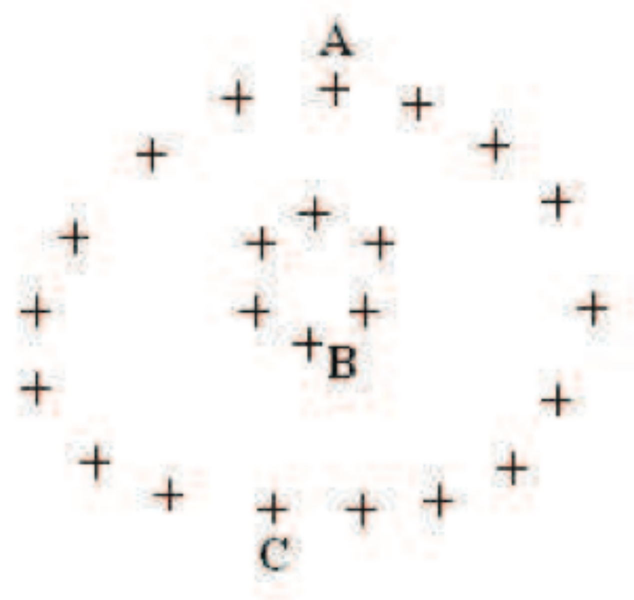
Mathematics of Data
II
Diffusion Geometry

姚 远
北京大学数学科学学院
2011.7.12

Data Distances



- We look for distance function such that
  - dist(A,C) is small
  - dist(A,B) is large
- Geodesic distance is one candidate, but hard to compute and sensitive to noise
- Any other distance with such properties but robust to stochastic noise?

Data Graph

- Given n points x_i, i=1,…,n, as vertices in V
- Similarity weight between x_i and x_j is $w_{ij}=w_{ji}$, e.g.

$$w_{ij} = k\left(\frac{\left\|x_i - x_j\right\|_{R^p}}{\sqrt{\varepsilon}}\right), \qquad k(t) = e^{-t^2/2}$$

- Undirected weighted graph $G(V,E,W)$, W=(w_{ij})

Random Walk on Graphs

- Degree $d_i = \Sigma_k w_{ik}$, D = diag(d_i)
- Random walk on *G(V,E,W)*
  - Transition probability $P = D^{-1} W$ where $p_{ij} = w_{ij}/d_i$
  - Stationary distribution $\pi_i \sim d_i$
  - Irreducible (G is connected)
  - Reversible $w_{ij} = w_{ji}$ ⇒ $\pi_i p_{ij} = \pi_j p_{ji}$

Symmetric Kernel

- $P = D^{-1}W$ is similar to $S = D^{-1/2}WD^{-1/2}$, as $P = D^{-1/2}SD^{1/2}$
- S is real symmetric, whence eigen-decomposition

$$S = V\Lambda V^T, \qquad \Lambda = diag(\lambda_i \in R)$$

$$\Rightarrow \quad P = D^{-1/2}V\Lambda V^T D^{1/2} = \Phi\Lambda\Psi^T, \qquad \Phi = D^{-1/2}V, \qquad \Psi = D^{1/2}V$$

Spectrum of P

- Eigenvalues of S and P are the same, so

$$|\lambda_i| \leq 1$$

- Φ and Ψ are right and left eigenvector matrix of P, respectively, $\Phi^T\Psi = V^TV = I$
- In particular, P 1 = 1, whence

$$\phi_1(i) = 1, \quad \psi_1(i) = \frac{d_i}{\sum_i d_i} = \pi_i$$

Diffusion Map

- Let λ_i be sorted by

$$1 = \lambda_1 \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$$

- Diffusion map of x_i is defined via right eigenvectors

$$\Phi_t(x_i) = \begin{pmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_n^t \phi_n(i) \end{pmatrix} \in R^n$$

Dimensionality Reduction

- $\lambda_1 = 1$ and $\phi_1 = 1$, so it does not distinguish points
- Threshold by δ, for those

$$\left|\lambda_i^t\right| \geq 1-\delta, \qquad i = 1,\ldots,m,$$

$$\left|\lambda_k^t\right| < 1-\delta, \qquad k > m$$

- Define

$$\Phi_t^\delta\left(x_i\right) = \begin{pmatrix} \lambda_2^t \phi_2(i) \\ \lambda_3^t \phi_3(i) \\ \vdots \\ \lambda_m^t \phi_m(i) \end{pmatrix} \in R^{m-1}$$

Diffusion Distance

- Define the diffusion distance between points at scale *t*

$$D_t(x_i, x_j) := \left\| \Phi_t(x_i) - \Phi_t(x_j) \right\|_{l^2} \cong \sum_{k=2}^{m} \lambda_k^t (\phi_k(x_i) - \phi_k(x_j))^2$$

- This is exactly the weighted 2-distance between diffusion profiles

$$D_t(x_i, x_j) := \left\| P_{i*}^t - P_{j*}^t \right\|_{l^2(1/d)} = \sum_{k=2}^{m} \frac{(P_{ik}^t - P_{jk}^t)^2}{d_k}$$

Lumpability of Markov Chains

- Let P be the transition matrix of a Markov chain defined on n states S={1,…,n}.
- $\Gamma=\{S_1,\dots,S_k\}$ is a partition of S into k macrostates.
- Sequences $\{x_0,\dots,x_t,\dots\}$ generated by P, i.e.

$$\text{Prob}(x_t=j\ ;\ x_{t-1}=i)= P_{ij}$$

- Induced dynamics: relabel x_t by y_t from corresponding states in partition Γ
- [Kemeny-Snell'76] P is called *lumpable* if

$$\text{Prob}(y_t=k_0;\ y_{t-1}=k_1,\ \dots,y_{t-m}=k_m) = \text{Prob}(y_t=k_0;\ y_{t-1}=k_1)$$

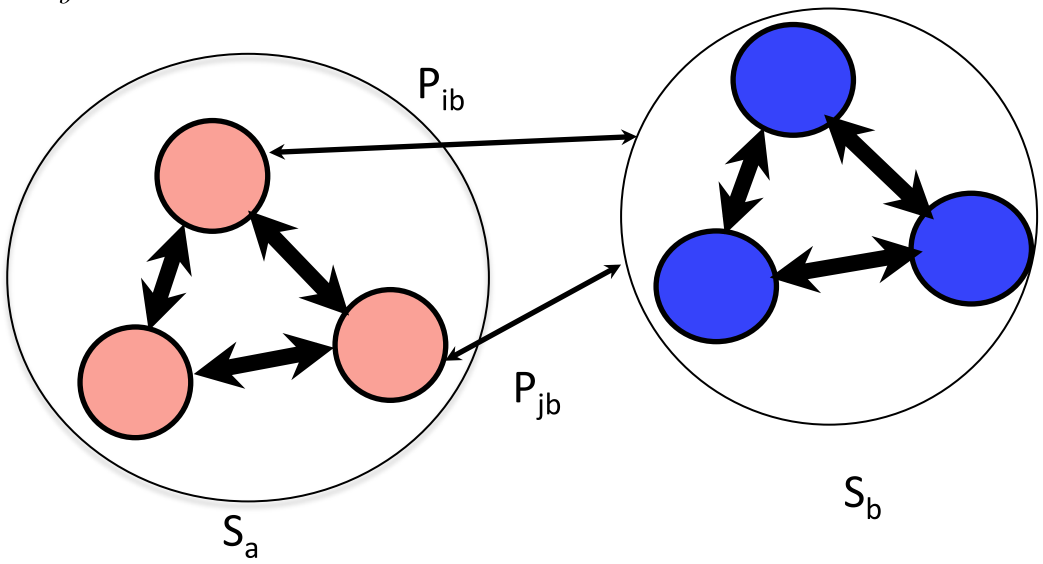
i.e. the induced dynamics is Markovian.

A Necessary and Sufficient Condition for Lumpability

- [Kemeny-Snell'76] P is *lumpable* w.r.t. partition $\Gamma=\{S_1,\ldots,S_k\}$ iff for any s, t chosen from P, and for any i, j lying in S_a, the following holds

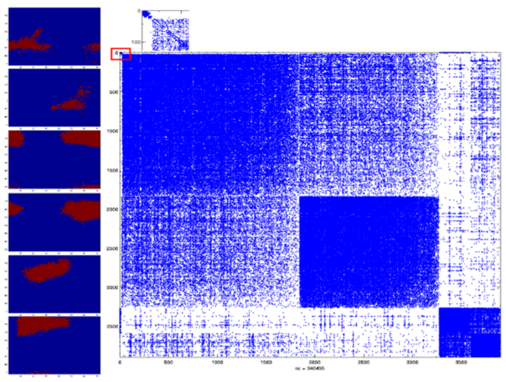
$$P_{ib} = P_{jb}$$

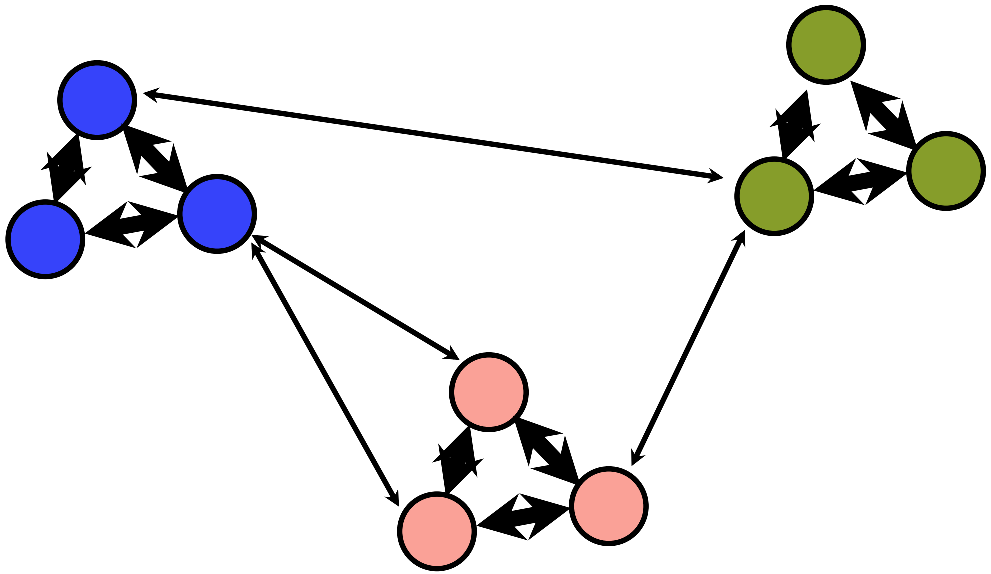
where $$P_{ib} = \sum_{k \in S_b} P_{ik}$$



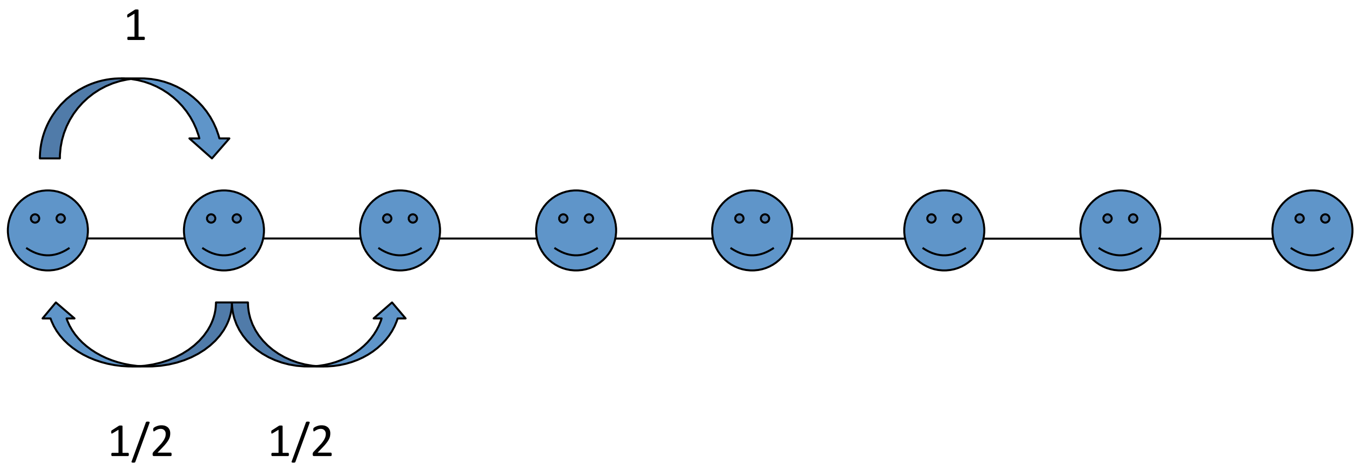
Spectral Theory of Lumpability

- [Meila-Shi 2001] P is *lumpable w.r.t. P* iff P has k independent piece-wise constant right eigenvectors in the span of characteristic functions of $\Gamma=\{S_1, \ldots, S_k\}$.
- Special case: If P is block diagonal, i.e. uncoupled Markov chain, then P is lumpable with piece-wise constant right eigenvectors associated with multiple eigenvalue 1.
- [e.g. Belkin-Shi-Yu 2007] If P is nearly block diagonal, then there are top-k eigenvectors which fix signs within the block.



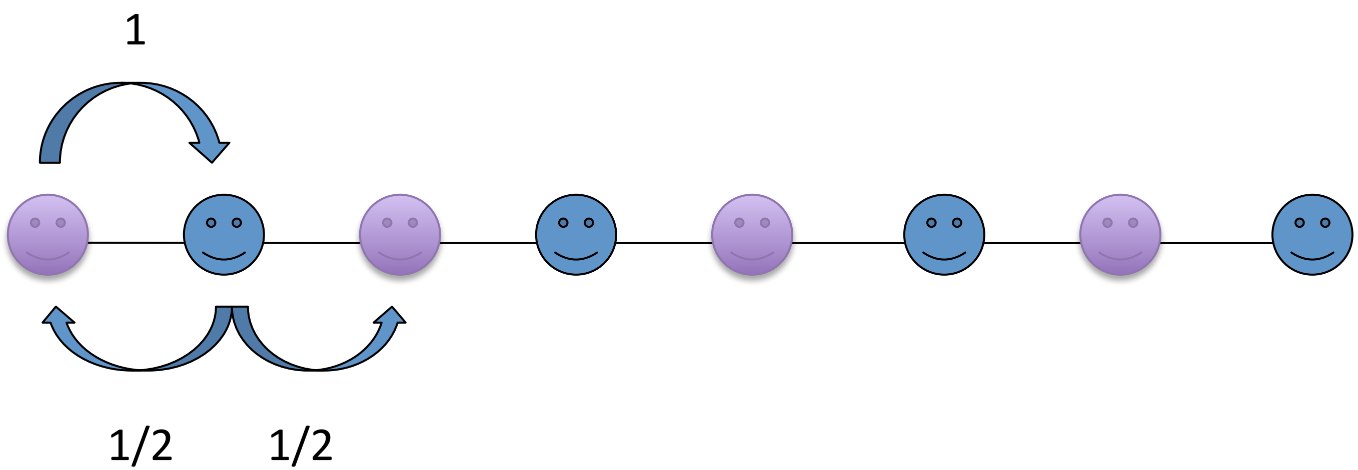


Example I



- Consider 2n nodes on a linear chain
- Markov Chain: a node will jump to its neighbors with equal probability
  - $P(i, i-1) = P(i, i+1) = ½$, for $2n>i>1$
  - $P(1,2) = P(2n,2n-1) = 1$

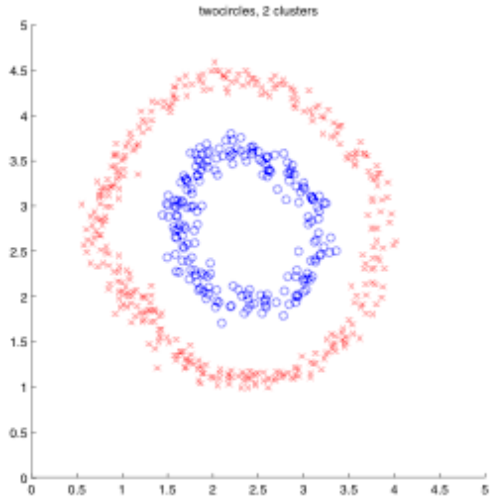
Example I



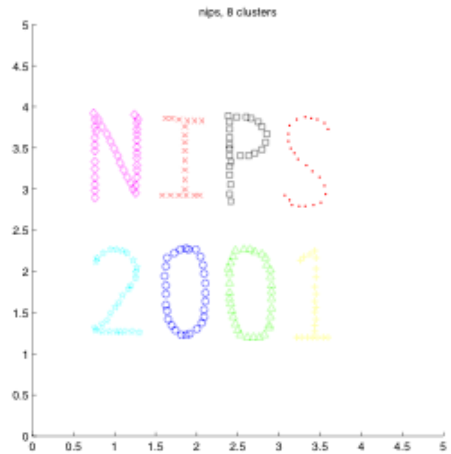
- P is lumpable w.r.t. $\Gamma^*=(S_{even},S_{odd})$
  - S_{even}: even nodes
  - S_{odd}: odd nodes
- Γ^* corresponds to eigenvector with eigenvalue -1

Spectral Clustering Algorithm

- Typical spectral algorithm to find lumpable states in nearly uncoupled systems [Ng-Jordan-Weiss NIPS'01]:
    1) Find top k right eigenvectors of P where a large spectral gap occurs, $v_1,\dots,v_k$
    2) Embed the data into R^k by those eigenvectors
    3) Use k-means (or alternatives) to find k clusters in R^k



(e)

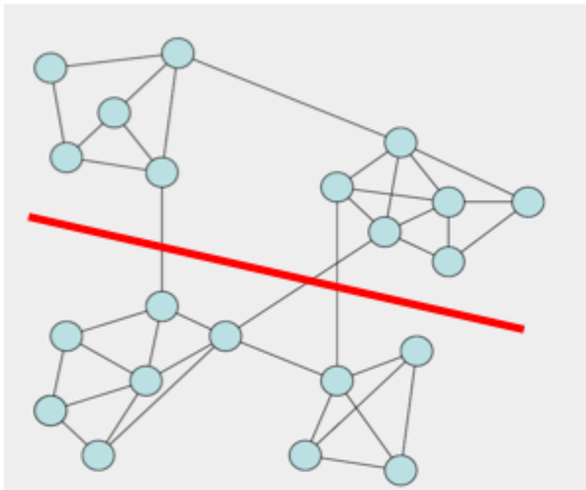


(a)

Graph Partition Problem

- goal: find a cut with the smallest Cheeger ratio (conductance)
  - For $S \subset V$, volume of S: $vol(S) = \sum_{v \in S} d_v$
  - $\partial S = \{(u,v) \in E : u \in S \& v \in S\}$
  - Cheeger ratio of S, $h(S) = \frac{|\partial S|}{\min\{vol(S), vol(G) - vol(S)\}}$

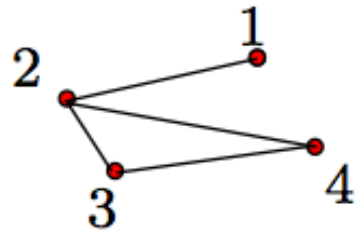
- applications
  - clustering
  - segmentation
  - task partitioning for parallel processing
  - a preprocessing step to divide-and-conquer algorithms



Graph Laplacian Operator

- given an undirected graph G=(V, E),
  - Adjacency matrix A:

$$A(u,v) = \left\{ \begin{array}{ll} 1 & \text{if } u \sim v \\ 0 & \text{o.w.} \end{array} \right\}$$



  - Diagonal degree matrix $D = diag(d_{v_1}, \cdots, d_{v_n})$
  - Graph Laplace Operator $L = D^{-1}(D - A)$
  - Tranistion probability matrix $W = D^{-1}A = I - L$,
  - $Wv = \lambda v$ implies $Lv = (1 - \lambda)v$
  - 1 is the largest eigenvalue for W; 0 is the smallest eigenvalue for L.

Graph Partition Problem

- Rayleigh quotient $R(f) = \frac{\sum_{u \sim v} (f(u) - f(v))^2}{\sum_u f^2(u) d_u}$ for $f \neq 0$
  - find a boolean function f minimizing $R(f)$ $\Leftarrow$ NP-complete
  - RELAXATION: find a real valued function f minimizing $R(f)$
  - $R(f) = \frac{<f,(D-A)f>}{<f,Df>}$
  - $\lambda_1 = \inf_f R(f) \;\Rightarrow\; \lambda_1$ and f are the first nonzero eigenvalue and eigenvector of L.

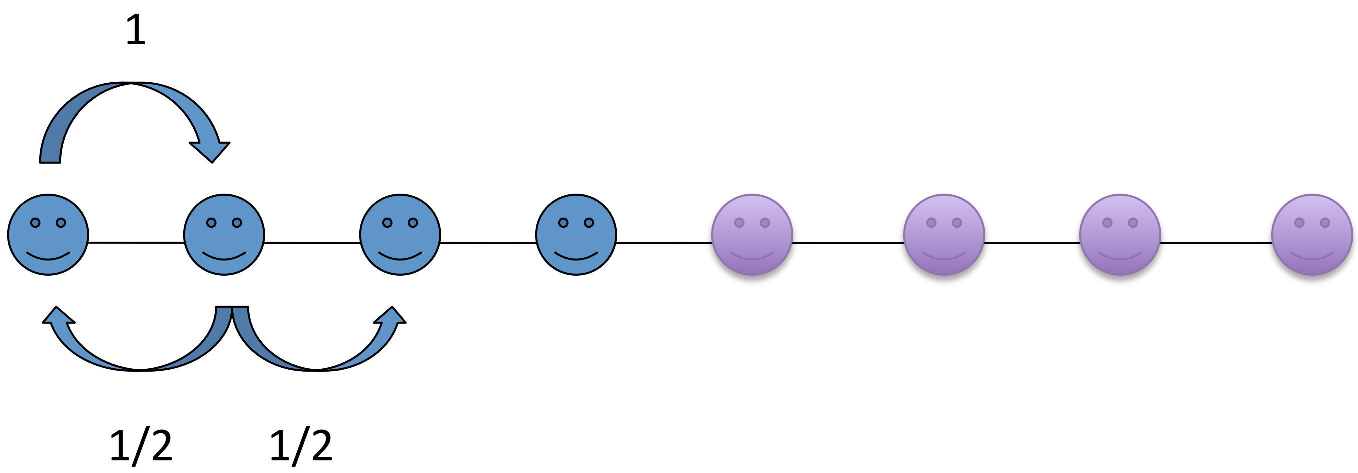
How good is this relaxation? Cheeger inequality

Cheeger Inequality

$$2h_G \geq \lambda_1 \geq \frac{h_f^2}{2} \geq \frac{h_G^2}{2}.$$

- f is the eigenvector of L corresponding to λ_1
- h_G is the smallest conductance (Cheeger ratio) of graph G
- h_f: the minimum Cheeger ratio determinded by a sweep of f
  - order the vertices: $f(v_1) \geq f(v_2) \geq \cdots \geq f(v_n)$.
  - $S_i = \{v_1, \cdots, v_i\}$
  - $h_f = \min_i h_{S_i}$
- find a partition whose conductance is within $2\sqrt{h_G}$

Example I



- One graph min-cut given by second largest right eigenvector of T
- n=8,
  - v_2=[0.4714 0.4247 0.2939 0.1049 -0.1049 -0.2939 -0.4247 -0.4714]
  - Eigenvalue is 0.9010

Connections to Manifold Learning

Given $x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$,
Find $y_1, \ldots, y_n \in \mathbb{R}^d$ where $d << N$

- ISOMAP (Tenenbaum, et al, 00)
- LLE (Roweis, Saul, 00)
- Laplacian Eigenmaps (Belkin, Niyogi, 01)
- Local Tangent Space Alignment (Zhang, Zha, 02)
- Hessian Eigenmaps (Donoho, Grimes, 02)
- Diffusion Maps (Coifman, Lafon, et al, 04)

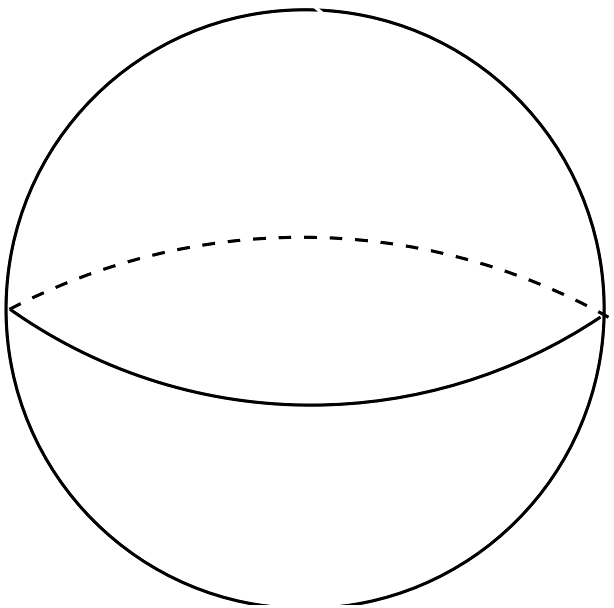
Related: Kernel PCA (Schoelkopf, et al, 98)

All you wanna know about differential geometry but were afraid to ask, in 9 easy slides

Embeded Manifolds

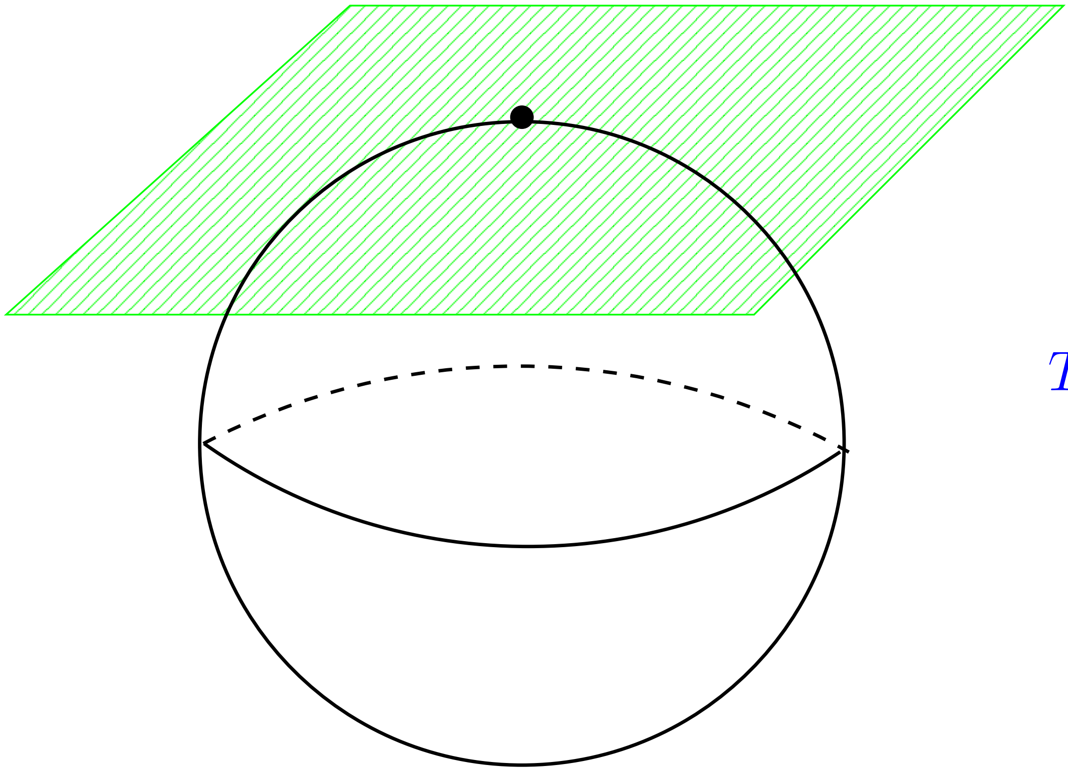
$$\mathcal{M}^k \subset \mathbb{R}^N$$

Locally (not globally) looks like Euclidean space.



$$S^2 \subset \mathbb{R}^3$$

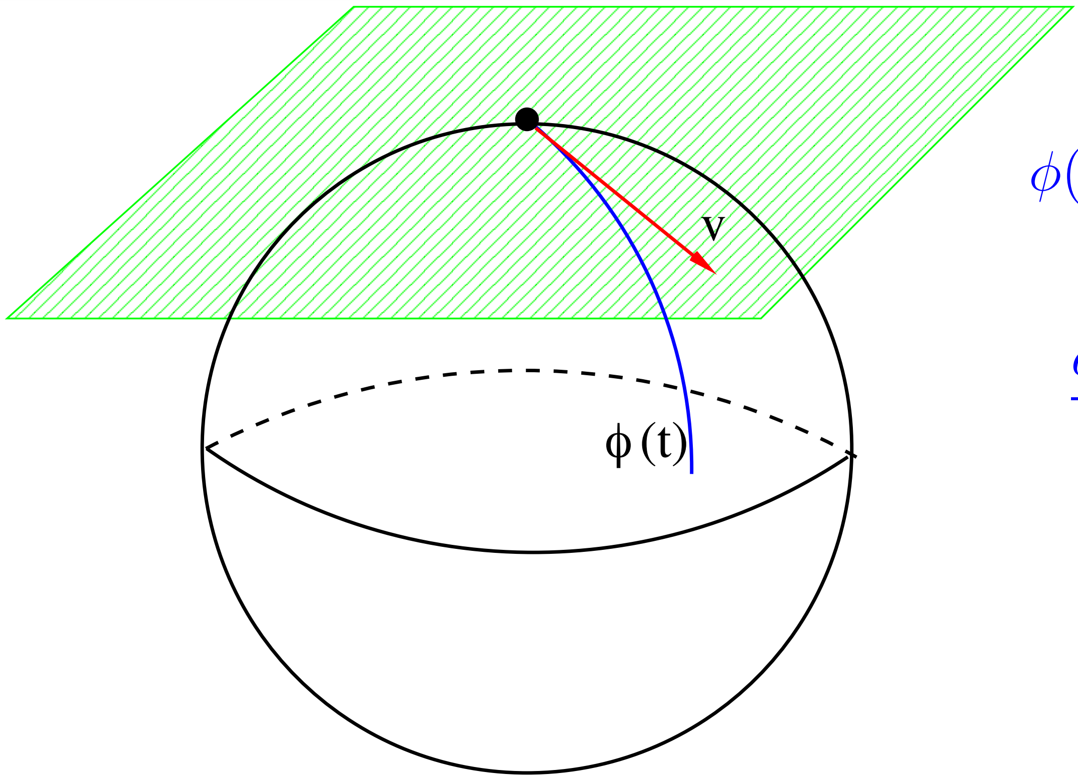
Tangent Space



$T_p\mathcal{M}^k \subset \mathbb{R}^N$

k-dimensional affine subspace of $\mathbb{R}^N$.

Tangent Vectors and Curves



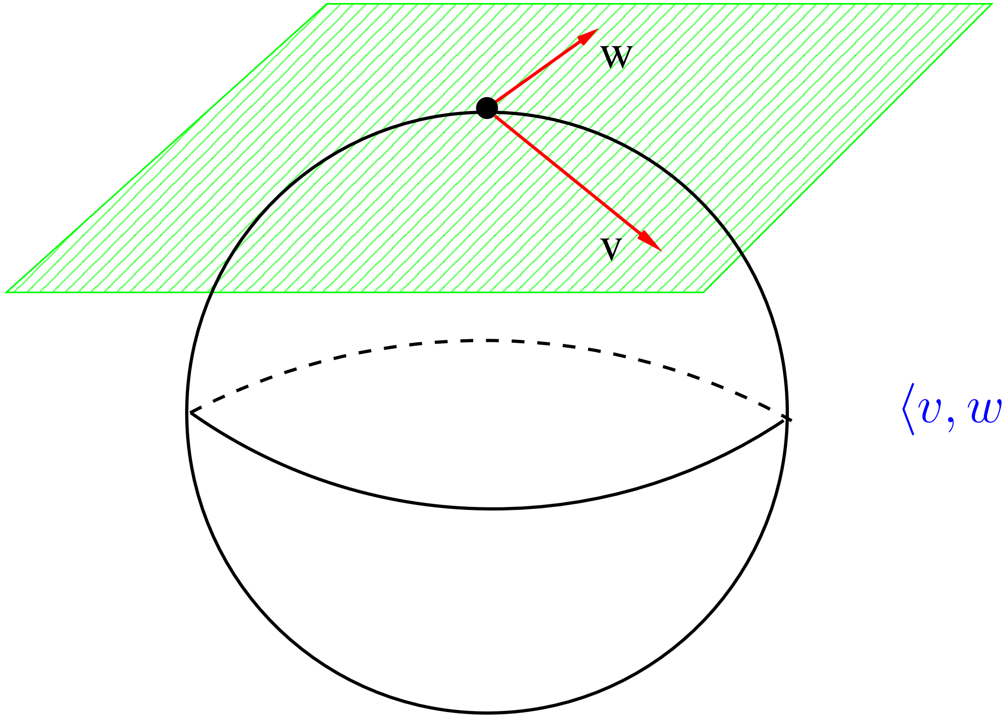
$$\phi(t) : \mathbb{R} \rightarrow \mathcal{M}^k$$

$$\left.\frac{d\phi(t)}{d\,t}\right|_0 = V$$

Tangent vectors <– – – –> curves.

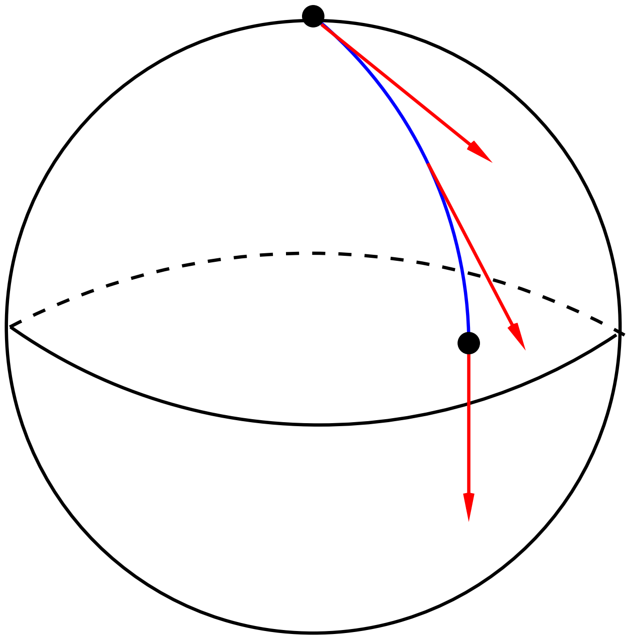
Riemannian Geometry

Norms and angles in tangent space.



$\langle v, w \rangle \qquad \|v\|,\ \|w\|$

Geodesics



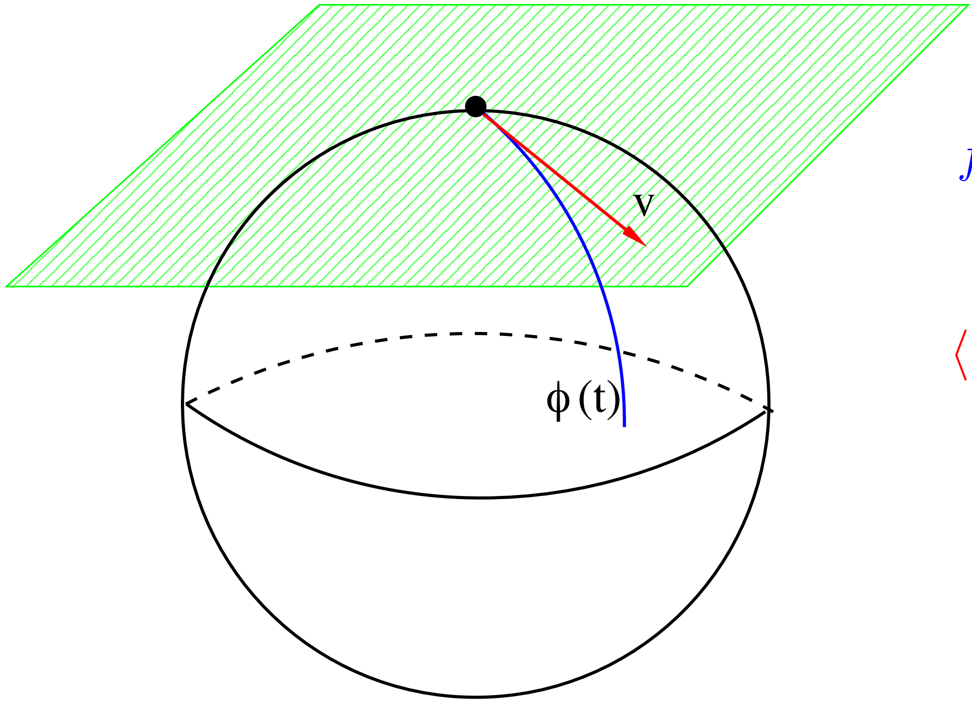
$$\phi(t) : [0, 1] \rightarrow \mathcal{M}^k$$

$$l(\phi) = \int_0^1 \left\| \frac{d\phi}{dt} \right\| dt$$

Can measure length using norm in tangent space.

Geodesic — shortest curve between two points.

Gradients



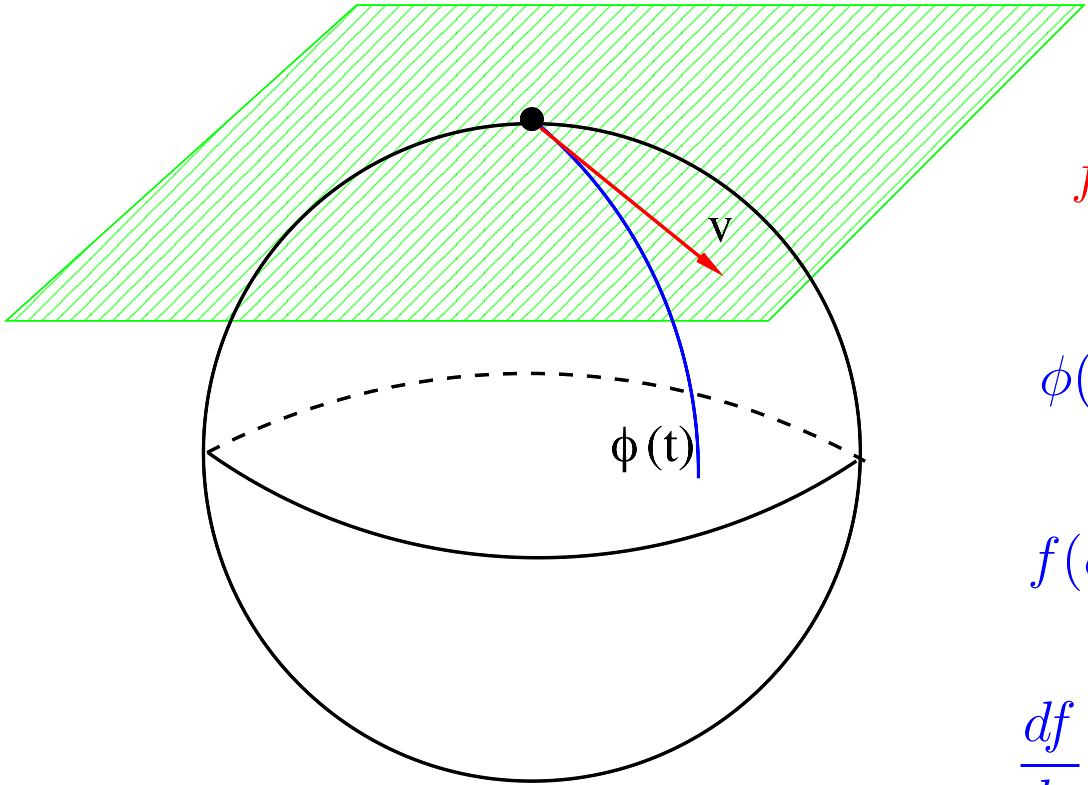
$$f : \mathcal{M}^k \longrightarrow \mathbb{R}$$

$$\langle \nabla f, v \rangle \equiv \frac{df}{dv}$$

Tangent vectors <— — —> Directional derivatives.

Gradient points in the direction of maximum change.

Tangent Vectors vs. Derivatives



$f : \mathcal{M}^k \to \mathbb{R}$

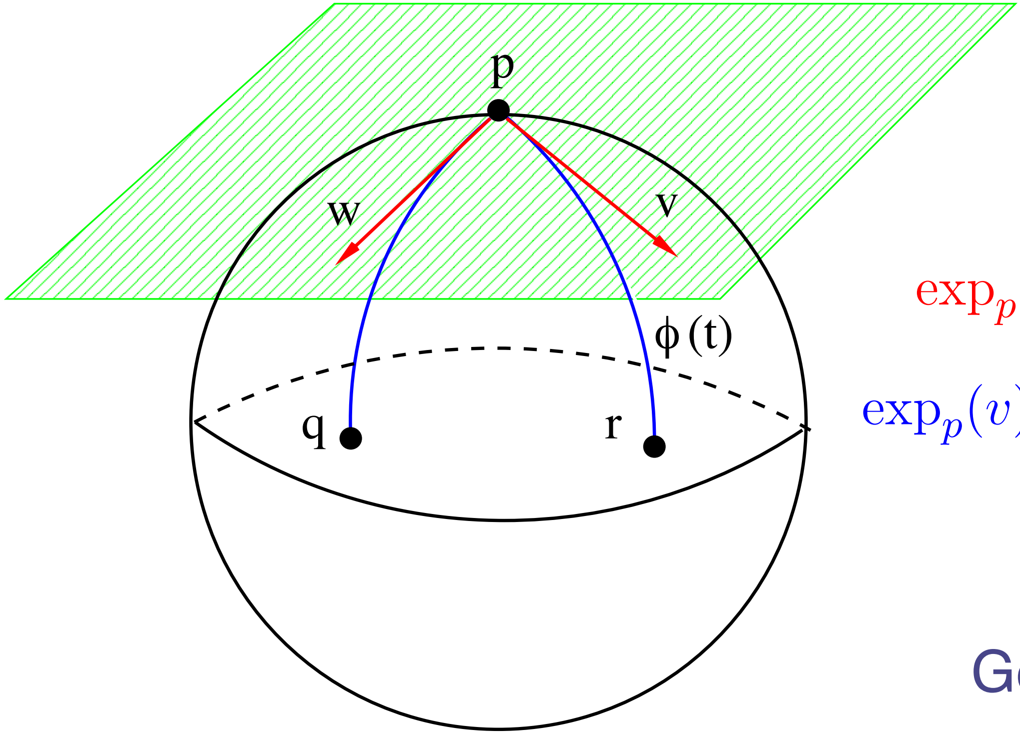
$\phi(t) : \mathbb{R} \to \mathcal{M}^k$

$f(\phi(t)) : \mathbb{R} \to \mathbb{R}$

$$\frac{df}{dv} = \left.\frac{d\,f(\phi(t))}{d\,t}\right|_0$$

Tangent vectors <— — —> Directional derivatives.

Exponential Maps



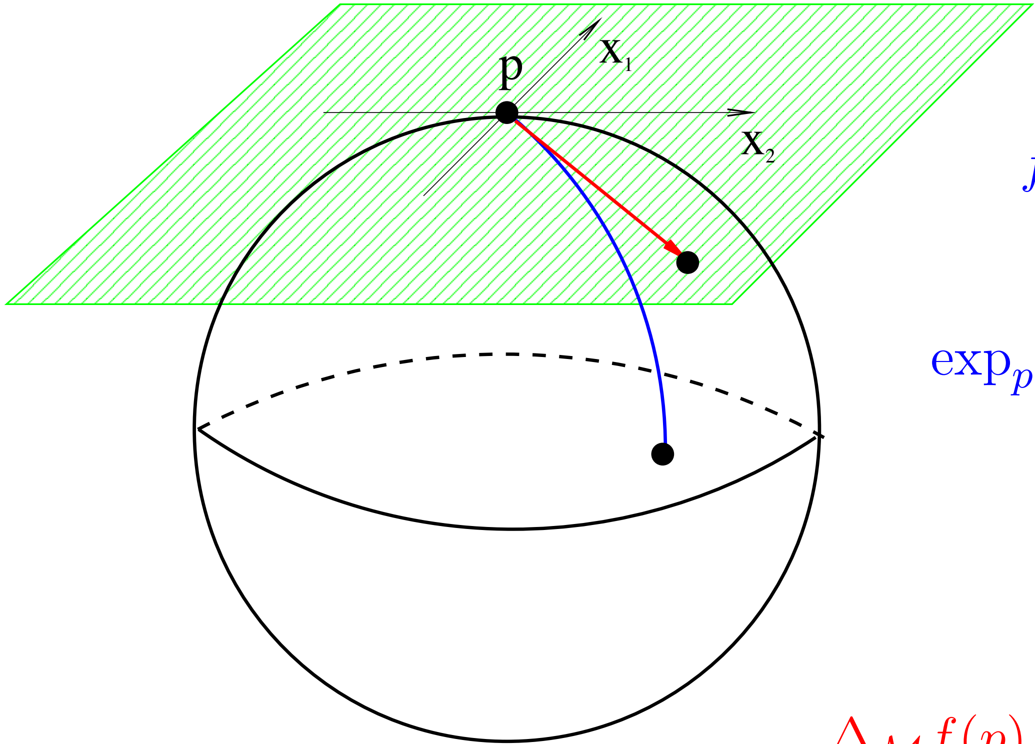
$\exp_p : T_p\mathcal{M}^k \rightarrow \mathcal{M}^k$

$\exp_p(v) = r \quad \exp_p(w) = q$

Geodesic $\phi(t)$

$$\phi(0) = p, \quad \phi(\|v\|) = q \qquad \left.\frac{d\phi(t)}{dt}\right|_0 = v$$

Laplacian-Beltrami Operator



$$f : \mathcal{M}^k \to \mathbb{R}$$

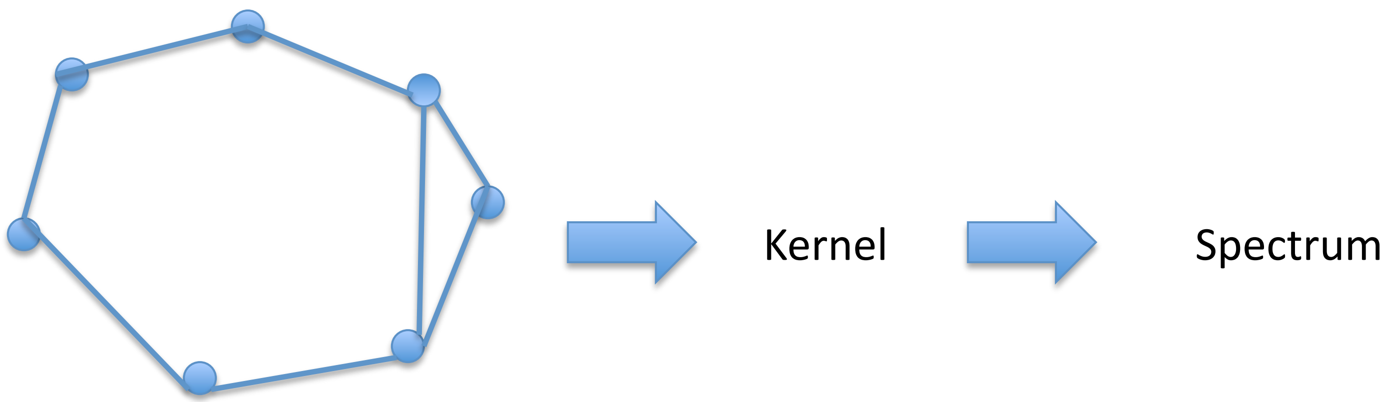
$$\exp_p : T_p\mathcal{M}^k \to \mathcal{M}^k$$

$$\Delta_{\mathcal{M}} f(p) \equiv \sum_i \frac{\partial^2 f(\exp_p(x))}{\partial x_i^2}$$

Orthonormal coordinate system.

Meta-Algorithm

1. Construct a neighborhood graph
2. Construct a positive semi-definite kernel
3. Find the eigen-decomposition



Recall: MDS

- Idea: Distances -> Inner Products -> Embedding
- Inner Product:

$$\|x-y\|^2 = \langle x,x\rangle + \langle y,y\rangle - 2\langle x,y\rangle$$

$$D_{ij} = K_{ii} + K_{jj} - 2K_{ij}$$

$$\Rightarrow \quad K = -\frac{1}{2}HDH^T, \qquad H = I - \frac{1}{n}11^T$$

- *K* is positive semi-definite with

$$K = U\Lambda U^T = YY^T, \qquad Y = U\Lambda^{1/2}$$

Recall: ISOMAP

1. Construct Neighborhood Graph.
2. Find shortest path (geodesic) distances.

   D_{ij} is $n \times n$

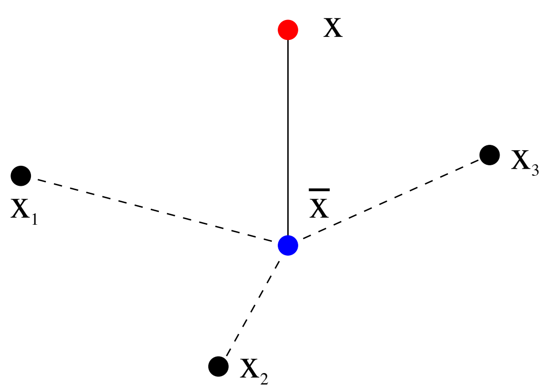
3. Embed using Multidimensional Scaling.

Recall: LLE (I)

1. Construct Neighborhood Graph.

2. Let $x_1, \ldots, x_n$ be neighbors of x. Project x to the span of $x_1, \ldots, x_n$.

3. Find barycentric coordinates of $\bar{x}$.



$$\bar{x} = w_1 x_1 + w_2 x_2 + w_3 x_3$$

$$w_1 + w_2 + w_3 = 1$$

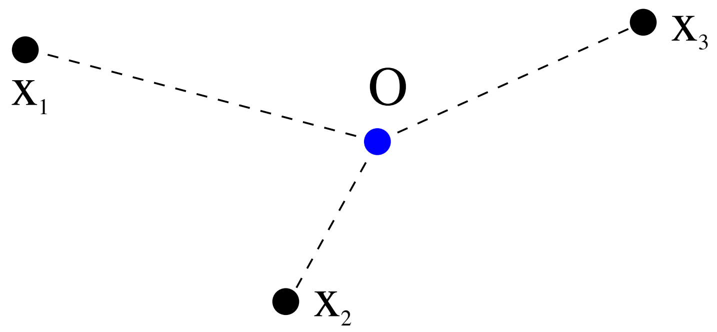
Weights w_1, w_2, w_3 chosen,
so that $\bar{x}$ is the center of mass.

Recall: LLE (II)

4. Construct sparse matrix W. i th row is barycentric coordinates of $\bar{x}_i$ in the basis of its nearest neighbors.

5. Use lowest eigenvectors of $(I - W)^t(I - W)$ to embed.

Laplacian and LLE



$$\sum w_i x_i = 0$$

$$\sum w_i = 1$$

Hessian H. Taylor expansion :

$$f(x_i) = f(0) + x_i^t \nabla f + \frac{1}{2} x_i^t H x_i + o(\|x_i\|^2)$$

$$(I - W)f(0) = f(0) - \sum w_i f(x_i) \approx f(0) - \sum w_i f(0) - \sum_i w_i x_i^t \nabla f - \frac{1}{2} \sum_i x_i^t H x_i =$$

$$= -\frac{1}{2} \sum_i x_i^t H x_i \approx -trH = \Delta f$$

Laplacian Eigenmaps (I) [Belkin-Niyogi]

Step 1 [*Constructing the Graph*]

$$e_{ij} = 1 \Leftrightarrow \mathbf{x}_i \text{ “close to” } \mathbf{x}_j$$

1. ϵ-neighborhoods. [*parameter* $\epsilon \in \mathbb{R}$] Nodes i and j are connected by an edge if

$$||\mathbf{x}_i - \mathbf{x}_j||^2 < \epsilon$$

2. n nearest neighbors. [*parameter* $n \in \mathbb{N}$] Nodes i and j are connected by an edge if i is among n nearest neighbors of j or j is among n nearest neighbors of i.

Laplacian Eigenmaps (II)

Step 2. [*Choosing the weights*].

1. Heat kernel. [*parameter* $t \in \mathbb{R}$]. If nodes i and j are connected, put

$$W_{ij} = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{t}}$$

2. Simple-minded. [*No parameters*]. $W_{ij} = 1$ if and only if vertices i and j are connected by an edge.

Laplacian Eigenmaps (III)

Step 3. [*Eigenmaps*] Compute eigenvalues and eigenvectors for the generalized eigenvector problem:

$$Lf = \lambda D f$$

D is diagonal matrix where

$$D_{ii} = \sum_j W_{ij}$$

$$L = D - W$$

Let $\mathbf{f}_0, \ldots, \mathbf{f}_{k-1}$ be eigenvectors.

Leave out the eigenvector $\mathbf{f}_0$ and use the next m lowest eigenvectors for embedding in an m-dimensional Euclidean space.

Justification

Find $y_1, \ldots, y_n \in R$

$$\min \sum_{i,j} (y_i - y_j)^2 W_{ij}$$

Tries to preserve locality

A Fundamental Identity

But

$$\frac{1}{2}\sum_{i,j}(y_i - y_j)^2 W_{ij} = \mathbf{y}^T L \mathbf{y}$$

$$\sum_{i,j}(y_i - y_j)^2 W_{ij} = \sum_{i,j}(y_i^2 + y_j^2 - 2y_i y_j) W_{ij}$$

$$= \sum_i y_i^2 D_{ii} + \sum_j y_j^2 D_{jj} - 2\sum_{i,j} y_i y_j W_{ij}$$

$$= 2\mathbf{y}^T L \mathbf{y}$$

Embedding as Eigenmaps

$$\lambda = 0 \rightarrow \mathbf{y} = \mathbf{1}$$

$$\min_{\mathbf{y}^T\mathbf{1}=0} \mathbf{y}^T L \mathbf{y}$$

Let $Y = [\mathbf{y}_1 \mathbf{y}_2 \ldots \mathbf{y}_m]$

$$\sum_{i,j} ||Y_i - Y_j||^2 W_{ij} = \text{trace}(Y^T L Y)$$

subject to $Y^T Y = I$.

Use eigenvectors of L to embed.

On the Manifold

smooth map $f : \mathcal{M} \to R$

$$\int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 \approx \sum_{i \sim j} W_{ij}(f_i - f_j)^2$$

Recall standard gradient in $\mathbb{R}^k$ of $f(z_1, \ldots, z_k)$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial z_1} \\ \frac{\partial f}{\partial z_2} \\ \cdot \\ \cdot \\ \frac{\partial f}{\partial z_k} \end{bmatrix}$$

Stokes Theorem

A Basic Fact

$$\int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 = \int f \cdot \Delta_{\mathcal{M}} f$$

This is like

$$\sum_{i,j} W_{ij}(f_i - f_j)^2 = \mathbf{f}^T \mathbf{L} \mathbf{f}$$

where

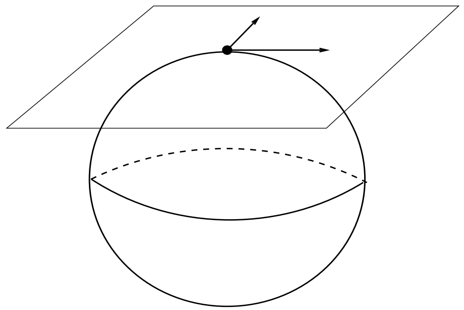
$\Delta_{\mathcal{M}} f$ is the manifold Laplacian

Manifold Laplacian

Recall ordinary Laplacian in $\mathbb{R}^k$
This maps

$$f(x_1, \ldots, x_k) \to \left(- \sum_{i=1}^{k} \frac{\partial^2 f}{\partial x_i^2} \right)$$

Manifold Laplacian is the same on the tangent space.



Manifold Laplacian Eigenvectors

Eigensystem

$$\Delta_{\mathcal{M}} f = \lambda_i \phi_i$$

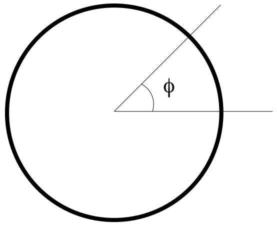
$\lambda_i \geq 0$ and $\lambda_i \to \infty$

$\{\phi_i\}$ form an orthonormal basis for $L^2(\mathcal{M})$

$$\int \|\nabla_{\mathcal{M}} \phi_i\|^2 = \lambda_i$$

Manifold Laplacian is non-compact!

Example: Circle



$$-\frac{d^2u}{dt^2} = \lambda u \text{ where } u(0) = u(2\pi)$$

Eigenvalues are

$$\lambda_n = n^2$$

Eigenfunctions are

$$\sin(nt), \cos(nt)$$

Spherical Harmonics in high-D sphere!

Spectral Growth

$$\lambda_1 \leq \lambda_2 \ldots \leq \lambda_j \leq \ldots$$

Then

$$A + \frac{2}{d}\log(j) \leq \log(\lambda_j) \leq B + \frac{2}{d}\log(j+1)$$

Example: on S^1

$$\lambda_j = j^2 \implies \log(\lambda_j) = \frac{2}{1}\log(j)$$

(Li and Yau; Weyl's asymptotics)

From Graph to Manifolds

$$f : \mathcal{M} \to \mathbb{R} \quad x \in \mathcal{M} \quad x_1, \ldots, x_n \in \mathcal{M}$$

Graph Laplacian:

$$L_n^t(f)(x) = f(x) \sum_j e^{-\frac{\|x - x_j\|^2}{t}} - \sum_j f(x_j) e^{-\frac{\|x - x_j\|^2}{t}}$$

Theorem [pointwise convergence] $t_n = n^{-\frac{1}{k+2+\alpha}}$

$$\lim_{n \to \infty} \frac{(4\pi t_n)^{-\frac{k+2}{2}}}{n} L_n^{t_n} f(x) = \Delta_{\mathcal{M}} f(x)$$

Belkin 03, Lafon Coifman 04, Belkin Niyogi 05, Hein et al 05

From Graph to Manifolds

Theorem [convergence of eigenfunctions]

$$\lim_{t \to 0, n \to \infty} Eig[L_n^{t_n}] \to Eig[\Delta_{\mathcal{M}}]$$

Belkin Niyogi 06

Heat Diffusion Map

- Gaussian kernel

$$K_\varepsilon(x,y) = \exp\left(-\frac{\|x-y\|^2}{\varepsilon^2}\right)$$

- Normalize kernel

$$K^{(\alpha)}(x,y) = \frac{K_\varepsilon(x,y)}{p^\alpha(x)p^\alpha(y)} \quad where \quad p(x) = \int K_\varepsilon(x,y)d\mu(y)$$

- Renormalized kernel

$$A_\varepsilon(x,y) = \frac{K^{(\alpha)}(x,y)}{\sqrt{d^{(\alpha)}(x)}\sqrt{d^{(\alpha)}(y)}} \quad where \quad d^{(\alpha)}(x) = \int K^{(\alpha)}(x,y)d\mu(y)$$

  – α=1, Laplacian-Beltrami operator, separate geometry from density
  – α=0, classical normalized graph Laplacian
  – α=1/2, backward Fokkar-Planck operator

Coifman-Lafon 2006. Diffusion Maps.

Heat Diffusion Distance

Heat diffusion operator H^t. $H^t = \exp(-tL_n) \quad where \quad L_n = I - D^{-1/2} W D^{-1/2}$

δ_x and δ_y initial heat distributions.

Diffusion distance between x and y:

$$\|H^t \delta_x - H^t \delta_y\|_{L^2}$$

Difference between heat distributions after time t.

Note: Another choice of eigenmaps

- Normalized positive semi-definite Laplacian

$$L_n = D^{-1/2}(D - W)D^{-1/2} = I - D^{-1/2}WD^{-1/2}$$

- ϕ_i is an eigenvector of L_n with eigenvalue λ_i
- Normalized Laplacian eigenmaps:

$$Y = \begin{pmatrix} \lambda_1^{1/2}\phi_1 & \lambda_2^{1/2}\phi_2 & \ldots & \lambda_d^{1/2}\phi_d \end{pmatrix}$$

Connections to Markov Chain

- $L = D\text{-}W$: unnormalized graph Laplacian
- $L_n = D^{-1/2} L D^{-1/2}$: normalized graph Laplacian
- $P = I - D^{-1}L = D^{-1}W$ is the markov matrix
- v is generalized eigenvector of L: $L v = \lambda D v$
- v is also a right eigenvector of P with eigenvalue $1-\lambda$
- $D^{1/2} v$ is eigenvectors of L_n with eigenvalue λ
- P is lumpable iff v is piece-wise constant
- So *v is the most often choice of* Laplacian eigenmaps and Diffusion Map

Two Assumptions on ISOMAP

(ISO1) *Isometry.* The mapping ψ preserves geodesic distances. That is, define a distance between two points m and m' on the manifold according to the distance travelled by a bug walking along the manifold M according to the shortest path between m and m'. Then the isometry assumption says that

$$G(m, m') = |\theta - \theta'|, \qquad \forall m \leftrightarrow \theta, m' \leftrightarrow \theta',$$

where $|\cdot|$ denotes Euclidean distance in $\mathbb{R}^d$.

(ISO2) *Convexity.* The parameter space Θ is a convex subset of $\mathbb{R}^d$. That is, if θ, θ' is a pair of points in Θ, then the entire line segment $\{(1-t)\theta + t\theta' : t \in (0,1)\}$ lies in Θ.

Convexity is hard to meet: consider two balls in an image which never intersect, whose center coordinate space (x_1,y_1,x_2,y_2) must have a hole.

Relaxations (Donoho-Grimes'2003)

(LocISO1) *Local Isometry.* In a small enough neighborhood of each point m, geodesic distances to nearby points m' in M are identical to Euclidean distances between the corresponding parameter points θ and θ'.

(LocISO2) *Connectedness.* The parameter space Θ is a open connected subset of $\mathbb{R}^d$.

Hessian LLE

- Summary
  - Build graph from K Nearest Neighbors.
  - Estimate tangent Hessians.
  - Compute embedding based on Hessians.

$$f : X \to \Re \qquad Basis\left(null\left(\int \left\| H_f(x) \right\| dx\right)\right) = Basis(X)$$

- Predictions
  - Specifically set up to handle non-convexity.
  - Slower than LLE & Laplacian.
  - Will perform poorly in sparse regions.
  - Only method with convergence guarantees.

Note that: $\Delta(f) = trace\left(H(f)\right)$

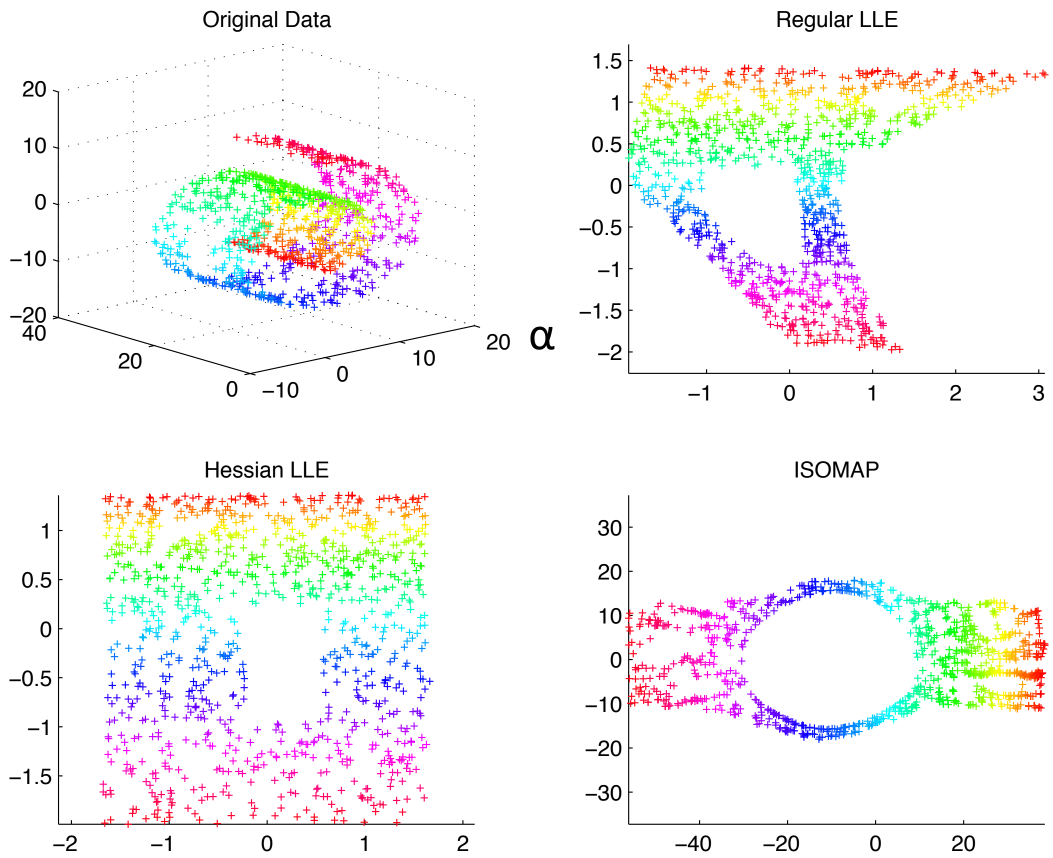
Convergence of Hessian LLE (Donoho-Grimes)

Theorem 1 *Suppose* $M = \psi(\Theta)$ *where* Θ *is an open connected subset of* $\mathbb{R}^d$*, and* ψ *is a locally isometric embedding of* Θ *into* $\mathbb{R}^n$*. Then* $\mathcal{H}(f)$ *has a* $d+1$ *dimensional nullspace, consisting of the constant function and a d-dimensional space of functions spanned by the original isometric coordinates.*

We give the proof in Appendix A.

Corollary 2 *Under the same assumptions as Theorem 1, the original isometric coordinates* θ *can be recovered, up to a rigid motion, by identifying a suitable basis for the null space of* $\mathcal{H}(f)$*.*

Comparisons on Swiss Roll with holes



Comparisons of Manifold Learning Techniques

- MDS
- PCA
- ISOMAP
- LLE
- Hessian LLE
- Laplacian LLE
- Diffusion Map
- Local Tangent Space Alignment
- Matlab codes: mani.m

Courtesy of Todd Wittman

Reference

- Tenenbaum, de Silva, and Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290:2319-2323, 22 Dec. 2000.
- Roweis and Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290:2323-2326, 22 Dec. 2000.
- M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15:1373–1396, 2003.
- D. L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. PNAS 100 (10): 5591–5596 2003.
- R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. PNAS 102 (21):7426-7431, 2005.
- M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. In P. Auer and R. Meir, editors, Proc. of the 18th Conf. on Learning Theory (COLT), pages 486–500, Berlin, 2005. Springer.
- Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. Ann. Statist. Volume 36, Number 2 (2008), 555-586.