbb{R}^n$.

The Courant-Fischer Theorem: Proof I

Using the Spectral Theorem to prove the Courant-Fischer Theorem.

> **Theorem (1.3.1 The Spectral Theorem)**
>
> ***If*** $\mathbf{M}$ *is an n-by-n, real, symmetric matrix,* ***then*** *there exist real numbers* $\lambda_1, \ldots, \lambda_n$ *and n mutually orthogonal unit vectors* $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ *and such that* $\boldsymbol{\psi}_i$ *is an eigenvector of* $\mathbf{M}$ *of eigenvalue* λ_i*, for each i.*

Main Steps:

- expanding a vector $\mathbf{x}$ in the basis of eigenvectors of $\mathbf{M}$
- use the properties of eigenvalues and eigenvectors to prove it

The Courant-Fischer Theorem: Proof II

$\mathbf{M} \in \mathbb{R}^{n \times n}$: a symmetric matrix, with eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. The corresponding **orthogonal** eigenvectors are $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_n$.
Then we may write $\mathbf{x} \in \mathbb{R}^n$ as:

$$\mathbf{x} = \sum_i c_i \boldsymbol{\psi}_i, \qquad c_i = \boldsymbol{\psi}_i^T \mathbf{x}$$

Why $\mathbf{x}$ can be expanded in this way? (Intuitively obvious, but we need a mathematical explanation.)

Let $\boldsymbol{\Psi}$ be a matrix whose columns are $\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_n\}$ — orthogonal vectors. By definition, $\boldsymbol{\Psi}$ is an orthogonal matrix.

$$\boldsymbol{\Psi}\boldsymbol{\Psi}^T = \boldsymbol{\Psi}^T\boldsymbol{\Psi} = \boldsymbol{I}$$

Therefore we have:

$$\sum_i c_i\boldsymbol{\psi}_i = \sum_i \boldsymbol{\psi}_i c_i = \sum_i \boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\mathbf{x} = \Big(\sum_i \boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\Big)\mathbf{x} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T\mathbf{x} = \mathbf{x}$$

and thus, since $\boldsymbol{\psi}_i^T\boldsymbol{\psi}_j = 1$ when $i = j$ and 0 otherwise,

$$\mathbf{x}^T\mathbf{x} = \Big(\sum_i c_i\boldsymbol{\psi}_i\Big)^T\Big(\sum_i c_i\boldsymbol{\psi}_i\Big) = \sum_{i,j} c_i^2\boldsymbol{\psi}_i^T\boldsymbol{\psi}_j = \sum_{i=1}^n c_i^2$$

The Courant-Fischer Theorem: Proof IV

Let's revisit the theorem to prove (Now we have $\mathbf{x}^T\mathbf{x}$, to prove it we need to consider $\mathbf{x}^T\mathbf{M}\mathbf{x}$):

Theorem (2.0.1 Courant-Fischer Theorem)

Let $\mathbf{M}$ be a symmetric matrix with eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Then,

$$\mu_k = \max_{\substack{\mathcal{S}\subseteq\mathbb{R}^n \\ \dim(\mathcal{S})=k}} \min_{\substack{\mathbf{x}\in\mathcal{S} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \min_{\substack{\mathcal{T}\subseteq\mathbb{R}^n \\ \dim(\mathcal{T})=n-k+1}} \max_{\substack{\mathbf{x}\in\mathcal{T} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

where the maximization and minimization are over subspaces $\mathcal{S}$ and $\mathcal{T}$ of $\mathbb{R}^n$.

In the textbook, Lemma 2.1.1 suggests that, in the previous example, for any $\mathbf{x} = \sum_i c_i \boldsymbol{\psi}_i$:

$$\mathbf{x}^T \mathbf{M} \mathbf{x} = \sum_{i=1}^{n} c_i^2 \mu_i$$

Again, $\boldsymbol{\psi}_i^T \boldsymbol{\psi}_j = 1$ when $i = j$ and 0 otherwise, also because $\mathbf{M}\boldsymbol{\psi}_i = \mu_i \boldsymbol{\psi}_i$,

$$\begin{aligned}
\mathbf{x}^T \mathbf{M} \mathbf{x} &= \Big(\sum_i c_i \boldsymbol{\psi}_i \Big)^T \mathbf{M} \Big(\sum_i c_i \boldsymbol{\psi}_i \Big) \\
&= \Big(\sum_i c_i \boldsymbol{\psi}_i \Big)^T \Big(\sum_i c_i \mu_i \boldsymbol{\psi}_i \Big) \\
&= \sum_{i,j} c_i^2 \mu_i \boldsymbol{\psi}_i^T \boldsymbol{\psi}_j \\
&= \sum_i c_i^2 \mu_i
\end{aligned}$$

Take a look again:

Theorem (2.0.1 Courant-Fischer Theorem)

Let $\mathbf{M}$ *be a symmetric matrix with eigenvalues* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. *Then,*

$$\mu_k = \max_{\substack{\mathcal{S} \subseteq \mathbb{R}^n \\ \dim(\mathcal{S}) = k}} \min_{\substack{\mathbf{x} \in \mathcal{S} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \min_{\substack{\mathcal{T} \subseteq \mathbb{R}^n \\ \dim(\mathcal{T}) = n-k+1}} \max_{\substack{\mathbf{x} \in \mathcal{T} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

where the maximization and minimization are over subspaces $\mathcal{S}$ *and* $\mathcal{T}$ *of* $\mathbb{R}^n$.

The Courant-Fischer Theorem: Proof VIII

We need the value of $\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$. In particular, we care about μ_k and subspace $\mathcal{S}$ where $\dim(\mathcal{S}) = k$. Also recall that we put $\{\mu_i\}_{i=1}^n$ in the **non-increasing** order.

$$\mathbf{x} = \sum_i^k c_i \boldsymbol{\psi}_i, \qquad c_i = \boldsymbol{\psi}_i^T \mathbf{x}$$

$$\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\sum_i^k c_i^2 \mu_i}{\sum_i^k c_i^2} \geq \frac{\sum_i^k c_i^2 \mu_k}{\sum_i^k c_i^2} = \mu_k$$

Therefore,

$$\min_{\substack{\mathbf{x}\in\mathcal{S} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \geq \mu_k$$

The Courant-Fischer Theorem: Proof IX

To prove the theorem, we also need to show that for all subspace $\mathcal{S} \subseteq \mathbb{R}^n$ where $\dim(\mathcal{S}) = k$,

$$\min_{\substack{\mathbf{x} \in \mathcal{S} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \leq \mu_k$$

For this part we bring up the subspace $\mathcal{T}$ of dimension $n - k + 1$, whose basis vectors are $\boldsymbol{\psi}_k, \ldots, \boldsymbol{\psi}_n$. Similarly, for $\mathbf{x} \in \mathcal{T}$, we have:

$$\max_{\substack{\mathbf{x} \in \mathcal{T} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \max_{\substack{\mathbf{x} \in \mathcal{T} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\sum_k^n c_i^2 \mu_i}{\sum_k^n c_i^2} \leq \frac{\sum_k^n c_i^2 \mu_k}{\sum_k^n c_i^2} = \mu_k$$

Every subspace $\mathcal{S}$ of dimension k has an intersection with $\mathcal{T}$ (dimension $n-k+1$), the intersection has dimension at least 1 ($(n-k+1)+k=n+1$).

$$\min_{\substack{\mathbf{x}\in\mathcal{S}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \min_{\substack{\mathbf{x}\in\mathcal{S}\cap\mathcal{T}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \max_{\substack{\mathbf{x}\in\mathcal{S}\cap\mathcal{T}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \max_{\substack{\mathbf{x}\in\mathcal{T}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \mu_k$$

The theorem is proved this way.

Counterexample in the non-Hermitian case

This example shows that when $\mathbf{M}$ is not symmetric, the properties is no longer guarantee to exist.

$$\mathbf{M} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \det(\lambda \mathbf{I} - \mathbf{M}) = \lambda^2 = 0$$

$$\frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \frac{x_1 x_2}{x_1^2 + x_2^2}$$

We can easily make it larger than 0, by, say, $\mathbf{x} = \mathbb{1}$. Then $\frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \frac{1}{2}$.

We prove the Spectral Theorem in a form that is almost identical to Courant-Fischer.

Main Steps:

- showing that the **Rayleigh quotient** and eigenvectors, eigenvalues have certain relation, starting from μ_1;
- use the conclusion in the first step to prove that a vector is an eigenvector, prove the Spectral Theorem by generalizing this characterization to **all** of the eigenvalues of $\mathbf{M}$

The **Rayleigh quotient** of a vector $\mathbf{x}$ with respect to a matrix $\mathbf{M}$ is defined to be:

$$\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

The Rayleigh quotient of an eigenvector is its corresponding eigenvalue: if $\mathbf{M}\boldsymbol{\psi} = \mu\boldsymbol{\psi}$, then (by default, $\boldsymbol{\psi} \neq \mathbf{0}$)

$$\frac{\boldsymbol{\psi}^T\mathbf{M}\boldsymbol{\psi}}{\boldsymbol{\psi}^T\boldsymbol{\psi}} = \frac{\boldsymbol{\psi}^T(\mathbf{M}\boldsymbol{\psi})}{\boldsymbol{\psi}^T\boldsymbol{\psi}} = \frac{\boldsymbol{\psi}^T(\mu\boldsymbol{\psi})}{\boldsymbol{\psi}^T\boldsymbol{\psi}} = \frac{\mu\boldsymbol{\psi}^T\boldsymbol{\psi}}{\boldsymbol{\psi}^T\boldsymbol{\psi}} = \mu$$

The first step is to prove the following theorem:

Theorem (2.2.1 (Rayleigh quotient and eigenvectors))

Let $\mathbf{M}$ *be a symmetric matrix and let vector* $\mathbf{x} \neq \mathbf{0}$ *maximize the* ***Rayleigh quotient*** *with respect to* $\mathbf{M}$*:*

$$\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

Then, $\mathbf{M}\mathbf{x} = \mu_1\mathbf{x}$*, where* μ_1 *is the largest eigenvalue of* $\mathbf{M}$*. Conversely, the minimum is achieved by eigenvectors of the smallest eigenvalue of* $\mathbf{M}$*.*

Observe that:

- the Rayleigh quotient is homogeneous (being homogeneous of degree k means:)

$$f(\alpha \mathbf{v}) = \alpha^k f(\mathbf{v})$$

- it suffices to consider unit vectors $\mathbf{x}$, the set of unit vectors is a closed and compact set

Rayleigh quotient's maximum is achieved, on the set of unit vectors.

Recall that: a function at its maximum and minimum has gradient $\mathbf{0}$ (zero vector).

The Spectral Theorem: Proof III

We can compute the gradient of the Rayleigh quotient.

$$\nabla \mathbf{x}^T\mathbf{x} = 2\mathbf{x} \qquad \nabla \mathbf{x}^T\mathbf{M}\mathbf{x} = 2\mathbf{M}\mathbf{x}$$

also recall the derivative rule:

$$(\frac{f}{g})' = \frac{gf' - fg'}{g^2}$$

$$\nabla\Big(\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\Big) = \frac{(\mathbf{x}^T\mathbf{x})(2\mathbf{M}\mathbf{x}) - (\mathbf{x}^T\mathbf{M}\mathbf{x})(2\mathbf{x})}{(\mathbf{x}^T\mathbf{x})^2}, \qquad \mathbf{x} \neq \mathbf{0}$$

when it is $\mathbf{0}$, $(\mathbf{x}^T\mathbf{x})\mathbf{M}\mathbf{x} = (\mathbf{x}^T\mathbf{M}\mathbf{x})\mathbf{x}$, $\mathbf{M}\mathbf{x} = \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\mathbf{x}$.

$$\mathbf{M}\mathbf{x} = \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\mathbf{x}$$

Recall that: *the Rayleigh quotient of an eigenvector is its corresponding eigenvalue.*

Also recall the definition of eigenvalues and eigenvectors.

The above equation holds iff $\mathbf{x}$ is an eigenvector of $\mathbf{M}$, with corresponding eigenvalue $\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$.

$\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$ has to be selected from the eigenvalues of $\mathbf{M}$.

Proved.

Theorem (2.2.2 (almost identical to the CF Theorem))

Let M be an n-dimensional real symmetric matrix. There exist numbers $\mu_1, \ldots, \mu_n$ and orthonormal vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ such that $\mathbf{M}\boldsymbol{\psi}_i = \mu_i \boldsymbol{\psi}_1$. Moreover,

$$\boldsymbol{\psi}_1 \in \arg \max_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{M} \mathbf{x}$$

and for $2 \leq i \leq n$,

$$\boldsymbol{\psi}_i \in \arg \max_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T \boldsymbol{\psi}_j = 0, j<i}} \mathbf{x}^T \mathbf{M} \mathbf{x},$$

similarly, $$\boldsymbol{\psi}_i \in \arg \min_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T \boldsymbol{\psi}_j = 0, j>i}} \mathbf{x}^T \mathbf{M} \mathbf{x}$$

The Spectral Theorem: Proof VI

To start with, we want to reduce to the case of positive definite matrices. In order to do that, we first modify $\mathbf{M}$ a bit.

$$\mu_n = \min_{\mathbf{x}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

we know μ_n exists from Theorem 2.2.1 we've just proved. Now we consider:

$$\widetilde{\mathbf{M}} = \mathbf{M} + (1 - \mu_n)\mathbf{I}$$

For $\forall \mathbf{x}$ such that $\|\mathbf{x}\| = 1$, we have:

$$\mathbf{x}^T \widetilde{\mathbf{M}} \mathbf{x} = \mathbf{x}^T \mathbf{M} \mathbf{x} + 1 - \mu_n = 1 + \left(\mathbf{x}^T \mathbf{M} \mathbf{x} - \min_{\mathbf{x}} \mathbf{x}^T \mathbf{M} \mathbf{x}\right) \geq 1$$

Therefore $\widetilde{\mathbf{M}}$ is positive definite.

Besides,

$$\widetilde{\mathbf{M}}\mathbf{x} = \mathbf{M}\mathbf{x} + (1 - \mu_n)\mathbf{x}$$

For $\forall \boldsymbol{\psi}, \mu$ where $\mathbf{M}\boldsymbol{\psi} = \mu\boldsymbol{\psi}$,

$$\widetilde{\mathbf{M}}\boldsymbol{\psi} = \mathbf{M}\boldsymbol{\psi} + (1 - \mu_n)\boldsymbol{\psi} = (\mu + 1 - \mu_n)\boldsymbol{\psi}$$

thus $\widetilde{\mathbf{M}}$ and $\mathbf{M}$ have the same eigenvectors.

Thus it suffices to prove the theorem for positive definite matrices. In other words, we treat $\mathbf{M}$ as if it is positive definite.

We proceed by induction on k. We construct $\boldsymbol{\psi}_{k+1}$ base on eigenvalues $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$ satisfying:

$$\boldsymbol{\psi}_i \in \arg \max_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T \boldsymbol{\psi}_j = 0, j<i}} \mathbf{x}^T \mathbf{M} \mathbf{x}$$

And define:

$$\mathbf{M}_k = \mathbf{M} - \sum_{i=1}^{k} \mu_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^T$$

For $j \leq k$ we have (because all the previous eigenvectors are all orthogonal to each other):

$$\mathbf{M}_k \boldsymbol{\psi}_j = \mathbf{M} \boldsymbol{\psi}_j - \sum_{i=1}^{k} \mu_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^T \boldsymbol{\psi}_j = \mu_j \boldsymbol{\psi}_j - \mu_j \boldsymbol{\psi}_j = \mathbf{0}$$

Hence, for vector $\mathbf{x}$ that are orthogonal to $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$,

$$\mathbf{M}\mathbf{x} = \mathbf{M}_k\mathbf{x} + \sum_{i=1}^{k} \mu_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^T \mathbf{x} = \mathbf{M}_k\mathbf{x}, \qquad \mathbf{x}^T\mathbf{M}\mathbf{x} = \mathbf{x}^T\mathbf{M}_k\mathbf{x}$$

and,

$$\arg \max_{\substack{\|\mathbf{x}\|=1 \\ \mathbf{x}^T\boldsymbol{\psi}_j=0, j<i}} \mathbf{x}^T\mathbf{M}\mathbf{x} \leq \arg \max_{\|\mathbf{x}\|=1} \mathbf{x}^T\mathbf{M}_k\mathbf{x}$$

For convenience we define $\mathbf{y} = \arg\max_{\|\mathbf{x}\|=1} \mathbf{x}^T\mathbf{M}_k\mathbf{x}$. From Theorem 2.2.1 we know that $\mathbf{y}$ is an eigenvector of $\mathbf{M}_k$. Let's say that the corresponding eigenvalue is μ. $\mathbf{M}_k$ and $\mathbf{M}$ have the same eigenvectors, thus $\mathbf{y}$ is an eigenvector of $\mathbf{M}$.

Now we will prove that we can set $\boldsymbol{\psi}_{k+1} = \mathbf{y}$ and $\mu_{k+1} = \mu$.

We prove it by showing $\mathbf{y}$ must be orthogonal to each $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$.

$$\widetilde{\mathbf{y}} = \mathbf{y} - \sum_{i=1}^{k} \boldsymbol{\psi}_i (\boldsymbol{\psi}_i^T \mathbf{y})$$

is the projection of $\mathbf{y}$ orthogonal to $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$. Since $\mathbf{M}_k \boldsymbol{\psi}_j = \mathbf{0}$ for $j \leq k$,

$$\widetilde{\mathbf{y}}^T \mathbf{M}_k \widetilde{\mathbf{y}} = \mathbf{y}^T \mathbf{M}_k \mathbf{y} = \mathbf{y}^T \mathbf{M} \mathbf{y}$$

If $\mathbf{y}$ is not orthogonal to $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$, some $\boldsymbol{\psi}_i^T \mathbf{y} \neq \mathbf{0}$, then $\|\widetilde{\mathbf{y}}\| < \|\mathbf{y}\|$. Because we assume positive definite of $\mathbf{M}$, there comes a contradiction.

$$\widetilde{\mathbf{y}}^T \mathbf{M}_k \widetilde{\mathbf{y}} = \widetilde{\mathbf{y}}^T \mathbf{M} \widetilde{\mathbf{y}} > 0$$

and also note that $\|\widetilde{\mathbf{y}}\| < \|\mathbf{y}\|$ (previous conclusion), for normalized $\widetilde{\mathbf{y}}$, $\widehat{\mathbf{y}} = \widetilde{\mathbf{y}}/\|\widetilde{\mathbf{y}}\|$, and $\mathbf{y}$ was an **unit** vector,

$$\widehat{\mathbf{y}}^T \mathbf{M} \widehat{\mathbf{y}} = \widehat{\mathbf{y}}^T \mathbf{M}_k \widehat{\mathbf{y}} = \frac{\widetilde{\mathbf{y}}^T \mathbf{M}_k \widetilde{\mathbf{y}}}{\|\widetilde{\mathbf{y}}\|^2}$$

$$> \frac{\widetilde{\mathbf{y}}^T \mathbf{M}_k \widetilde{\mathbf{y}}}{\|\mathbf{y}\|^2} = \widetilde{\mathbf{y}}^T \mathbf{M}_k \widetilde{\mathbf{y}} = \mathbf{y}^T \mathbf{M}_k \mathbf{y} = \mathbf{y}^T \mathbf{M} \mathbf{y}$$

There's a conflict with $\mathbf{y}$'s definition:

$$\mathbf{y} = \arg \max_{\|\mathbf{x}\|=1} \mathbf{x}^T \mathbf{M}_k \mathbf{x}$$

$\therefore \mathbf{y}$ must be orthogonal to $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$.

The Laplacian and Graph Drawing

Overview of Chap 3

Chapter 3 shows that Laplacian should reveal a lot about the structure of graphs, although not always guaranteed to work.

It mentions Hall's (Kenneth M. Hall) work a lot of times:
An r-dimensional quadratic placement algorithm

The idea of drawing graphs using eigenvectors demonstrated in Section 1.5.1 was suggested by Hall in 1970.

Recall that weighted undirected graph $G = (V, E, w)$, with positive weigh $w : E \to \mathbb{R}^+$, is defined this way:

$$\mathbf{L}_G \stackrel{\text{def}}{=} \mathbf{D}_G - \mathbf{M}_G, \qquad \mathbf{D}_G = \sum_b w_{a,b}$$

where $\mathbf{D}_G$ is the diffusion matrix, $\mathbf{M}_G$ is the adjacency matrix.

Given a vector $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \sum_{(a,b) \in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

Hall's Idea on Graph Drawing

Hall's idea on graph drawing suggests that we choose the first coordinates of the n vertices as $\mathbf{x} \in \mathbb{R}^n$ that minimizes:

$$\mathbf{x}^T \mathbf{L} \mathbf{x} = \sum_{(a,b) \in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

To avoid degenerating to $\mathbf{0}$, we have restriction:

$$\|\mathbf{x}\|^2 = \sum_{a \in V} \mathbf{x}(a)^2 = 1$$

To avoid degenerating to $\mathbb{1}/\sqrt{n}$, Hall suggested another constraint:

$$\mathbb{1}^T \mathbf{x} = \sum_{a \in V} \mathbf{x}(a) = 0$$

When there are multiple sets of coordinates, say $\mathbf{x}$ and $\mathbf{y}$; we require $\mathbf{x}^T \mathbf{y} = 0$, to avoid cases such as $\mathbf{x} = \mathbf{y} = \boldsymbol{\psi}_2$.

Hall's Idea on Graph Drawing

We will minimize the sum of the squares of the lengths of the edges in the embedding. e.g. 2-D case:

$$\sum_{(a,b)\in E} \left\| \begin{bmatrix} \mathbf{x}(a) \\ \mathbf{y}(a) \end{bmatrix} - \begin{bmatrix} \mathbf{x}(b) \\ \mathbf{y}(b) \end{bmatrix} \right\|^2$$

$$= \sum_{(a,b)\in E} (\mathbf{x}(a) - \mathbf{x}(b))^2 + (\mathbf{y}(a) - \mathbf{y}(b))^2$$

$$= \mathbf{x}^T \mathbf{L} \mathbf{x} + \mathbf{y}^T \mathbf{L} \mathbf{y}$$

is the objective we want to minimize.

Properties to Prove

Here are some of the very interesting properties of a graph that we would like to prove.

- If and only if the graph is connected, there is only one eigenvalue of its Laplacian equals to zero.
- When mapping each vertex to a set of coordinates, choosing the coordinates to be the eigenvectors of the graph Laplacian is optimal.

Lemma

Let $G = (V, E)$ be a graph, and let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of its Laplacian matrix, $\mathbf{L}$. Then, $\lambda_2 > 0$ if and only if G is connected.

Property #1: Proof I

First of all, there exists eigenvalue $\mathbf{0}$, because the all-one vector $\mathbb{1}$ satisfies:

$$\mathbf{L}\mathbb{1} = \mathbf{0}$$

To prove, if we view the Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{M}$ as an operator ($\mathbf{D} = \sum_{(a,b)\in E} w_{a,b}$), for each $\mathbf{x}$ we have its a^{th} entry of $\mathbf{Lx}$ being:

$$(\mathbf{Lx})(a) = d(a)\mathbf{x}(a) - \sum_{(a,b)\in E} w_{a,b}\mathbf{x}(b) = \sum_{(a,b)\in E} w_{a,b}\big(\mathbf{x}(a) - \mathbf{x}(b)\big)$$

It infers that $\mathbb{1}$ is an eigenvector corresponds to eigenvalue 0. Therefore, $\lambda_1 = 0$.

Property #1: Proof II

Next, we show that $\lambda_2 = 0$ if G is disconnected.

If G is disconnected, then we can split it into two graphs G_1 and G_2. Because we can safely reorder the vertices of a graph, we can have:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{G_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_{G_2} \end{bmatrix}$$

It has at least 2 orthogonal eigenvectors of eigenvalue zero:

$$\begin{bmatrix} \mathbf{0} \\ \mathbb{1} \end{bmatrix}, \text{ and } \begin{bmatrix} \mathbb{1} \\ \mathbf{0} \end{bmatrix}$$

Property #1: Proof III

On the other hand, for a eigenvector $\boldsymbol{\psi}$ of eigenvalue 0, $\mathbf{L}\boldsymbol{\psi} = \mathbf{0}$,

$$\boldsymbol{\psi}^T \mathbf{L} \boldsymbol{\psi} = \sum_{(a,b) \in E} w_{a,b} \big(\boldsymbol{\psi}(a) - \boldsymbol{\psi}(b)\big)^2 = 0$$

For every pair of vertices (a, b) connected by an edge, we have $\boldsymbol{\psi}(a) = \boldsymbol{\psi}(b)$. In a connected graph, all vertices are directly or indirectly connected, and thus $\boldsymbol{\psi}$ must be a constant vector.

Contradiction found.

Therefore, G must be disconnected when $\lambda_2 = 0$.

Theorem (3.2.1)

Let $\mathbf{L}$ *be a Laplacian matrix and let* $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *be orthonormal* [a] *vectors that are all orthogonal to* $\mathbb{1}$*. Then*

$$\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{L} \mathbf{x}_i \geq \sum_{i=2}^{k+1} \lambda_i$$

and this inequality is tight only when $\mathbf{x}^T \boldsymbol{\psi}_j = 0$ *for all* j *such that* $\lambda_j \geq \lambda_{k+1}$*.* λ_i *are the eigenvalues, the graph* G *is an undirected connected graph.*

[a] orthonormal = both orthogonal and normalized

Property #2: Proof I

We can order λ such that:

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

As is proved before, $\lambda_1 = 0$ and because G is connected, $\boldsymbol{\psi}_1$ is a constant vector.

Let $\mathbf{x}_{k+1} \ldots \mathbf{x}_n$ be vectors such that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is an orthogonal **basis**. It is done by choosing $\mathbf{x}_{k+1} \ldots \mathbf{x}_n$ to be an orthogonal basis of the space orthogonal to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Because they are orthogonal basis, (think of orthogonal matrix)

$$\sum_{j=1}^{n} (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = \sum_{j=1}^{n} (\mathbf{x}_i^T \boldsymbol{\psi}_j)^2 = 1, \qquad i = 1, 2, \ldots, n$$

Property #2: Proof II

Because of that $\boldsymbol{\psi}_1^T \mathbf{x}_i \propto \mathbb{1}^T \mathbf{x}_i = 0$, and that $\sum_{j=1}^n (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 1$,

$$\sum_{j=2}^{n} (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 1$$

Previously, $\mathbf{x}^T \mathbf{M} \mathbf{x} = \sum_i c_i^2 \mu_i$, $c_i = \boldsymbol{\psi}_i^T \mathbf{x}$, $\mathbf{x} = \sum_i c_i \boldsymbol{\psi}_i$. Here,

$$\begin{aligned}
\mathbf{x}_i^T \mathbf{L} \mathbf{x}_i &= \sum_{j=2}^{n} \lambda_j (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = \lambda_{k+1} + \sum_{j=2}^{n} (\lambda_j - \lambda_{k+1}) (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \\
&\geq \lambda_{k+1} + \sum_{j=2}^{k+1} (\lambda_j - \lambda_{k+1}) (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2
\end{aligned}$$

It is tight only when $\boldsymbol{\psi}_j^T \mathbf{x}_i = 0$ for $\lambda_j \geq \lambda_{k+1}$.

Property #2: Proof III

$$\lambda_{k+1} + \sum_{j=2}^{n} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \geq \lambda_{k+1} + \sum_{j=2}^{k+1} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2$$

Quick proof of when the above inequality is tight:

$$\lambda_{k+1} + \sum_{j=2}^{n} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = \lambda_{k+1} + \sum_{j=2}^{k+1} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2$$

$$\sum_{j=k+2}^{n} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 0$$

That is $\boldsymbol{\psi}_j^T \mathbf{x}_i = 0$ for $j > k+1$. When $j > k+1$, $\lambda_j \geq \lambda_{k+1}$.

Property #2: Proof IV

To prove the Theorem 3.2.1, we sum up over i:

$$\begin{aligned}\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{L} \mathbf{x}_i &\geq k\lambda_{k+1} + \sum_{i=1}^{k}\sum_{j=2}^{k+1}(\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \\ &= k\lambda_{k+1} + \sum_{j=2}^{k+1}(\lambda_j - \lambda_{k+1}) \sum_{i=1}^{k}(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \\ &\geq k\lambda_{k+1} + \sum_{j=2}^{k+1}(\lambda_j - \lambda_{k+1}) = \sum_{j=2}^{k+1} \lambda_j\end{aligned}$$

because: $\lambda_j - \lambda_{k+1} \leq 0$, and, $\sum_{i=1}^{k}(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \leq \sum_{i=1}^{n}(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 1$.

The two properties are saying that:

- Eigenvalues of graphs Laplacian can easily reveal the graph's connectivity. The amount of eigenvalue 0 is exactly the amount of independent components in a graph. For a connected graph, only $\lambda_1 = 0$, $\lambda_2 > 0$. If the graph is disconnected, $\lambda_2 = 0$. If the graph contains 3 disconnected subgraphs, $\lambda_3 = 0$. etc.
- When visualizing a graph, using its eigenvectors ($\boldsymbol{\psi}_1$ excluded) as vertices' coordinates, will be an optimal choice.

Intuitive Understanding of Graph Laplacian

In this part I record the vivid example Prof. Cho provided during our reading group.

This is a very nice example that helps us understand the (physical) meaning of a graph's Laplacian better.

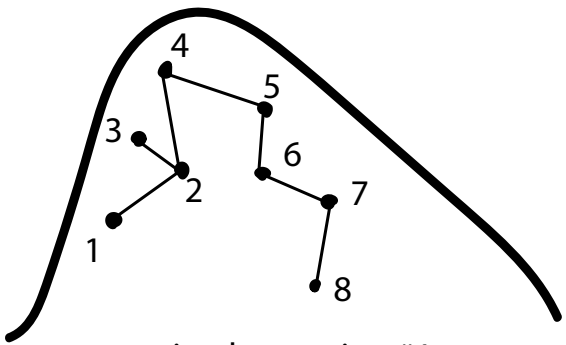
In other words, this is an **intuitive** explanation of what we've learned from the first three chapters.

Imagine that we are going to estimate the (absolute) height $\mathbf{h} \in \mathbb{R}^n$ of some selected points on a mountain. Let's say that there are n points to estimate in total.

Climbing up and down in the mountain, we have no clue what is its exact height, but we know k relative heights (e.g. relative height between vertices 1 and 2 is $\Delta_{1,2} = h_1 - h_2$). We denote the record of each relative height (the **edges**) as $\mathbf{m} \in \mathbb{R}^k$.

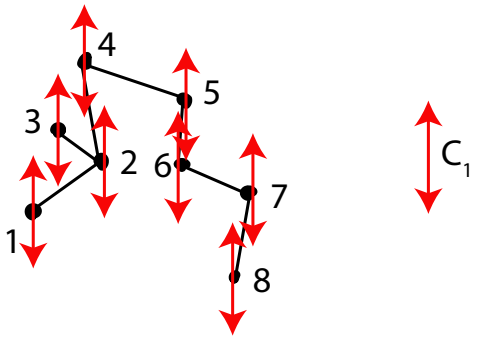
We denote the starting and ending of the nodes by an Incidence Matrix $\mathbf{I}_G \in \mathbb{R}^{k \times n}$.

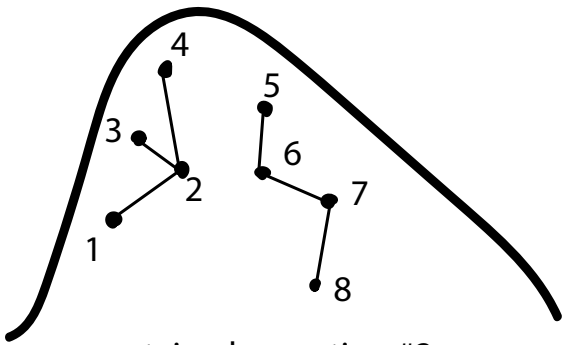
Illustration



mountain observation #1

#edge: k = 7
#node: n = 8
degree of freedom: 1





mountain observation #2

#edge: k = 6
#node: n = 8
degree of freedom: 2

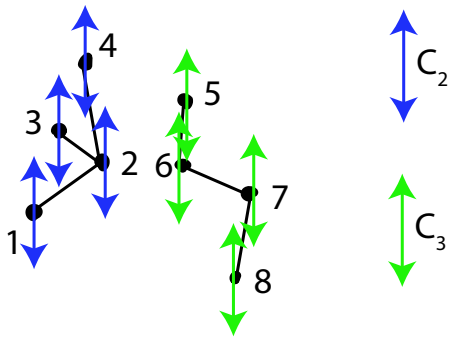


Figure: Illustration of the examples Prof. Cho brought up.

Illustration: mountain observation #1

$$\mathbf{m} = \mathbf{I}_G \mathbf{h}$$

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \end{bmatrix}$$

Illustration: mountain observation #2

$$\mathbf{m} = \mathbf{I}_G \mathbf{h}$$

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \end{bmatrix}$$

The problem is formally defined this way:

$$\mathbf{m} = \mathbf{I}_G \mathbf{h}$$

Knowing $\mathbf{m}, \mathbf{I}_G$, solving $\mathbf{h}$.

It is solved by minimizing over $\mathbf{h}$:

$$\|\mathbf{I}_G \mathbf{h} - \mathbf{m}\|^2$$

Recall that for any $\mathbf{Ax} = \mathbf{b}$ the solution is $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$, since $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$.

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

In this case, it means that:

$$\mathbf{I}_G^T\mathbf{I}_G\mathbf{h} = \mathbf{I}_G^T\mathbf{m}$$

Recall that the graph Laplacian $\mathbf{L}_G = \mathbf{I}_G^T\mathbf{I}_G$, therefore we have:

$$\mathbf{L}_G\mathbf{h} = \mathbf{I}_G^T\mathbf{m}$$

Just for convenience, we introduce a known value $\mathbf{b} = \mathbf{I}_G^T\mathbf{m} \in \mathbb{R}^n$.

$$\mathbf{L}_G\mathbf{h} = \mathbf{b}$$

Now we consider the graph itself:

- #1: The graph is connected, but we will never know the **exact** absolute height of the mountain. Because whatever **h** value we result in, since we only know the nodes' relative height, it makes sense if we move the entire graph up and down alone the vertical direction. That is, after adding a constant value C_1 to every entry in **h**, we still result in a valid solution.
- #2: Similarly, this time we have 2 separate subgraphs, therefore, each subgraph could be moved up and down independently. Let's say that nodes in the two subgraphs can be shifted alone the vertical direction by C_2 and C_3 distance respectively.

This is why we say that the degree of freedom in case #1 is 1, and that in case #2 is 2.

Eigenvalue Zero: λ_1 in case #1

Here, consider case #1, since we all know that we can add a constant vector to a solution $\mathbf{h}$ and the resulting vector is still a valid solution, we have:

$$\mathbf{L}_G \mathbf{h} = \mathbf{b}$$

$$\mathbf{L}_G \left(\mathbf{h} + C_1 \mathbb{1} \right) = \mathbf{b}$$

$$\therefore \; C_1 \mathbb{1} = \mathbf{0} = 0 \times C_1 \mathbb{1}$$

Therefore, It has an eigenvalue 0 with any constant vector being its corresponding eigenvector $\boldsymbol{\psi}_1 = C_1 \mathbb{1}$ where $C_1 \in \mathbb{R}$. $\lambda_1 = 0$.

Eigenvalue Zero: λ_2 in case #2

In case #2, we denote the two subgraphs as A and B respectively. Where we use $\mathbb{1}_A \in \mathbb{R}^n$ to denote a vector indicating whether or not a vertex is included in subgraph A (1 for yes, 0 for no). $\mathbb{1}_B \in \mathbb{R}^n$ is defined in the same way, but it is for subgraph B.

$$\mathbb{1}_A = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbb{1}_B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\mathbf{L}_G\mathbf{h} = \mathbf{b}$$

$$\mathbf{L}_G\left(\mathbf{h} + C_2\mathbb{1}_A\right) = \mathbf{b}$$

$$\mathbf{L}_G\left(\mathbf{h} + C_3\mathbb{1}_B\right) = \mathbf{b}$$

$$\therefore C_2\mathbb{1}_A = \mathbf{0} = 0 \times C_2\mathbb{1}$$

$$C_3\mathbb{1}_B = \mathbf{0} = 0 \times C_3\mathbb{1}$$

Therefore, $C_2\mathbb{1}_A$, $C_3\mathbb{1}_B$ are both eigenvectors of eigenvalue equals to 0, $C_2, C_3 \in \mathbb{R}$. Thus there must be $\lambda_1 = \lambda_2 = 0$.

We realize that the degree of freedom is directly reflected as how many eigenvalues (of the graph Laplacian) are 0.

Prof. Cho also shared some results of plotting a circuit graph ($(i, i+1)$ linked and also $(1, n)$).

There, he shows that we can run Python examples. Some of the very useful tools are built-in functions in numpy (np) and matplotlib (plt).

Useful Library Functions

```
1 np.linalg.eigh(...)
2 plt.plot(...)
```

The results generally agrees with our previous theories. e.g. Setting number of nodes $n = 1000$, plot the $2 - d$ graph by using coordinates: $\boldsymbol{\psi}_2 = V[:, 998]$ and $\boldsymbol{\psi}_3 = V[:, 997]$, what we draw ends up in an oval shape.

Example

```
plt.plot(V[:,998], V[:,997])
```

Moreover, in this case we have $\lambda_1 = 0$, $\lambda_2 = \lambda_3$, $\lambda_4 = \lambda_5$, $\ldots$; $\boldsymbol{\psi}_{2i}$ and $\boldsymbol{\psi}_{2i+1}$ $(i = 1, 2, \ldots, \text{floor}(n/2))$ correspond to the sine and cosine under the same frequency respectively.

Also observed that $\lambda_2 : \lambda_4 \approx 2 : 3$. Note that in code examples like this, $\lambda_1 \approx 0$, but aren't likely to be exactly 0, could be at e.g. e^{-15} level.