Machine Learning and the Geometry of Data

Mikhail Belkin
Dept. of Computer Science and Engineering,
Department of Statistics
Ohio State University

Collaborators: Partha Niyogi, Qichao Que, Vikas Sindwani, Jian Sun, Yusu Wang, Xueyuan Zhou

Mathematics of Machine Learning

Choose the pattern ϕ to "fit" the data.

Key: controlling the space of predictors ϕ. [Avoiding the curse of dimensionality]
[typically using statistical assumptions to validate models]

- Spaces of low VC-dimension, parameterized families.

Linear methods, parametric families, neural networks…

- Smoothness

Kernel methods, splines, regularization in RKHS, Support Vector Machines.

- Sparsity

Wavelets, LASSO, compressed sensing, L_1 regularization.

- Geometry -- understanding the shape of the domain.

Graph methods, Laplacian-based methods, diffusions, topological methods.

Mathematics needed: Functional analysis, probability/statistics, combinatorics, graph theory, approximation theory, differential geometry, topology, algorithms and numerical methods.

Geometry and Manifold learning

Two main points:

1. Natural data is non-uniform and concentrates along lower dimensional structures.

2. The shape of the data can be exploited for learning patterns.

The notion of a Riemannian manifold is a very general and powerful mathematical framework for describing geometry.

Note: in high dimension only nearest neighbors make sense.

► In many domains (e.g., speech, some vision problems) data explicitly lies on a manifold.

Ubiquity of manifolds

- In many domains (e.g., speech, some vision problems) data explicitly lies on a manifold.

- For all sources of high-dimensional data, true dimensionality is much lower than the number of features.

Ubiquity of manifolds

- In many domains (e.g., speech, some vision problems) data explicitly lies on a manifold.

- For all sources of high-dimensional data, true dimensionality is much lower than the number of features.

- Much of the data is highly nonlinear.

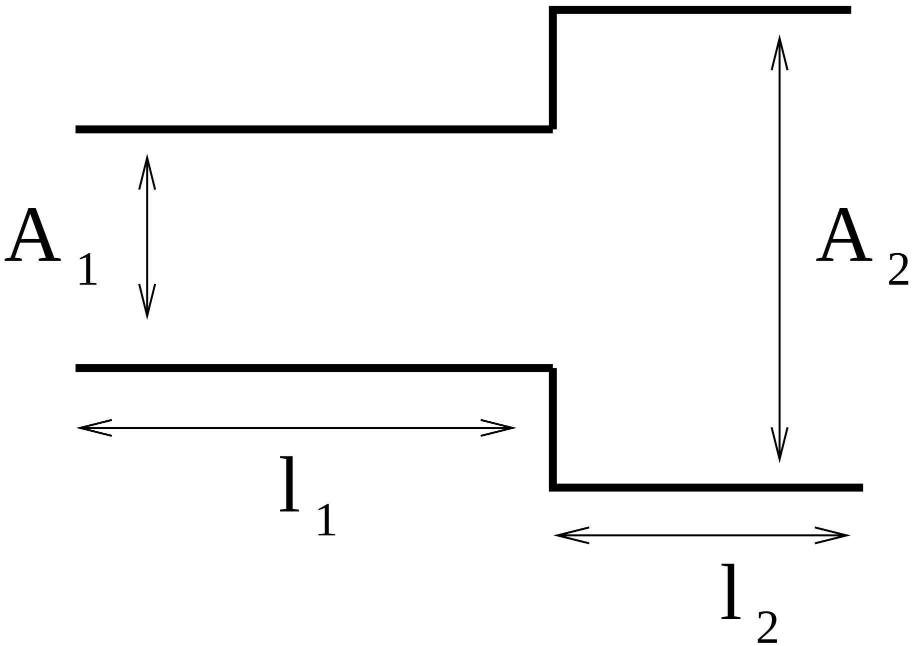
Ubiquity of manifolds

- In many domains (e.g., speech, some vision problems) data explicitly lies on a manifold.

- For all sources of high-dimensional data, true dimensionality is much lower than the number of features.

- Much of the data is highly nonlinear.

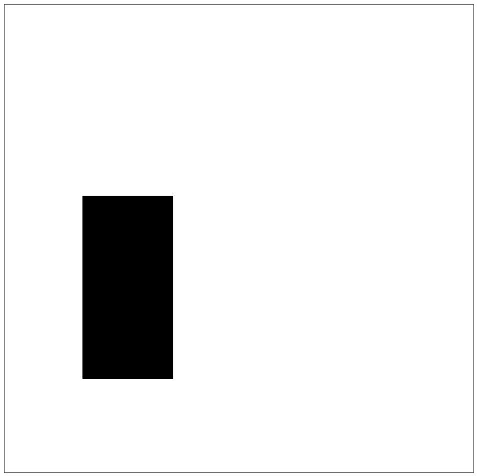
Manifolds (Riemannian manifolds with a measure + noise) provide a natural mathematical language for thinking about high-dimensional data.



Vocal tract modeled as a sequence of tubes.
(e.g. Stevens, 1998)

$$f : \mathbb{R}^2 \longrightarrow [0, 1]$$

$$\mathcal{F} = \{f | f(x, y) = v(x - t, y - r)\}$$



$$g : S^2 \times S^2 \times S^2 \to \mathbb{R}^3$$

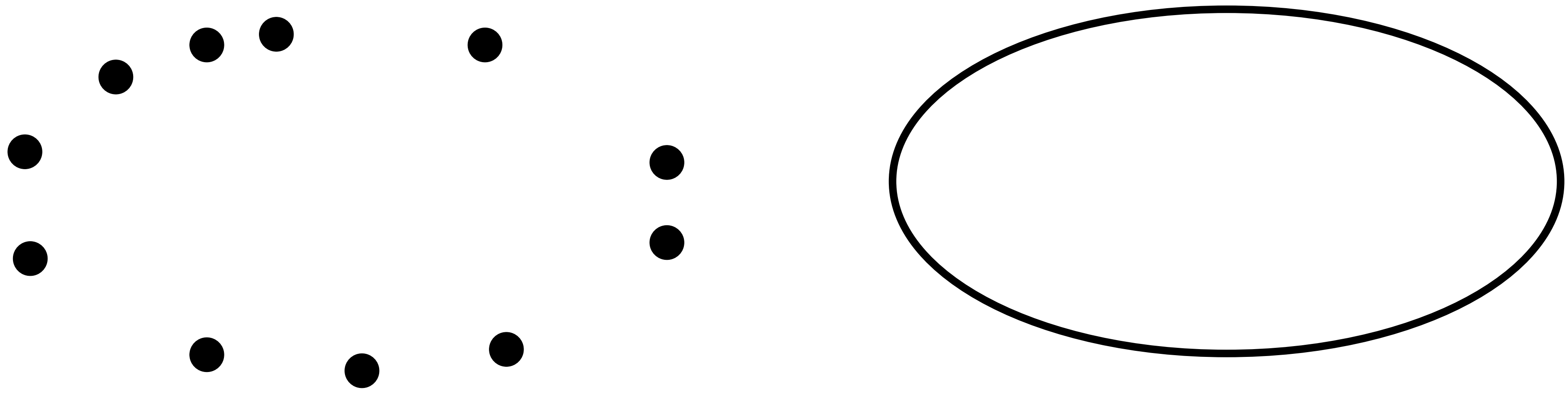
$$\langle(\theta_1, \phi_1), (\theta_2, \phi_2), (\theta_3, \phi_3)\rangle \to (x, y, z$$

Data ——— Probability Distribution

Graph ——— Manifold

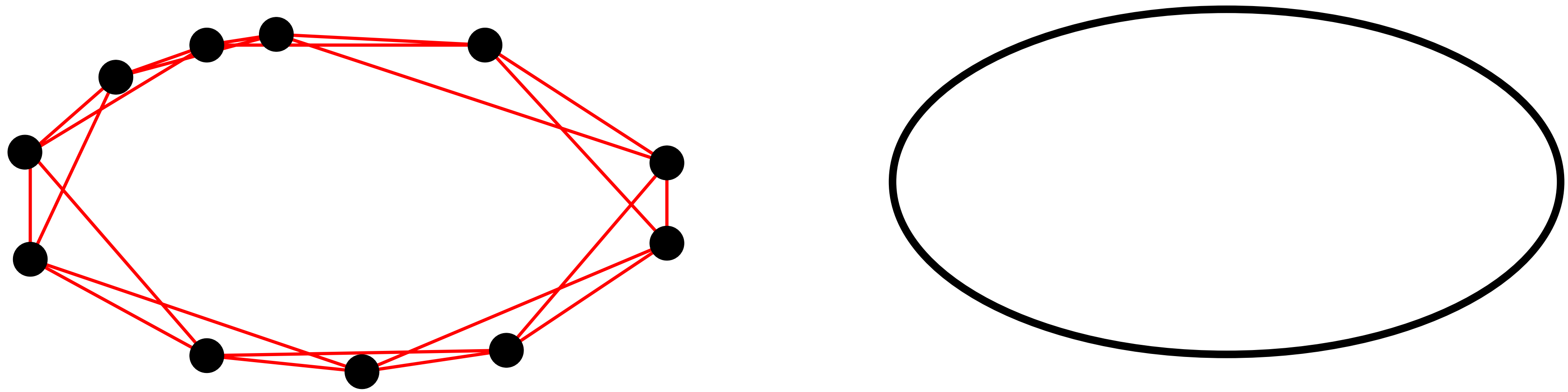
Data —— Probability Distribution

Graph —— Manifold



Data —— Probability Distribution

Graph —— Manifold

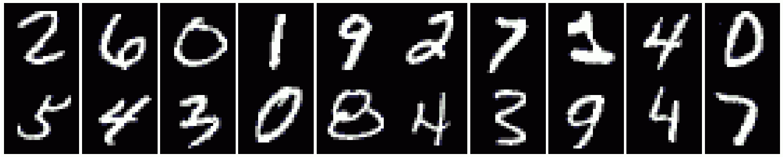


Graph extracts underlying geometric structure.

Problems of machine learning

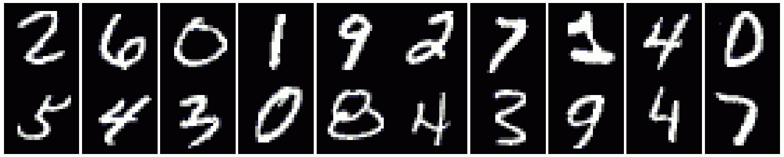
- Classification / regression.
- Data representation / dimensionality reduction.
- Clustering.

- Classification / regression.
- Data representation / dimensionality reduction.
- Clustering.



Problems of machine learning

- Classification / regression.
- Data representation / dimensionality reduction.
- Clustering.



Common intuition – similar objects have similar labels.

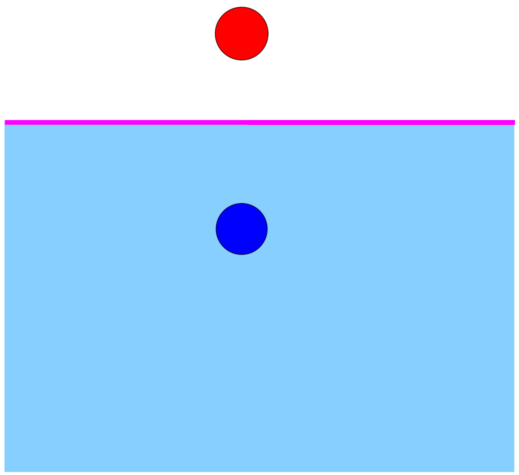
How does shape of the data affect the notion of similarity?

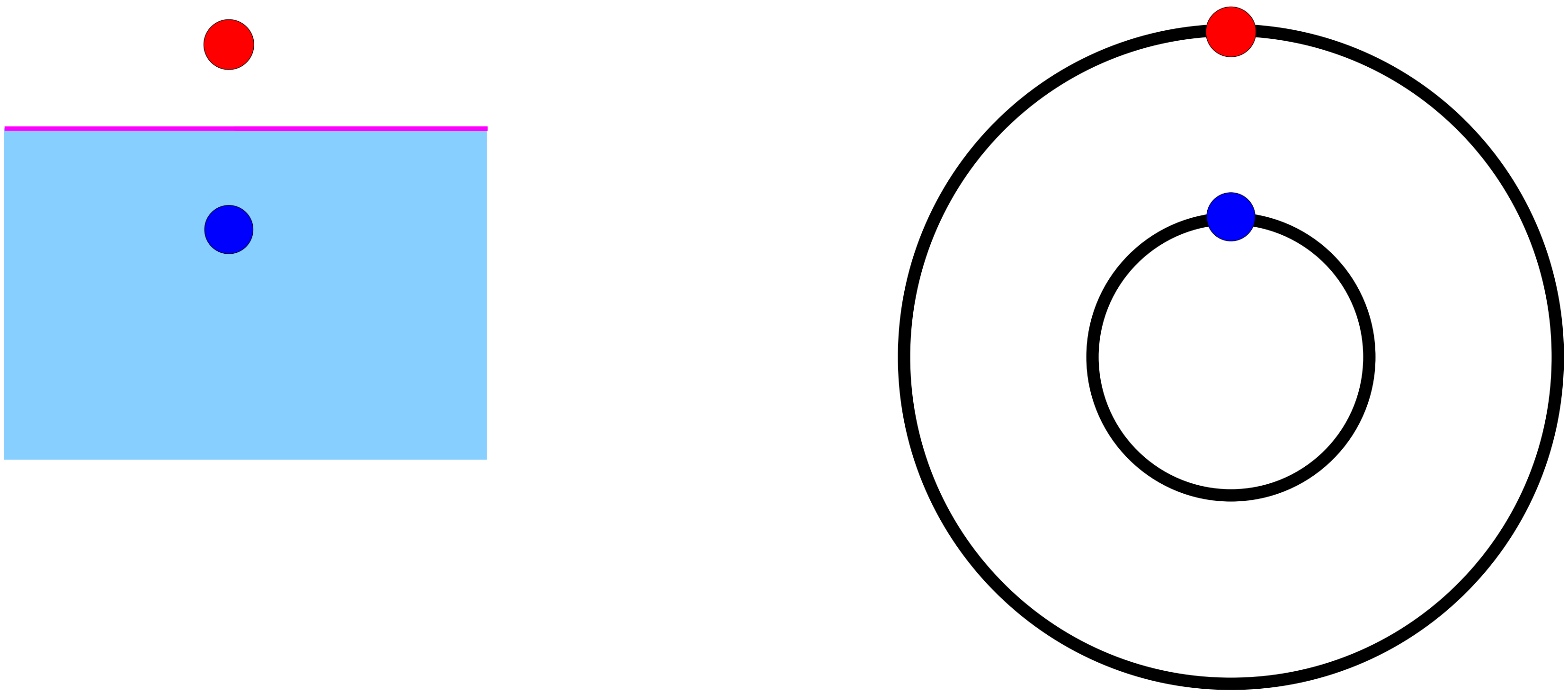
- Manifold assumption.
- Cluster assumption.

Reflect our understanding of structure of natural data.

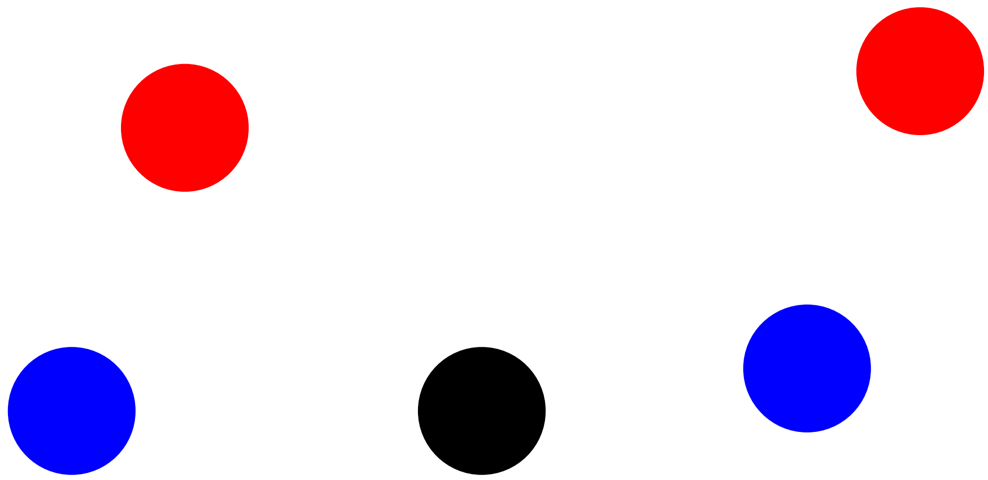
Intuition

Intuition

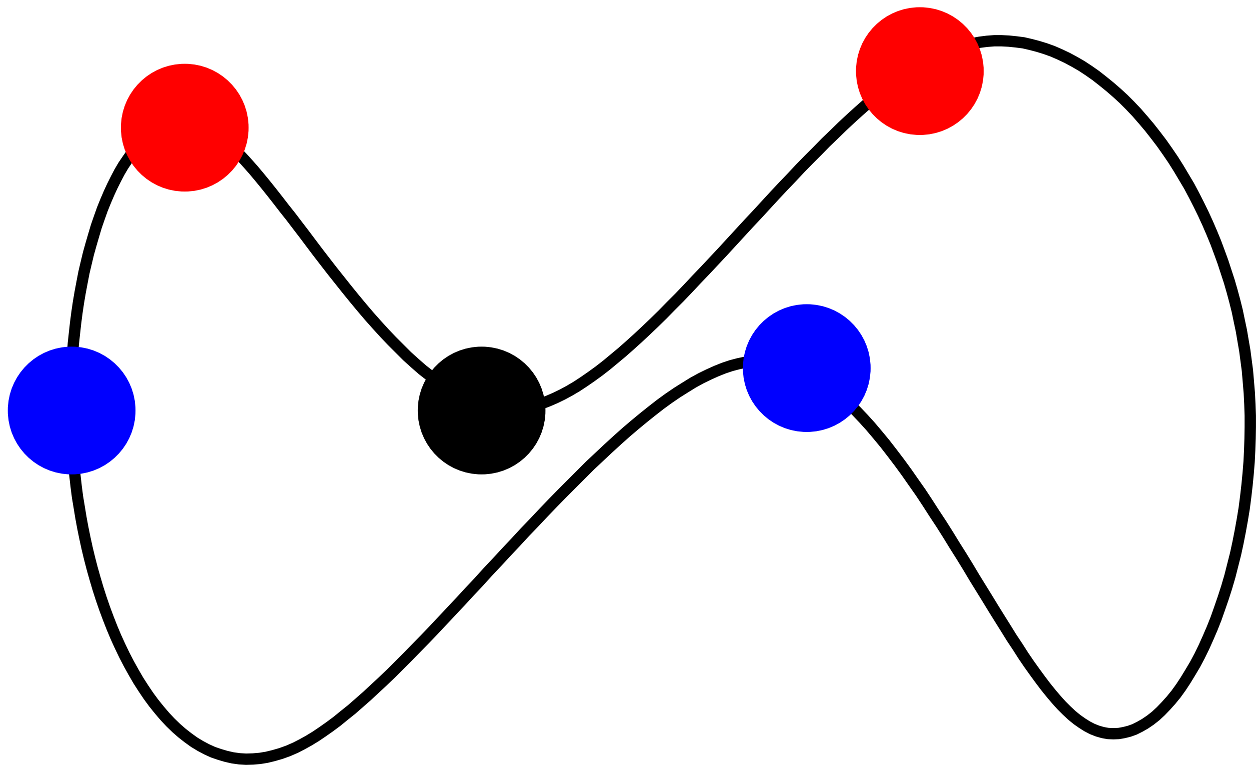




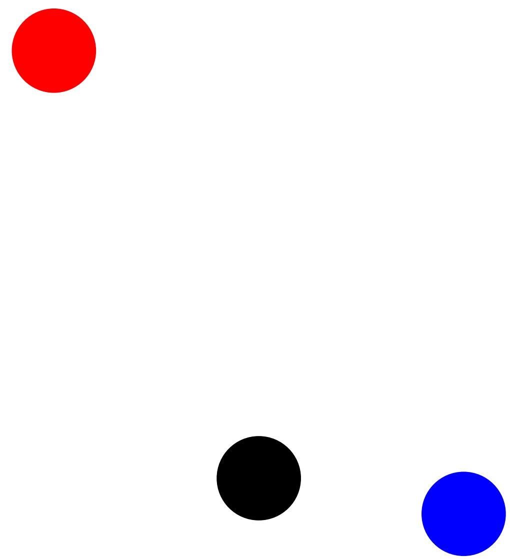
Manifold assumption



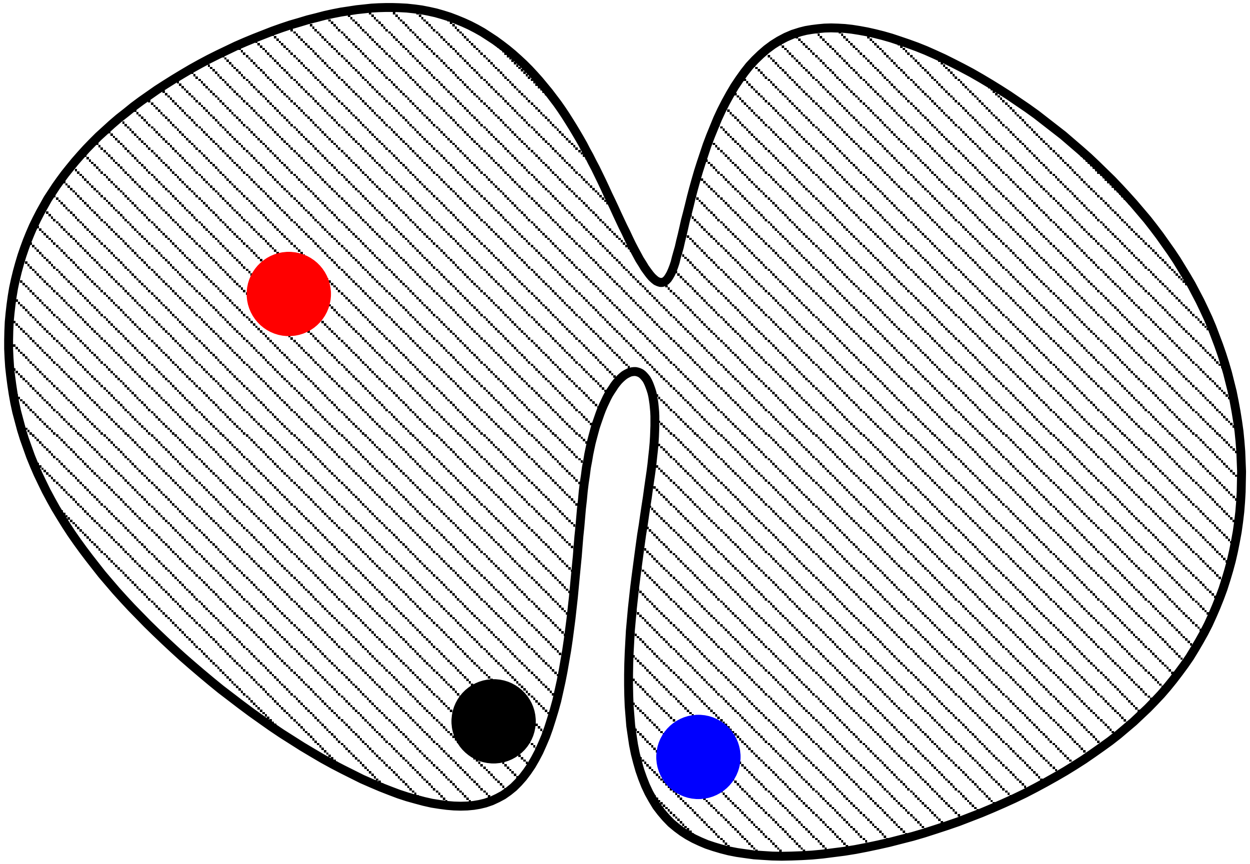
Manifold assumption

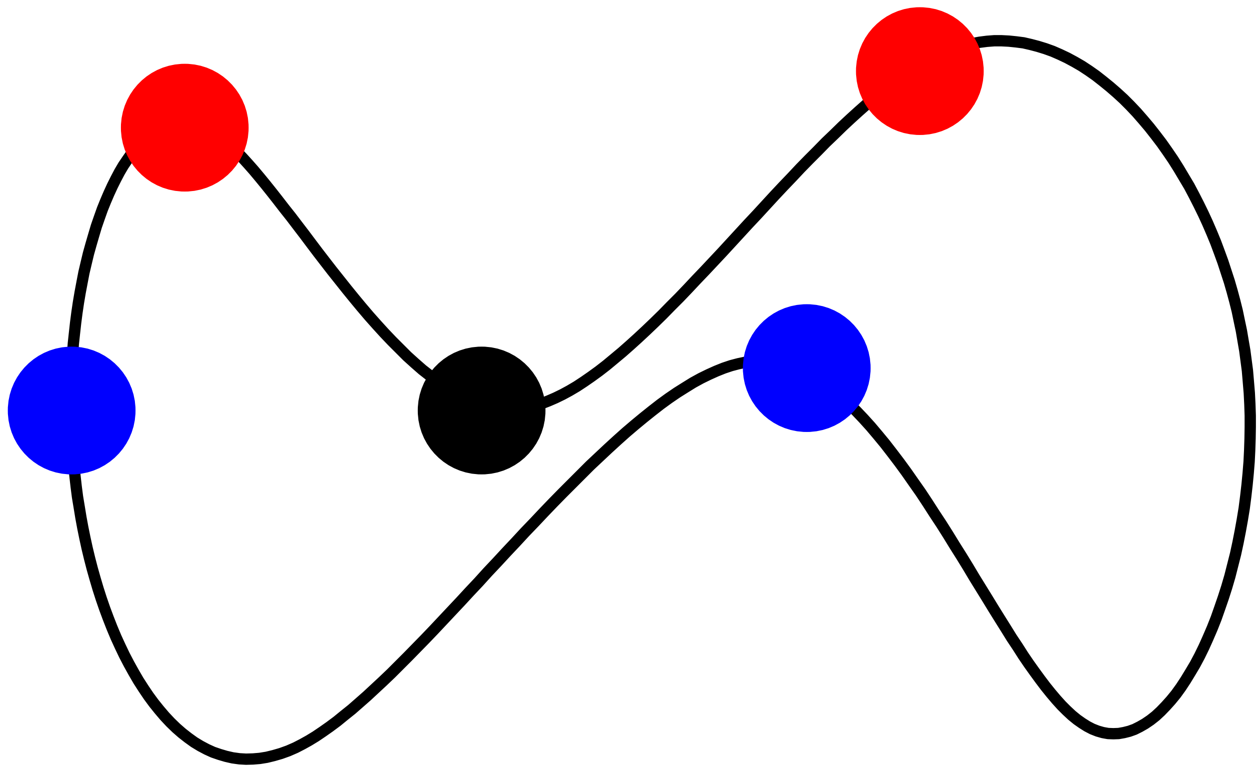


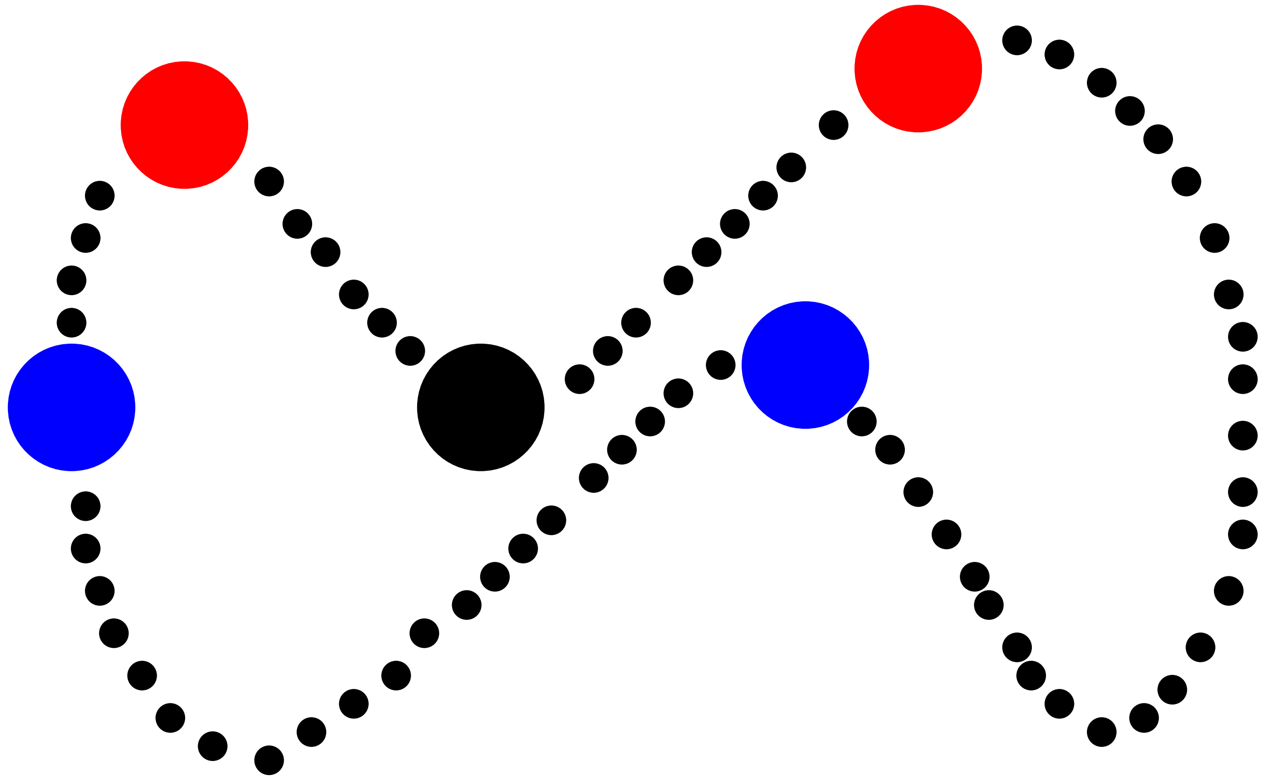
Cluster assumption



Cluster assumption

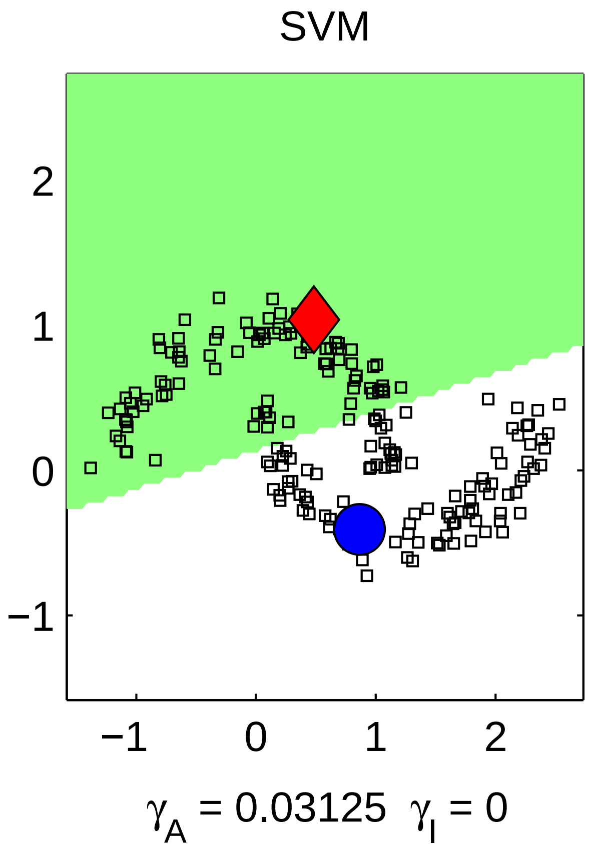




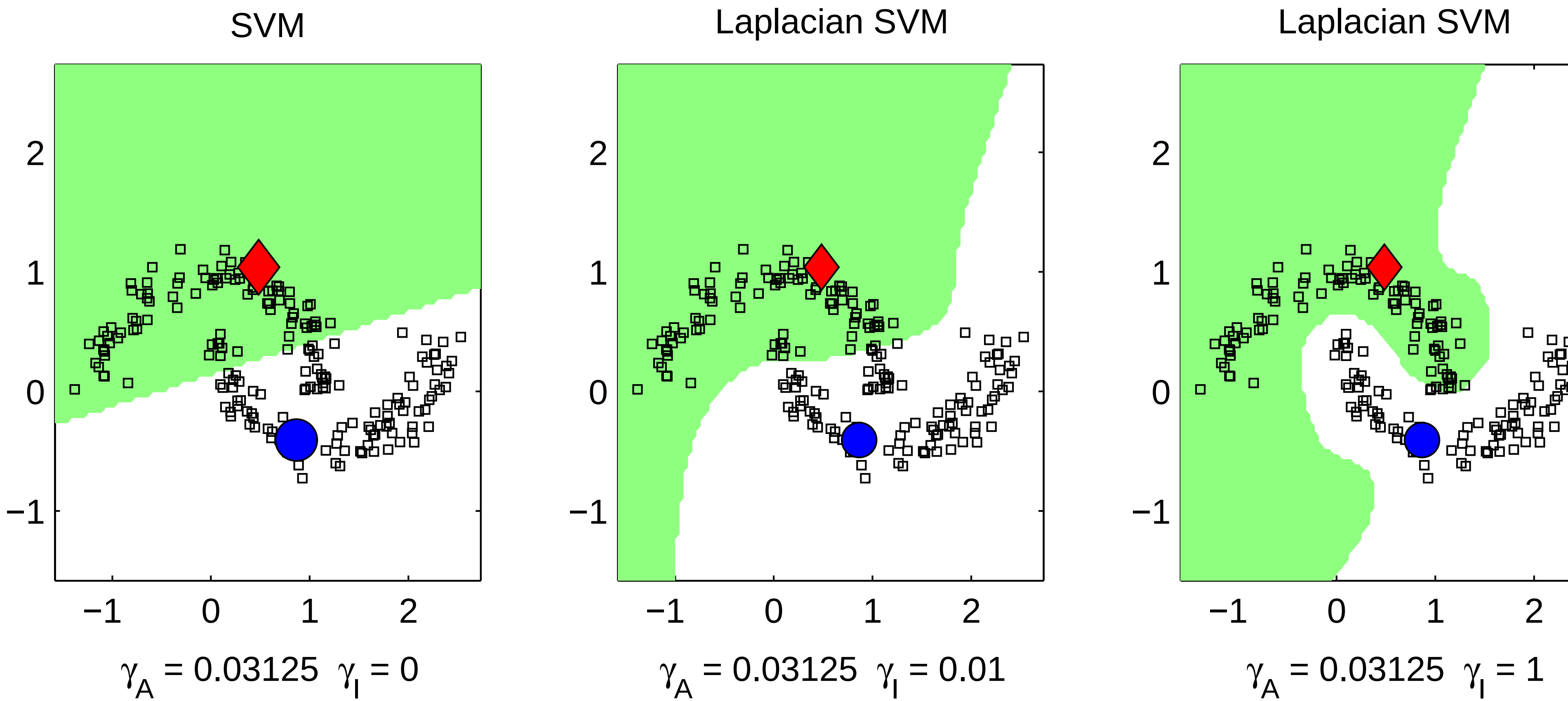


Unlabeled data to estimate geometry.

Toy example



Toy example


SVM
Laplacian SVM
Laplacian SVM
2
1
0
−1
−1
0
1
2
$\gamma_A = 0.03125 \quad \gamma_I = 0$
$\gamma_A = 0.03125 \quad \gamma_I = 0.01$
$\gamma_A = 0.03125 \quad \gamma_I = 1$

Manifold assumption

Manifold/geometric assumption:
functions of interest are smooth with respect to the underlying geometry.

Manifold/geometric assumption:
functions of interest are smooth with respect to the underlying geometry.

Probabilistic setting:
Map $X \to Y$. Probability distribution P on $X \times Y$.

Regression/(two class)classification: $X \to \mathbb{R}$.

Manifold assumption

Manifold/geometric assumption:
functions of interest are smooth with respect to the underlying geometry.

Probabilistic setting:
Map $X \to Y$. Probability distribution P on $X \times Y$.

Regression/(two class)classification: $X \to \mathbb{R}$.

Probabilistic version:
conditional distributions $P(y|x)$ are smooth with respect to the marginal $P(x)$.

Function $f : X \to \mathbb{R}$. Penalty at $x \in X$:

$$\frac{1}{\delta^k} \int_{\text{small } \delta} (f(x) - f(x+\delta))^2 p(x) d\delta \approx \|\nabla f\|^2 p(x)$$

Total penalty – Laplace operator:

$$\int_X \|\nabla f\|^2 p(x) = \langle f, \Delta_p f \rangle_X$$

Function $f : X \to \mathbb{R}$. Penalty at $x \in X$:

$$\frac{1}{\delta^k} \int_{\text{small } \delta} (f(x) - f(x+\delta))^2 p(x) d\delta \approx \|\nabla f\|^2 p(x)$$

Total penalty – Laplace operator:

$$\int_X \|\nabla f\|^2 p(x) = \langle f, \Delta_p f \rangle_X$$

Two-class classification – conditional $P(1|x)$.

Manifold assumption: $\langle P(1|x), \Delta_p P(1|x) \rangle_X$ is small.

Laplace operator is a fundamental geometric object.

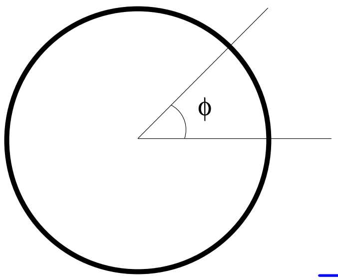
$$\Delta f = -\sum_{i=1}^{k} \frac{\partial^2 f}{\partial x_i^2}$$

The only differential operator invariant under translations and rotations.

Heat, Wave, Schroedinger equations.

Fourier analysis.

Laplacian on the circle



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

Same as in $\mathbb{R}$ with periodic boundary conditions.

Eigenvalues:

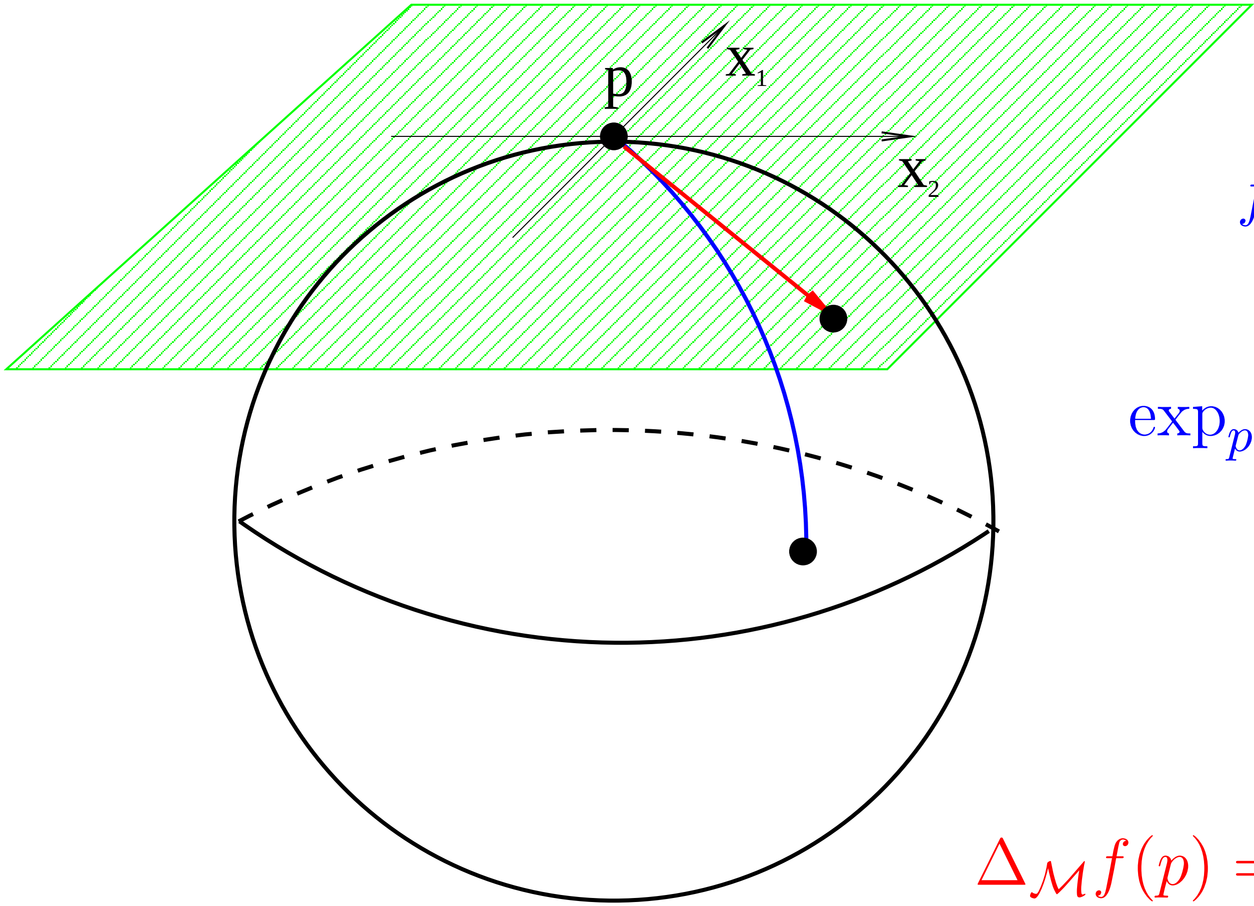
$$\lambda_n = n^2$$

Eigenfunctions:

$$\sin(n\phi), \cos(n\phi)$$

Fourier analysis.

Laplace-Beltrami operator



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

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

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

Generalization of Fourier analysis.

Laplace-Beltrami operator

Eigenfunctions of the Laplace-Beltrami operator provide a basis for L_2 functions on the manifold ordered by smoothness according to the eigenvalue.

The span of a few bottom eigenvectors $(e_1 \dots e_k)$ is a natural space of predictors for fitting data.

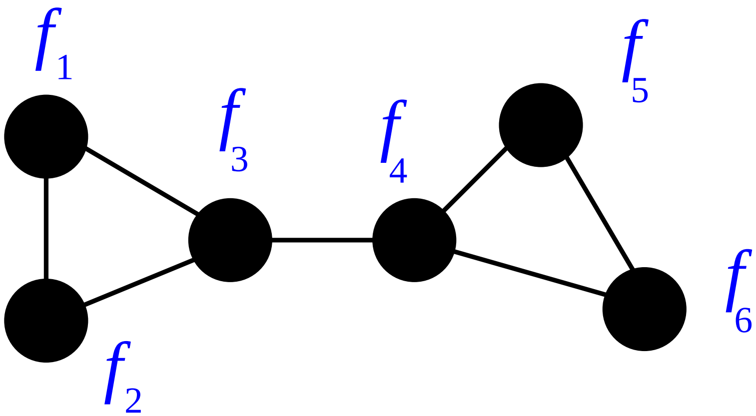
Data (x_i, y_i). Simplest learning method:

$$\min_{a_i, i=1..k} \sum_{j=1}^{n} \left(\sum_{i=1}^{k} a_i e_i - y_i \right)^2$$

Predictor: $\phi(\mathrm{x}) = \sum_{i=1}^{k} a_i e_i(x)$

What to do when the manifold is not known?

Algorithmic framework: Laplacian



$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

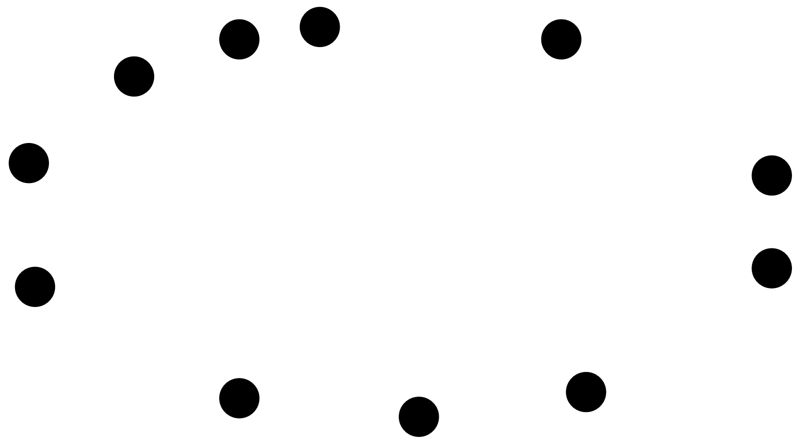
Natural smoothness functional (analogue of grad):

$$\mathcal{S}(\mathbf{f}) = (f_1-f_2)^2+(f_1-f_3)^2+(f_2-f_3)^2+(f_3-f_4)^2+(f_4-f_5)^2+(f_4-f_5)^2+(f_5-f_6)^2$$

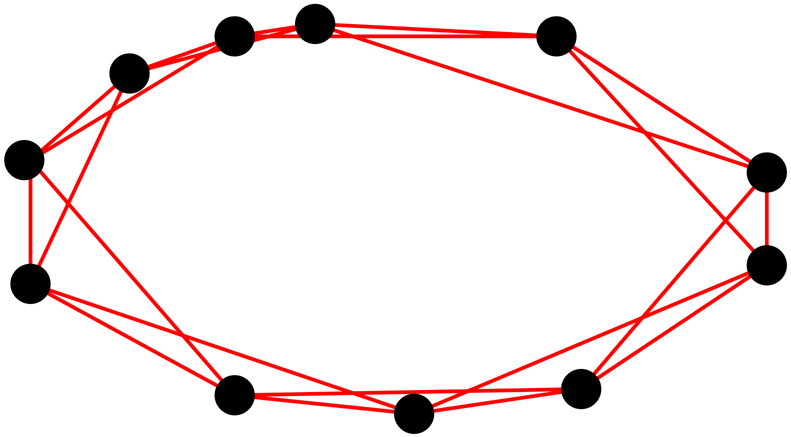
Basic fact:

$$\mathcal{S}(\mathbf{f}) = \sum_{i\sim j}(f_i - f_j)^2 = \frac{1}{2}\,\mathbf{f}^t\,\mathbf{L}\,\mathbf{f}$$

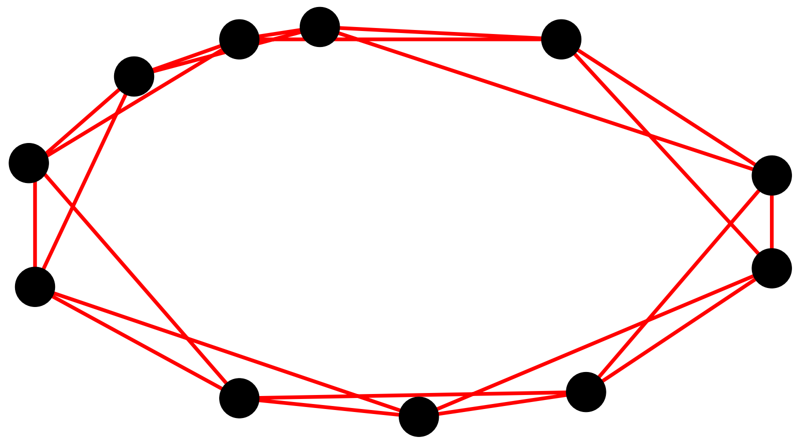
Algorithmic framework

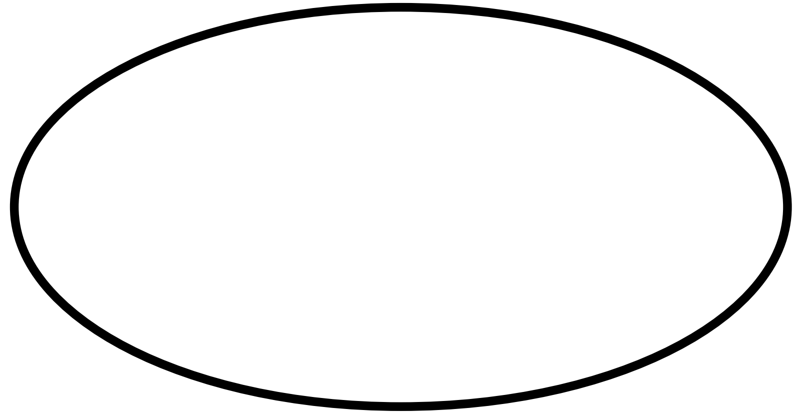


Algorithmic framework



Algorithmic framework

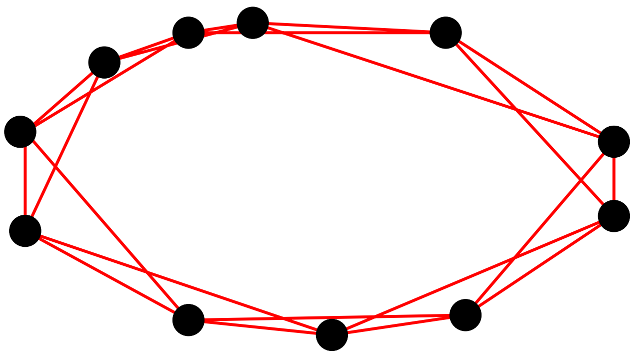




$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$$

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

$$\mathbf{f}^t\,\mathbf{L}\,\mathbf{f} = 2\sum_{i\sim j} e^{-\frac{\|x_i - x_j\|^2}{t}} (f_i - f_j)^2$$



$f : G \to \mathbb{R}$

Minimize $\sum_{i\sim j} w_{ij}(f_i - f_j)^2$

Preserve adjacency.

Solution: $Lf = \lambda f$ (slightly better $Lf = \lambda Df$)
Lowest eigenfunctions of L ($\tilde{L}$).

Laplacian Eigenmaps

Belkin Niyogi 01

Related work: LLE: Roweis, Saul 00; Isomap: Tenenbaum, De Silva, Langford 00

Hessian Eigenmaps: Donoho, Grimes, 03; Diffusion Maps: Coifman, et al, 04

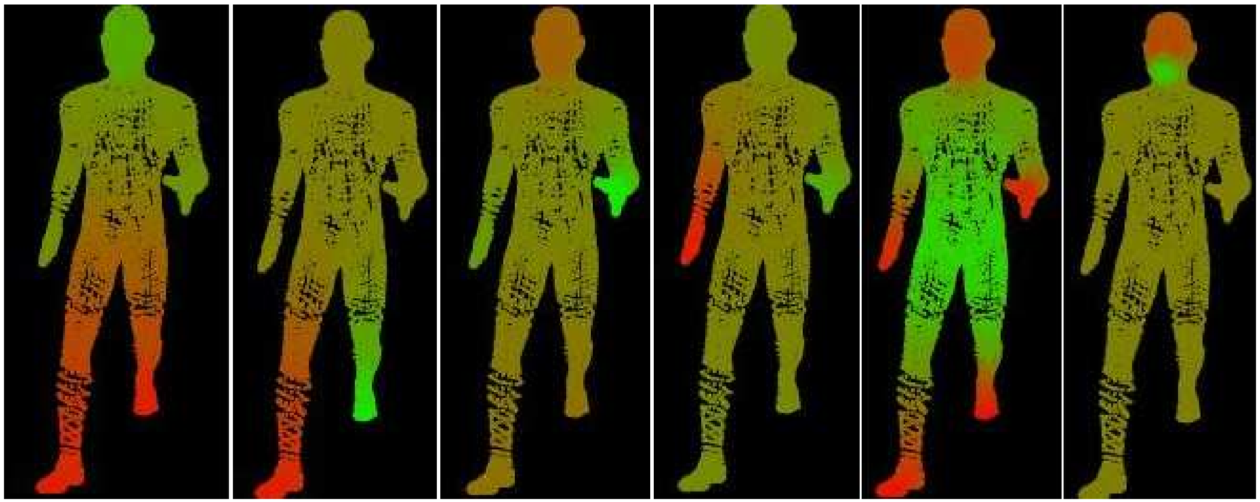
Laplacian Eigenmaps

- Visualizing spaces of digits and sounds.

Partiview, Ndaona, Surendran 04

- Machine vision: inferring joint angles.

Corazza, Andriacchi, Stanford Biomotion Lab, 05, Partiview, Surendran



Isometrically invariant representation. [link]

- Reinforcement Learning: value function approximation. Mahadevan, Maggioni, 05

Learning from labeled and unlabeled data.

- Unlabeled data is everywhere. Need to use it.
- Natural learning is semi-supervised.

Learning from labeled and unlabeled data.

- Unlabeled data is everywhere. Need to use it.
- Natural learning is semi-supervised.

Labeled data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \in \mathbb{R}^N \times \mathbb{R}$
Unlabeled data: $\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u} \in \mathbb{R}^N$

Need to reconstruct

$$f_{L,U} : \mathbb{R}^N \to \mathbb{R}$$

Estimate $f : \mathbb{R}^N \to \mathbb{R}$
Data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$

Regularized least squares (hinge loss for SVM):

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_K^2$$

fit to data + smoothness penalty

$\|f\|_K$ incorporates our smoothness assumptions.
Choice of $\|\ \|_K$ is important.

$$\textbf{Solve :} \qquad f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_K^2$$

$\|f\|_K$ is a Reproducing Kernel Hilbert Space norm with kernel $K(\mathbf{x}, \mathbf{y})$.

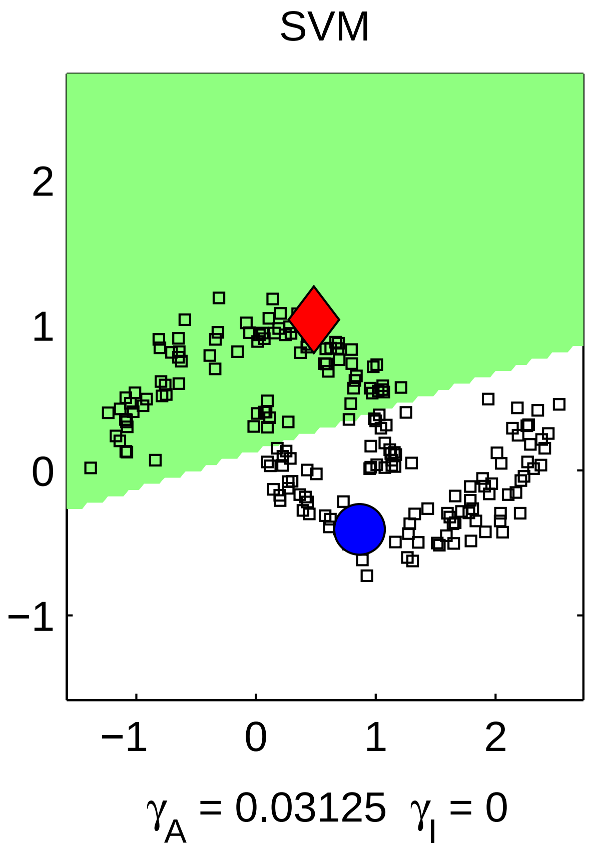
Can solve explicitly (via Representer theorem):

$$f^*(\cdot) = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}_i, \cdot)$$

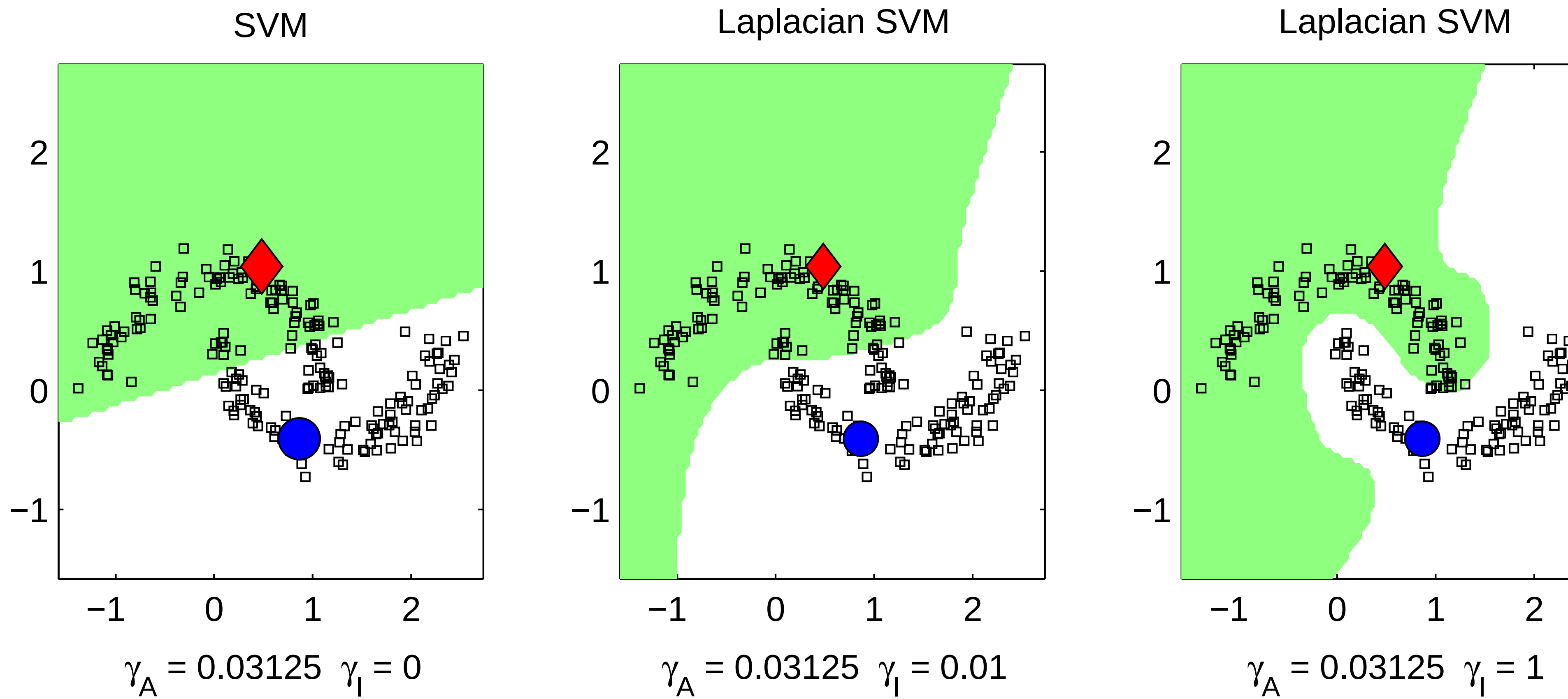
$$[\alpha_1, \ldots, \alpha_l]^t = (\mathbf{K} + \lambda I)^{-1} [y_1, \ldots, y_l]^t$$

$$(\mathbf{K})_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$$

Toy example



Toy example


SVM
Laplacian SVM
Laplacian SVM
2
1
0
−1
−1
0
1
2
γ_A = 0.03125 γ_I = 0
γ_A = 0.03125 γ_I = 0.01
γ_A = 0.03125 γ_I = 1

Manifold regularization

Estimate $f : \mathbb{R}^N \to \mathbb{R}$

Labeled data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$
Unlabeled data: $\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}$

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum (f(\mathbf{x}_i) - y_i)^2 + \lambda_A \|f\|_K^2 + \lambda_I \|f\|_I^2$$

fit to data + extrinsic smoothness + intrinsic smoothness

Empirical estimate:

$$\|f\|_I^2 = \frac{1}{(l+u)^2} [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{l+u})] \, L \, [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{l+u})]^t$$

Belkin Niyogi Sindhwani 04

Representer theorem (discrete case):

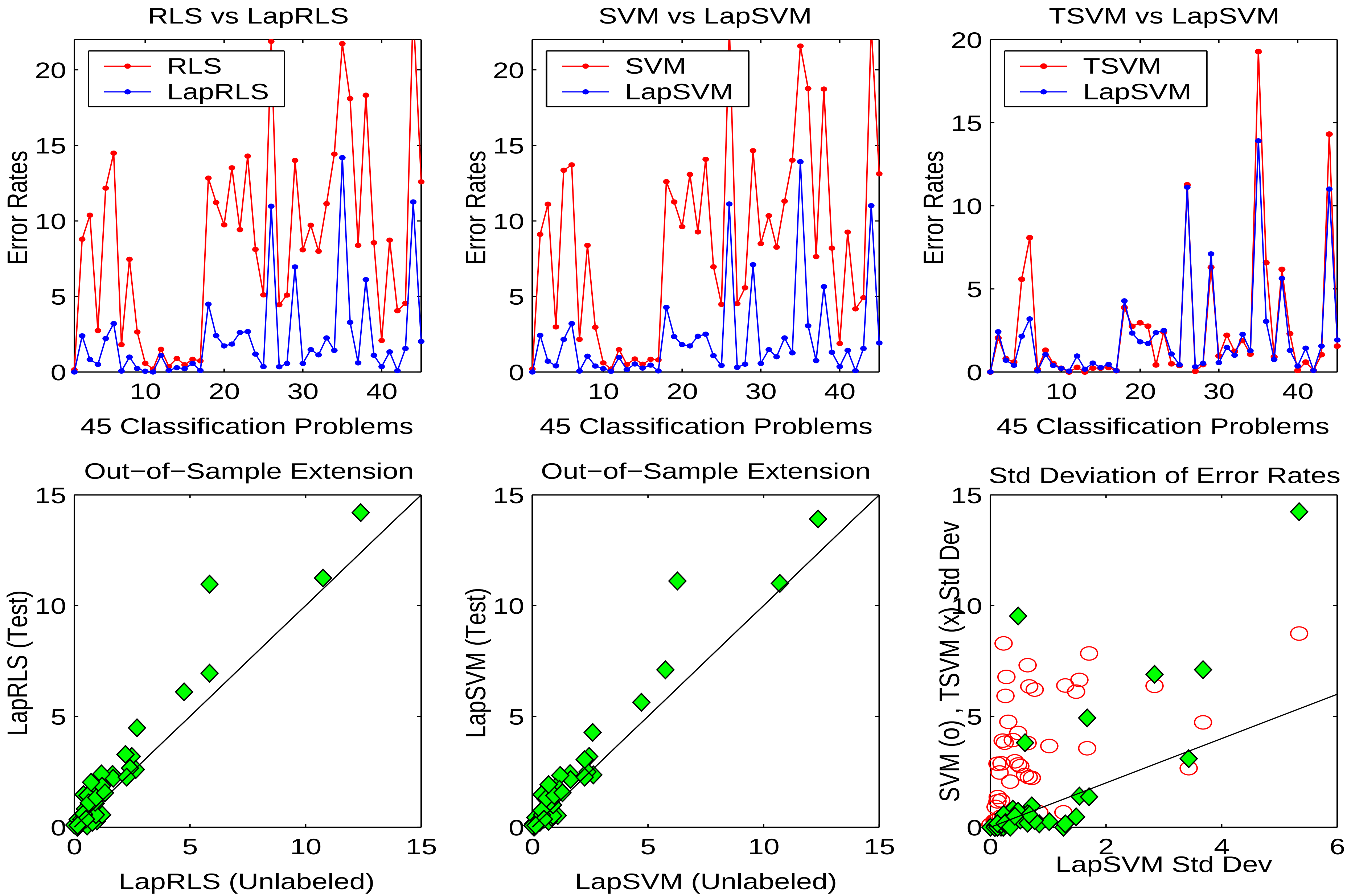
$$f^*(\cdot) = \sum_{i=1}^{l+u} \alpha_i K(\mathbf{x}_i, \cdot)$$

Explicit solution for quadratic loss:

$$\bar{\alpha} = \left(J\mathbf{K} + \lambda_A l I + \frac{\lambda_I l}{(u+l)^2}\mathbf{L}\mathbf{K}\right)^{-1} [y_1, \ldots, y_l, 0, \ldots, 0]^t$$

$$(\mathbf{K})_{ij} = K(\mathbf{x}_i, \mathbf{x}_j), \qquad J = diag\,(\underbrace{1, \ldots, 1}_{l}, \underbrace{0, \ldots, 0}_{u})$$

Experimental results: USPS



Experimental comparisons

Dataset $\rightarrow$ Algorithm $\downarrow$	g50c	Coil20	Uspst	mac-win	WebKB (link)	WebKB (page)	WebKB (page+link)
SVM (full labels)	3.82	0.0	3.35	2.32	6.3	6.5	1.0
SVM (l labels)	8.32	24.64	23.18	18.87	25.6	22.2	15.6
Graph-Reg	17.30	6.20	21.30	11.71	22.0	10.7	6.6
TSVM	6.87	26.26	26.46	7.44	**14.5**	**8.6**	7.8
Graph-density	8.32	6.43	16.92	10.48	-	-	-
∇TSVM	5.80	17.56	17.61	5.71	-	-	-
LDS	5.62	4.86	15.79	**5.13**	-	-	-
LapSVM	**5.44**	**3.66**	**12.67**	10.41	18.1	10.5	**6.4**

Key theoretical question

What is the connection between point-cloud Laplacian L and Laplace-Beltrami operator $\Delta_{\mathcal{M}}$?

Analysis of algorithms:

Eigenvectors of L $\quad \stackrel{?}{\longleftrightarrow} \quad$ Eigenfunctions of $\Delta_{\mathcal{M}}$

Theorem [convergence of eigenfunctions]

$$Eig[L_n^{t_n}] \to Eig[\Delta_{\mathcal{M}}]$$

(Convergence in probability)

number of data points $n \to \infty$
width fo the Gaussian $t_n \to 0$

Previous work. Point-wise convergence.

Belkin, 03 Belkin, Niyogi 05,06; Lafon Coifman 04,06;Hein Audibert Luxburg, 05; Gine Kolchinskii, 06

Convergence of eigenfunctions for a fixed$t$**:**

Kolchniskii Gine 00, Luxburg Belkin Bousquet 04

Heat equation in $\mathbb{R}^n$:

$u(x,t)$ – heat distribution at time t.
$u(x,0) = f(x)$ – initial distribution. $x \in \mathbb{R}^n, t \in \mathbb{R}$.

$$\Delta_{\mathbb{R}^n} u(x,t) = \frac{du}{dt}(x,t)$$

Solution – convolution with the heat kernel:

$$u(x,t) = (4\pi t)^{-\frac{n}{2}} \int_{\mathbb{R}^n} f(y) e^{-\frac{\|x-y\|^2}{4t}} dy$$

Proof idea (pointwise convergence)

Functional approximation:

Taking limit as $t \to 0$ and writing the derivative:

$$\Delta_{\mathbb{R}^n} f(x) = \frac{d}{dt}\left[(4\pi t)^{-\frac{n}{2}} \int_{\mathbb{R}^n} f(y) e^{-\frac{\|x-y\|^2}{4t}} dy\right]_0$$

Proof idea (pointwise convergence)

Functional approximation:

Taking limit as $t \to 0$ and writing the derivative:

$$\Delta_{\mathbb{R}^n} f(x) = \frac{d}{dt}\left[(4\pi t)^{-\frac{n}{2}} \int_{\mathbb{R}^n} f(y) e^{-\frac{\|x-y\|^2}{4t}} dy\right]_0$$

$$\Delta_{\mathbb{R}^n} f(x) \approx -\frac{1}{t}(4\pi t)^{-\frac{n}{2}}\left(f(x) - \int_{\mathbb{R}^n} f(y) e^{-\frac{\|x-y\|^2}{4t}} dy\right)$$

Proof idea (pointwise convergence)

Functional approximation:

Taking limit as $t \to 0$ and writing the derivative:

$$\Delta_{\mathbb{R}^n} f(x) = \frac{d}{dt}\left[(4\pi t)^{-\frac{n}{2}} \int_{\mathbb{R}^n} f(y) e^{-\frac{\|x-y\|^2}{4t}} dy\right]_0$$

$$\Delta_{\mathbb{R}^n} f(x) \approx -\frac{1}{t}(4\pi t)^{-\frac{n}{2}} \left(f(x) - \int_{\mathbb{R}^n} f(y) e^{-\frac{\|x-y\|^2}{4t}} dy\right)$$

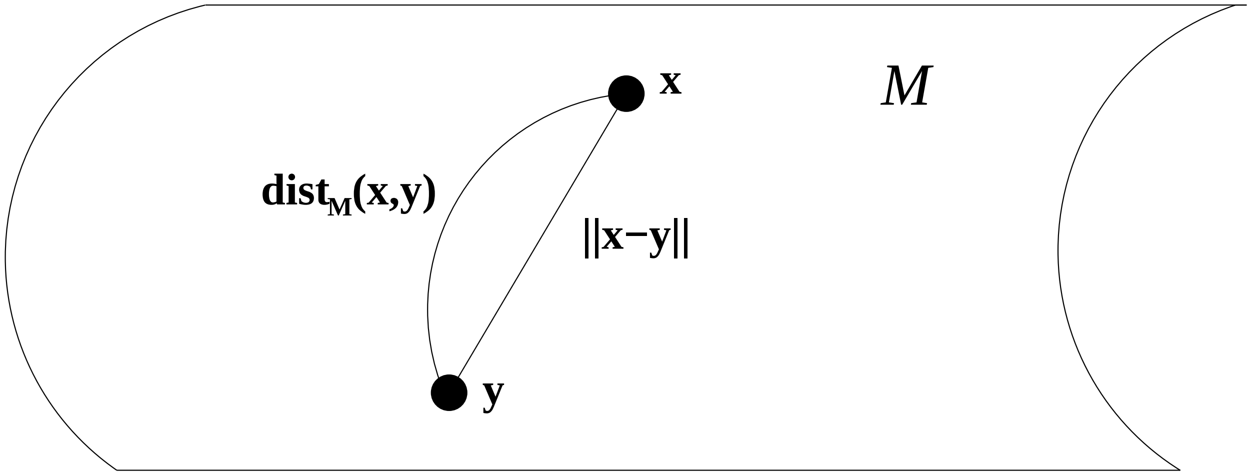
Empirical approximation:

Integral can be estimated from empirical data.

$$\Delta_{\mathbb{R}^n} f(x) \approx -\frac{1}{t}(4\pi t)^{-\frac{n}{2}} \left(f(x) - \sum_{x_i} f(x_i) e^{-\frac{\|x-x_i\|^2}{4t}}\right)$$

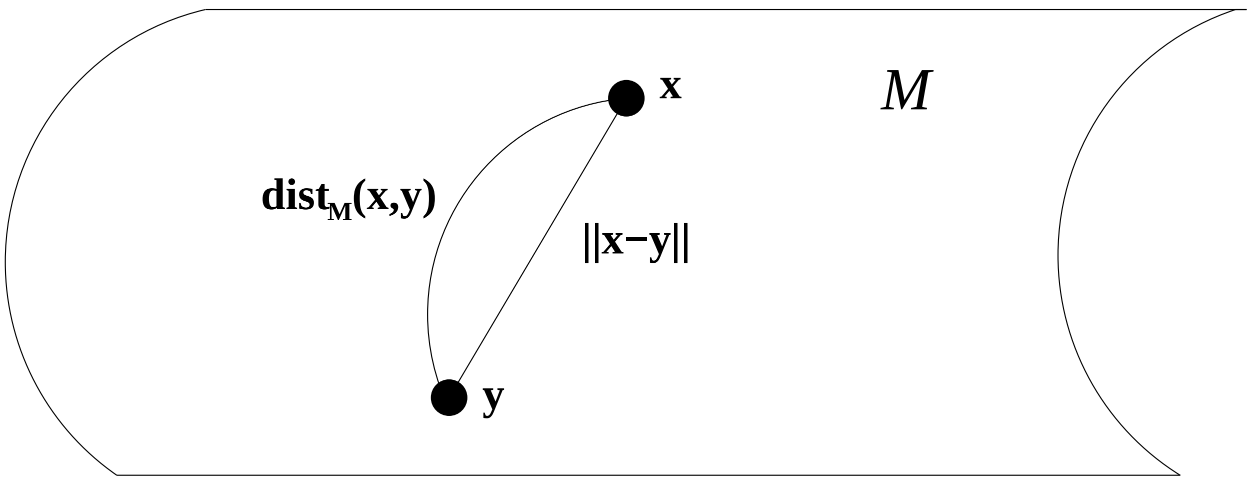
Some difficulties arise for manifolds:

- Do not know distances.
- Do not know the heat kernel.



Some difficulties arise for manifolds:

- Do not know distances.
- Do not know the heat kernel.



Careful analysis needed.

Let H^t be the heat operator.

$$H^t = \exp(-t\Delta_{\mathcal{M}})$$

L^t approximates $\frac{1-H^t}{t}$

Non-uniform convergence:

$$\frac{1 - H^t}{t} \not\to \Delta_{\mathcal{M}}$$

Convergence of eigenfunctions

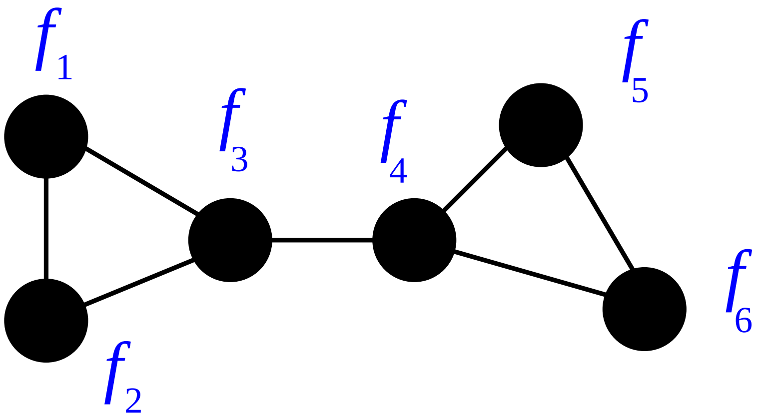
Observe that H^t has the same eigenfunctions as $\Delta_{\mathcal{M}}$.

Show that L^t is a relatively bounded and small perturbation of H^t.

$$\frac{\|(H^t - L^t)(f)\|_2}{\|H^t(f)\|_2} \ll 1$$

for small t.

Enough for convergence.



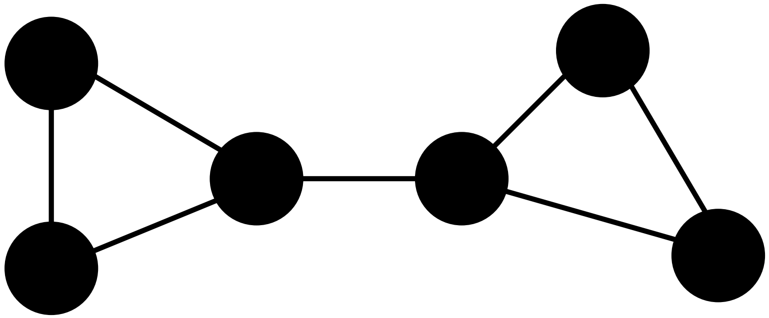
$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

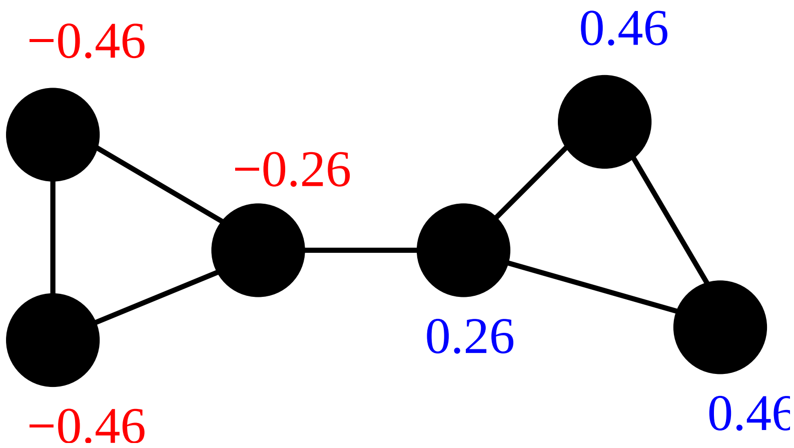
$$\underset{S}{\operatorname{argmin}} \sum_{i \in S,\ j \in V-S} w_{ij} = \underset{f_i \in \{-1,1\}}{\operatorname{argmin}} \sum_{i \sim j} (f_i - f_j)^2 = \frac{1}{8} \underset{f_i \in \{-1,1\}}{\operatorname{argmin}} \mathbf{f}^t \, \mathbf{L} \, \mathbf{f}$$

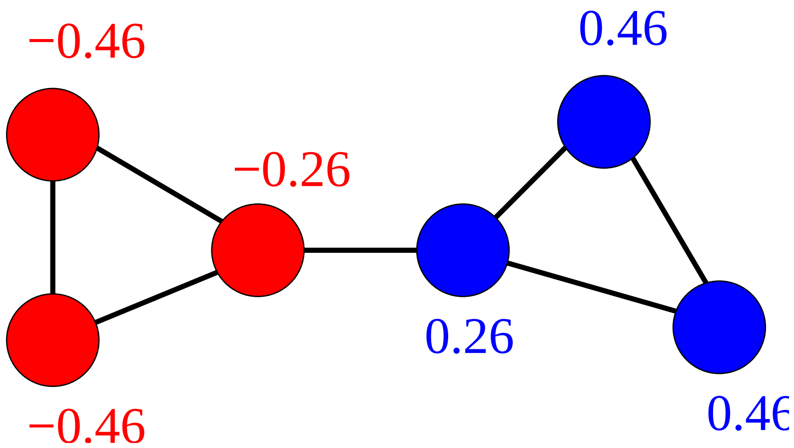
Relaxation gives eigenvectors.

$$\mathbf{L}v = \lambda v$$

Spectral clustering



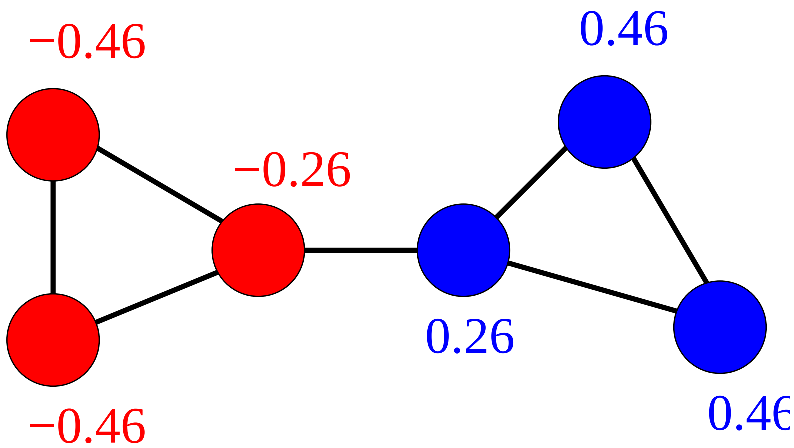

−0.46
0.46
−0.26
0.26
−0.46
0.46



$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

Unnormalized clustering:

$$L\mathbf{e_1} = \lambda_1 \mathbf{e_1} \qquad \mathbf{e_1} = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46]$$



$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

Unnormalized clustering:

$$L\mathbf{e_1} = \lambda_1 \mathbf{e_1} \qquad \mathbf{e_1} = [-0.46, -0.46, -0.26, 0.26, \ 0.46, \ 0.46]$$

Normalized clustering:

$$L\mathbf{e_1} = \lambda_1 D\mathbf{e_1} \qquad \mathbf{e_1} = [-0.31, -0.31, -0.18, 0.18, \ 0.31, \ 0.31]$$

Consistency of spectral clustering

Limit behavior of spectral clustering.

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \qquad\qquad n \to \infty$$

Sampled from probability distribution P on X.

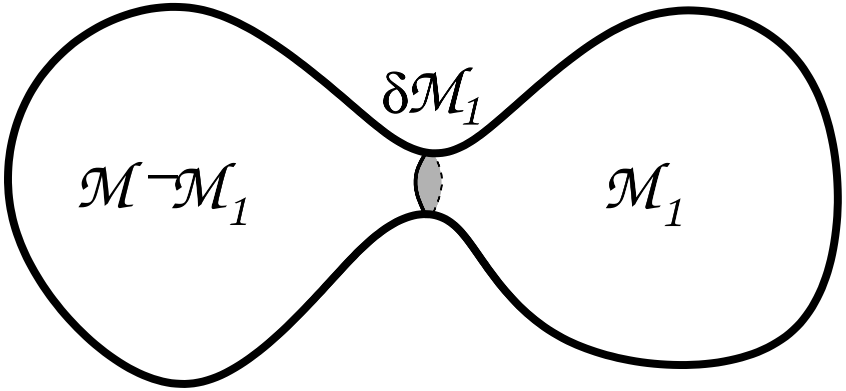
Theorem 1:
Normalized spectral clustering (bisectioning) is consistent.

Theorem 2:
Unnormalized spectral clustering may not converge depending on the spectrum of L and P.

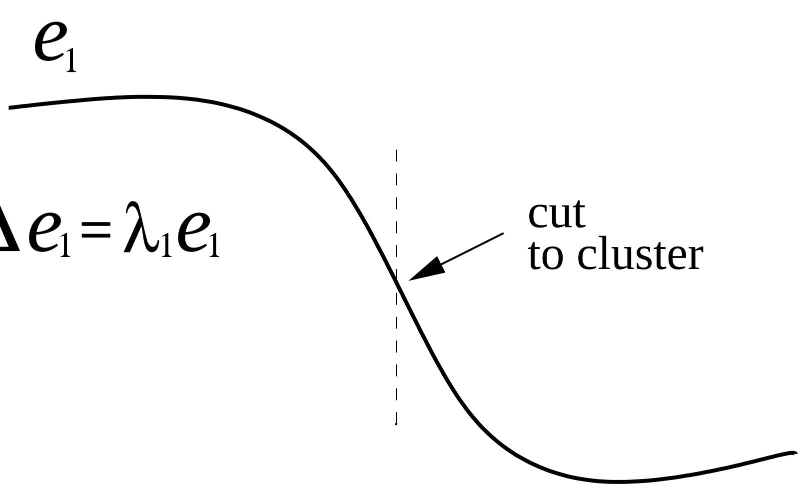
von Luxburg Belkin Bousquet 04

Continuous spectral clustering

Laplacian eigenfunction as a relaxation of the isoperimetric problem.



$$h = \inf \frac{\mathrm{vol}^{n-1}(\delta\mathcal{M}_1)}{\min\left(\mathrm{vol}^n(\mathcal{M}_1), \mathrm{vol}^n(\mathcal{M}-\mathcal{M}_1)\right)}$$



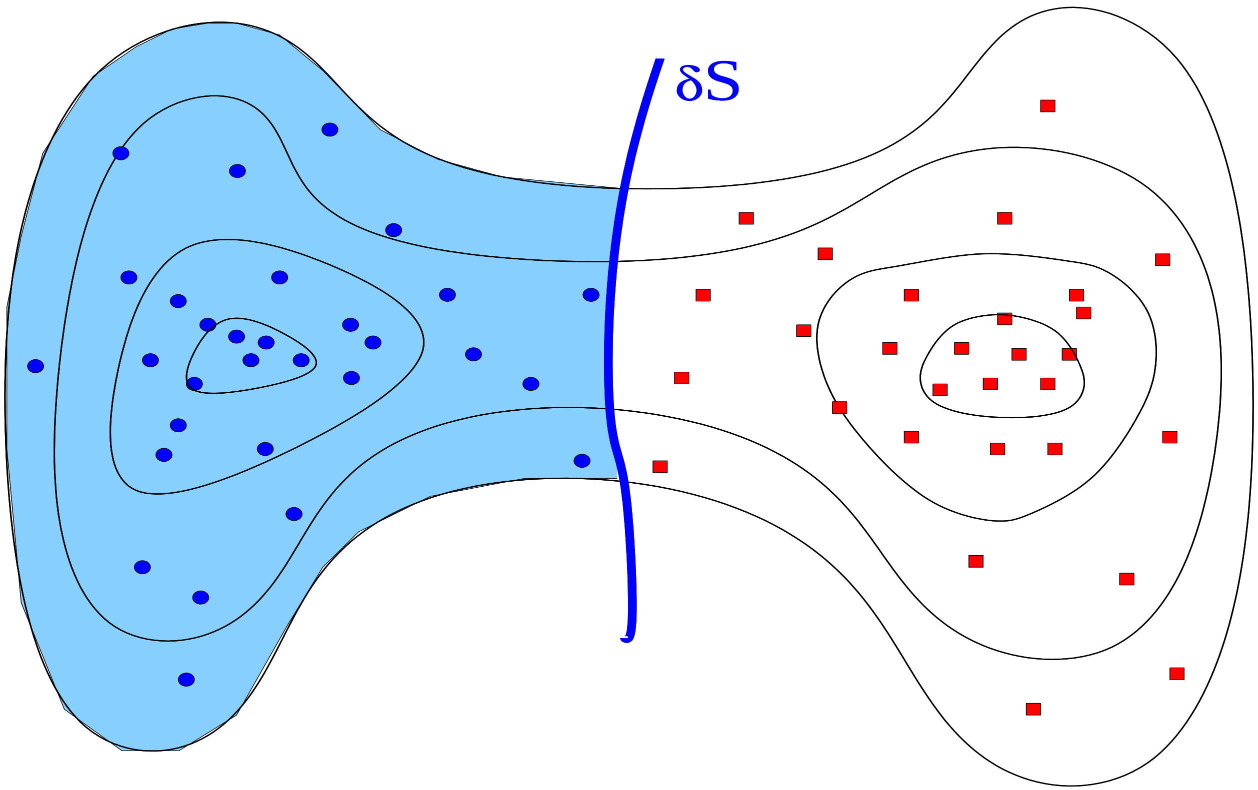
$$\Delta e_1 = \lambda_1 e_1$$

$$0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots$$

$$h \leq \frac{\sqrt{\lambda_1}}{2}$$

[Cheeger]

Estimating volumes of cuts



$$\sum_{i\in\text{blue}} \sum_{j\in\text{red}} \frac{w_{ij}}{\sqrt{d_j d_j}}$$

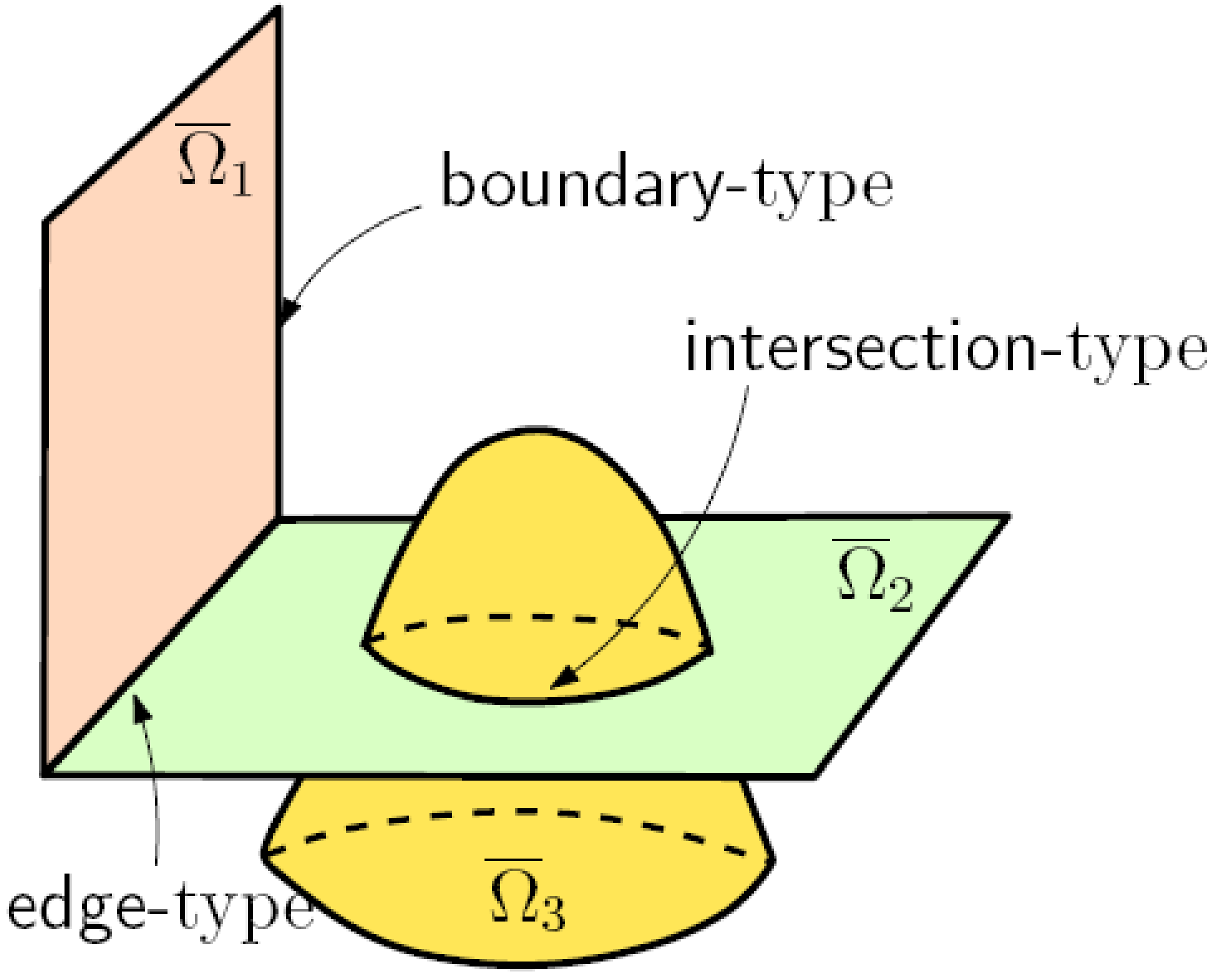
$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{4t}}$$

$$d_i = \sum_j w_{ij}$$

$$\mathrm{vol}(\delta S) \approx \frac{2}{N}\frac{1}{(4\pi t)^{n/2}}\sqrt{\frac{\pi}{t}}\,\mathbf{1}_S^t\, L\, \mathbf{1}_S$$

L is the normalized graph Laplacian and $\mathbf{1}_S$ is the indicator vector of points in S. (Narayanan Belkin Niyogi, 06)

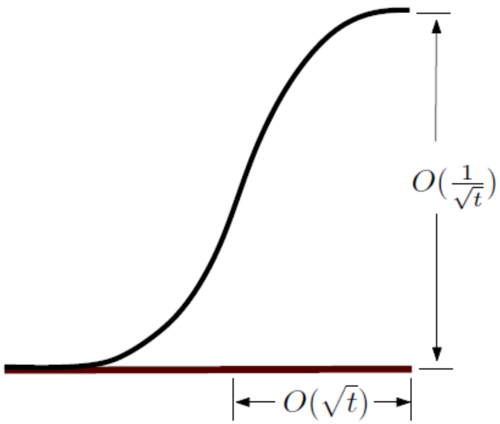
Singular manifolds.



Singular manifolds

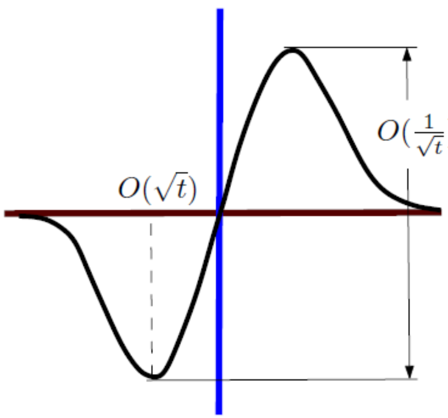
Operator scaling:

$$L_t f = \frac{1}{\sqrt{t}} \phi \, D_n$$



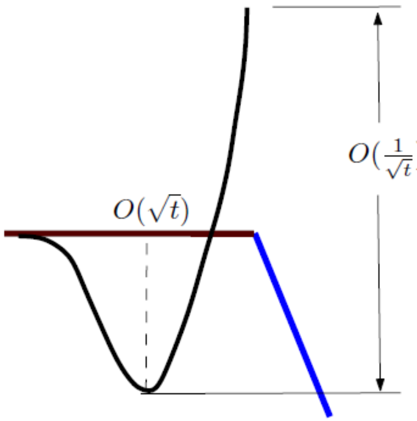
Boundary

$$\phi = e^{-r^2}$$



Intersection

$$\phi = r e^{-r^2}$$



Edge

$$\phi = e^{-r^2} + r e^{-r^2}$$

[Belkin, Que, Wang, Zhou 12]

Conclusion

1. Geometry controls many aspects of inference.

2. Our methods should adapt to geometry.
Graph-based representation of data is good at that.

3. Laplace operator – graph Laplacian is a useful tool
for various inferential tasks.