A Geometric Perspective on Machine Learning

Partha Niyogi

The University of Chicago

Collaborators: M. Belkin, X. He, H. Narayanan, V. Sindhwani, S. Smale, S. Weinberger

High Dimensional Data

When can we avoid the curse of dimensionality?

- Smoothness
- Sparsity
- *Geometry*

Manifold Learning

Learning when data $\sim \mathcal{M} \subset \mathbb{R}^N$

- Clustering: $\mathcal{M} \to \{1, \ldots, k\}$

  connected components, min cut

- Classification: $\mathcal{M} \to \{-1, +1\}$

  P on $\mathcal{M} \times \{-1, +1\}$

- Dimensionality Reduction: $f : \mathcal{M} \to \mathbb{R}^n \quad n << N$

- $\mathcal{M}$ unknown: what can you learn about $\mathcal{M}$ from data?

  e.g. dimensionality, connected components

  holes, handles, homology

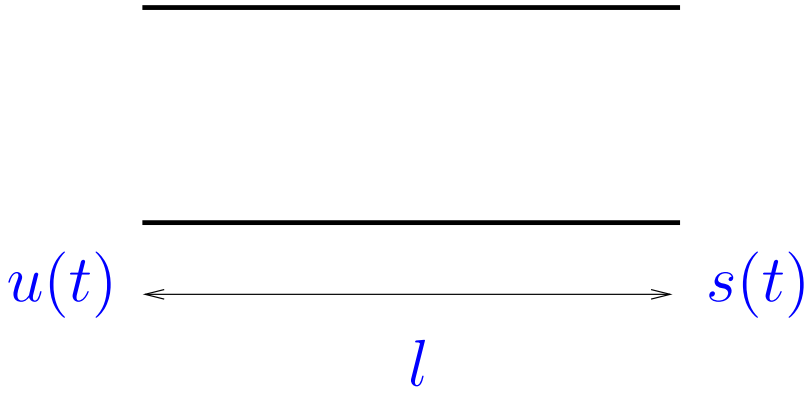
  curvature, geodesics

Take Home Message

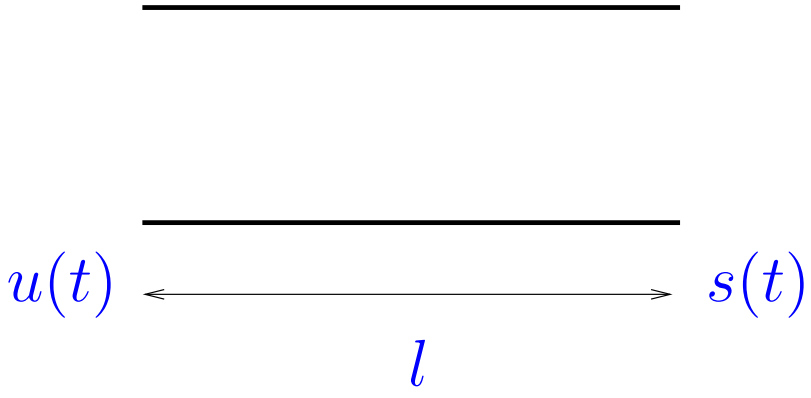
- Geometrically motivated approach to learning

  nonlinear, nonparametric, high dimensions
- Emphasize the role of the Laplacian and Heat Kernel
  - Semi-supervised regression and classification
  - Clustering
  - Homology

An Acoustic Example



An Acoustic Example



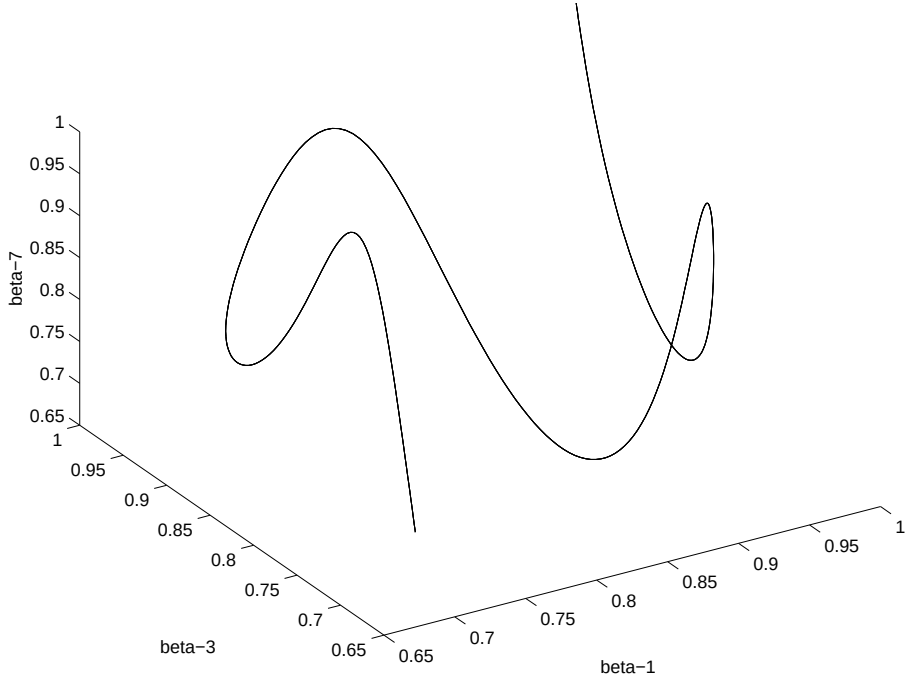
One Dimensional Air Flow

(i) $\frac{\partial V}{\partial x} = -\frac{A}{\rho c^2}\frac{\partial P}{\partial t}$ (ii) $\frac{\partial P}{\partial x} = -\frac{\rho}{A}\frac{\partial V}{\partial t}$

$V(x,t) =$ volume velocity

$P(x,t) =$ pressure

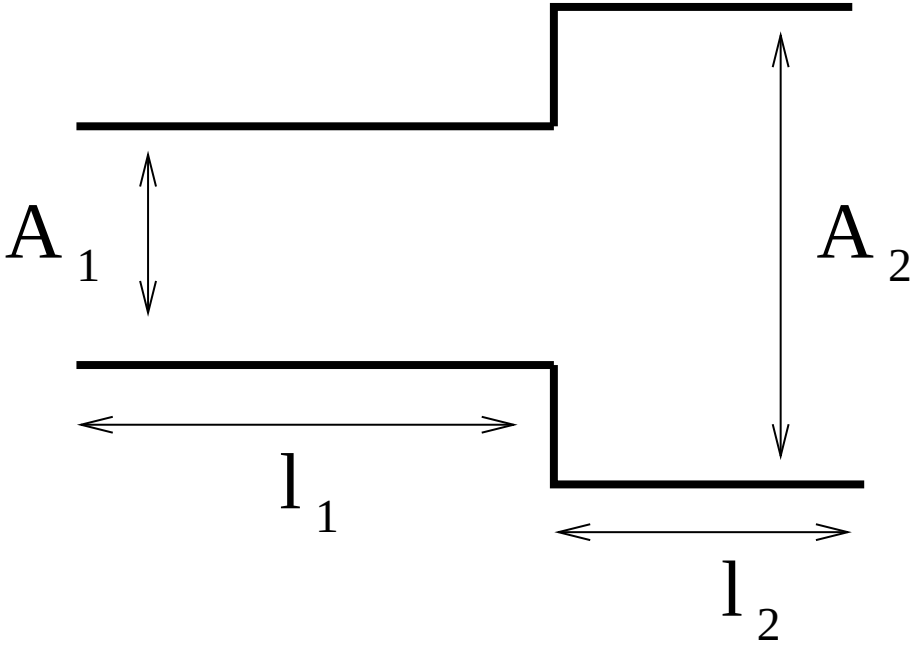
Solutions



$$u(t) = \sum_{n=1}^{\infty} \alpha_n \sin(n\omega_0 t) \in l_2$$

$$s(t) = \sum_{n=1}^{\infty} \beta_n \sin(n\omega_0 t) \in l_2$$

Acoustic Phonetics



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

Jansen and Niyogi (2005)

Formal Justification

- Speech

  speech $\in l_2$ generated by vocal tract

  Jansen and Niyogi (2005)

- Vision

  group actions on object leading to different images

  Donoho and Grimes (2004)

- Robotics

  configuration spaces in joint movements

- Graphics

Manifold + Noise may be generic model in high dimensions.

Pattern Recognition

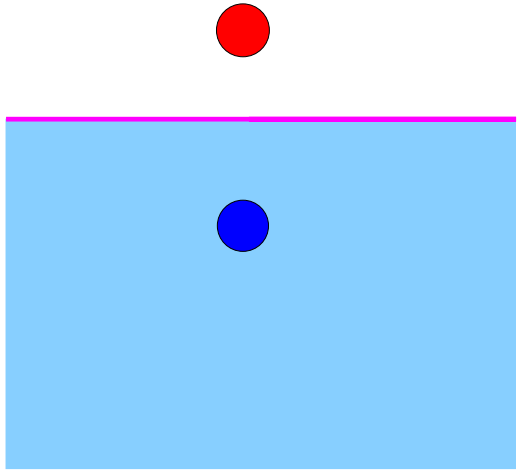
P on $X \times Y$ $\qquad\qquad$ $X = \mathbb{R}^N; Y = \{0, 1\}, \mathbb{R}$

(x_i, y_i) labeled examples

find $f : X \to Y$ $\qquad$ Ill Posed

Simplicity

Simplicity



Regularization Principle

$$f = \arg\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma \|f\|_K^2$$

Splines

Ridge Regression

SVM

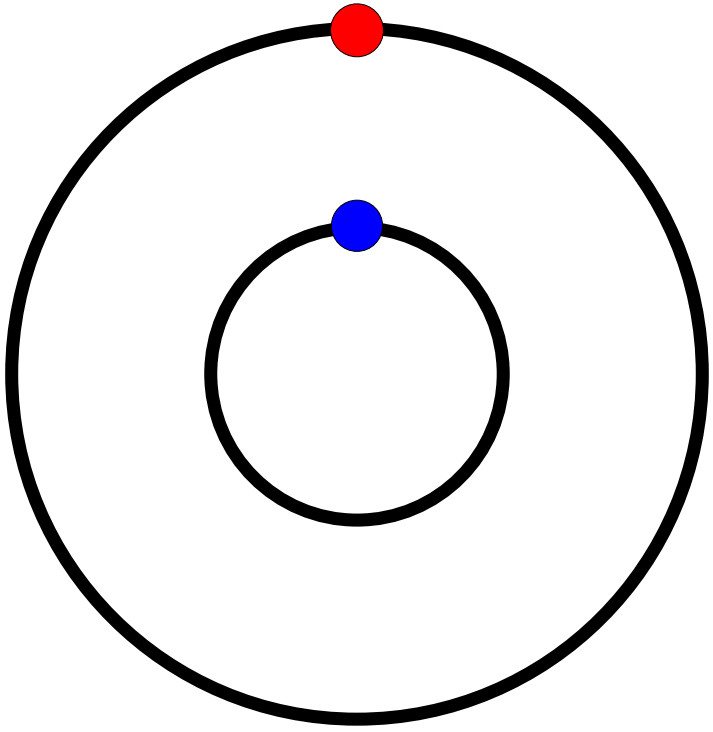
- $K : X \times X \to \mathbb{R}$ is a p.d. kernel

  e.g. $e^{-\frac{\|x-y\|^2}{\sigma^2}}$, $(1 + x \cdot y)^d$, etc.

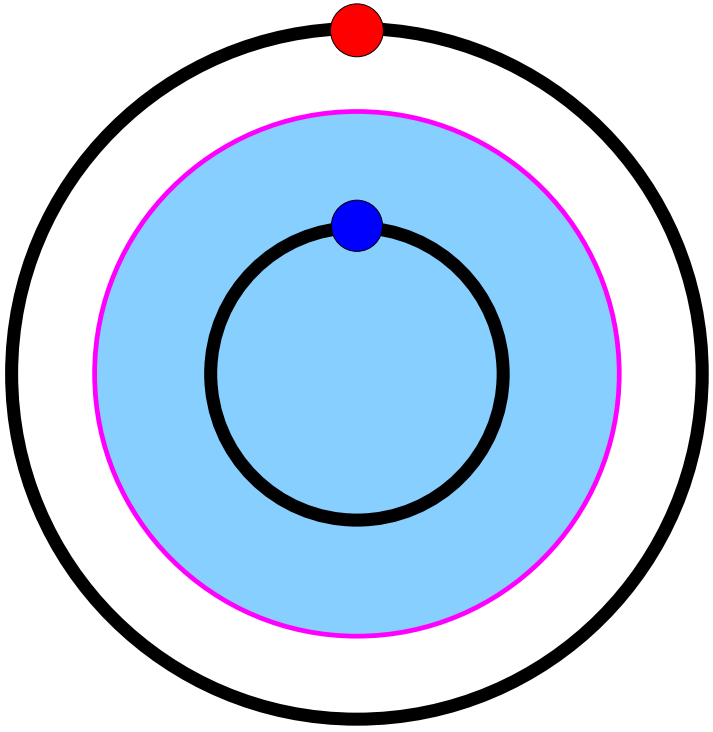
- H_K is a corresponding RKHS

  e.g., certain *Sobolev* spaces, polynomial families, etc.

Simplicity is Relative



Simplicity is Relative

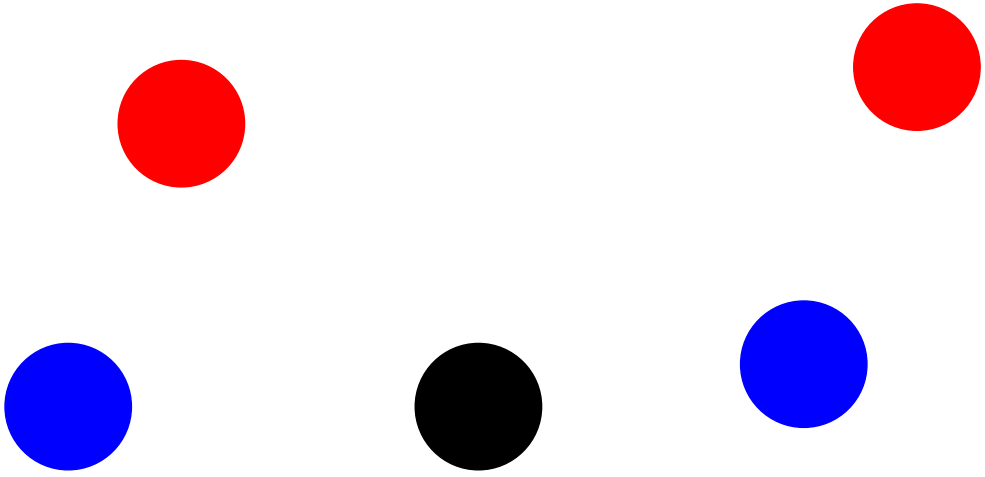


Intuitions

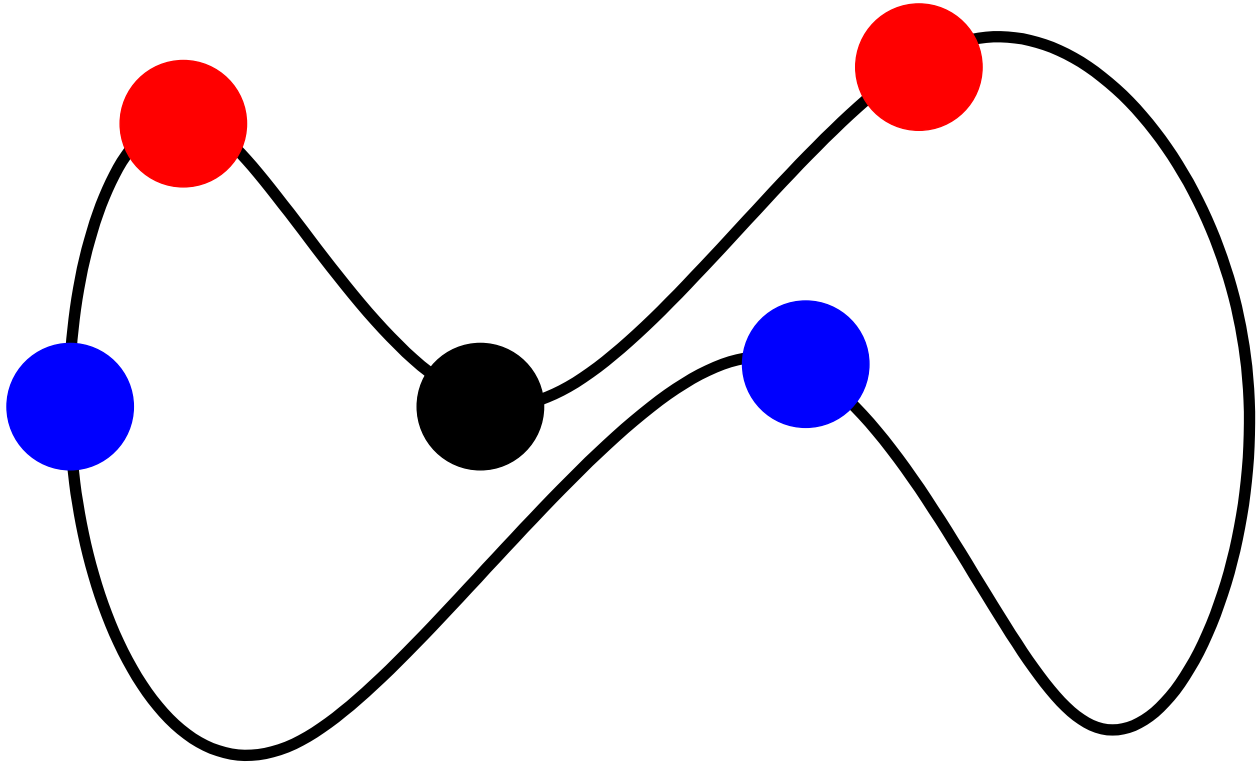
- supp P_X has manifold structure
- *geodesic* distance v/s *ambient* distance
- geometric structure of data should be incorporated
- f versus $f_{\mathcal{M}}$

Geometry and similarity

Geometry and similarity

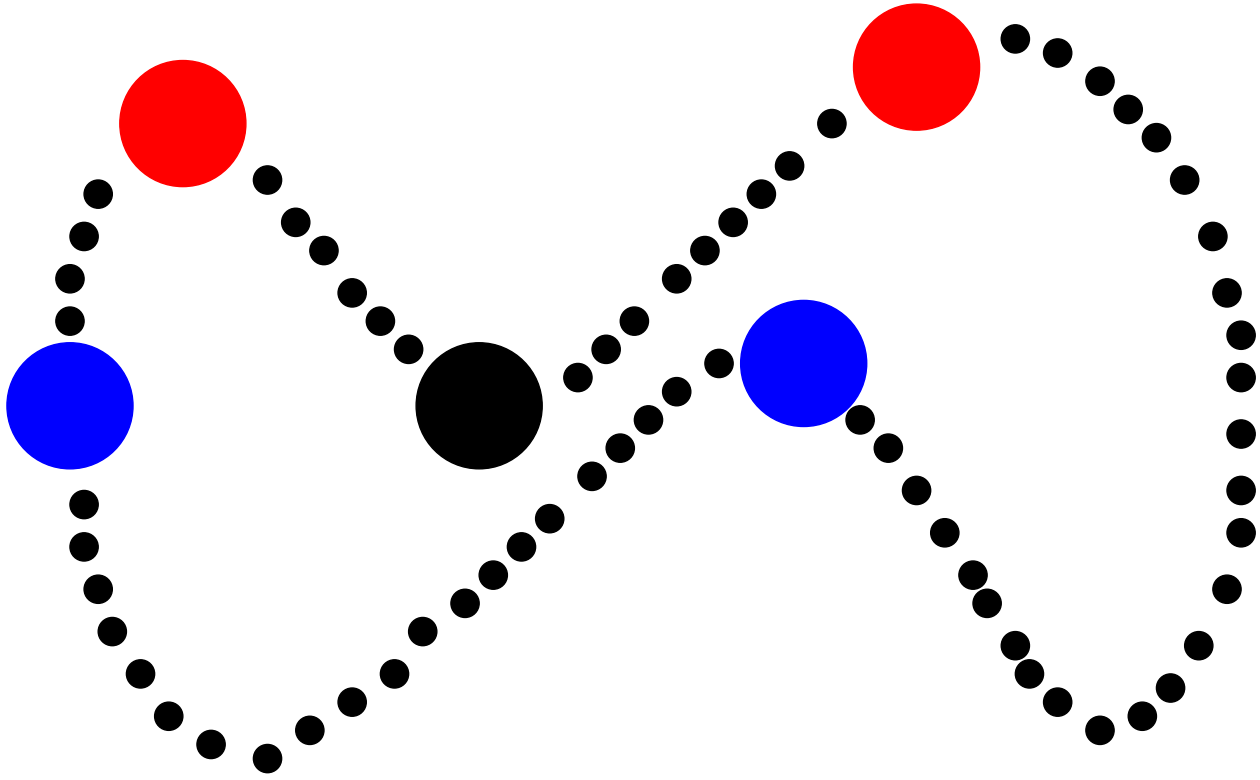


Geometry and similarity



Geometry is important.

Geometry and similarity



Geometry is important.
Unlabeled data to estimate geometry.

Manifold Regularization

$$\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2$$

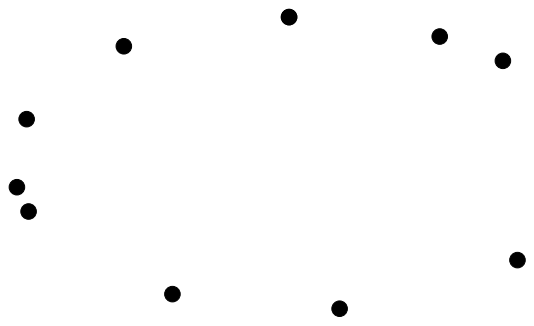
$$\|f\|_I^2 = \begin{cases} \text{Laplacian} & \int \langle \text{grad}_{\mathcal{M}} f, \text{grad}_{\mathcal{M}} f \rangle = \int f \Delta_{\mathcal{M}} f \\ \text{Iterated Laplacian} & \int f \Delta_{\mathcal{M}}{}^i f \\ \text{Heat kernel} & \mathrm{e}^{-\Delta_{\mathcal{M}} t} \\ \text{Differential Operator} & \int f(Df) \end{cases}$$

Representer Theorem: $f = \sum_{i=1}^{n} \alpha_i K(x, x_i) + \int_{\mathcal{M}} \alpha(y) K(x, y)$

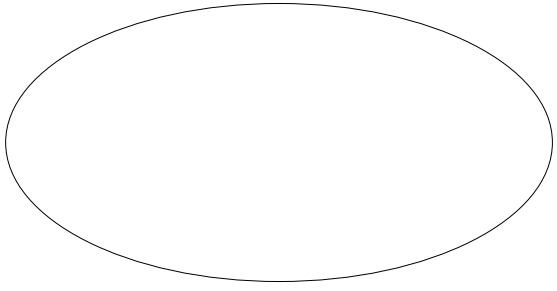
Belkin, Niyogi, Sindhwani (2004)

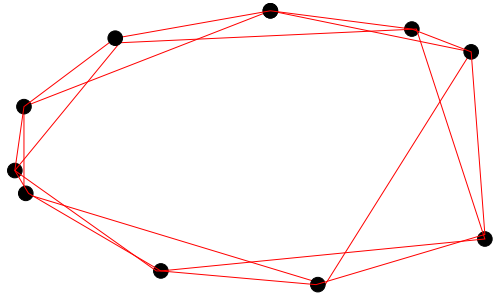
Approximating $\|f\|_I^2$

$\mathcal{M}$ is unknown but $x_1 \dots x_M \in \mathcal{M}$



$$\|f\|_I^2 = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle \quad \approx \quad \sum_{i \sim j} W_{ij} (f(x_i) - f(x_j))^2$$





Manifolds and Graphs

$$\mathcal{M} \approx G = (V, E)$$

$$e_{ij} \in E \text{ if } \|x_i - x_j\| < \epsilon$$

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

$$\Delta_{\mathcal{M}} \approx L = D - W$$

$$\int \langle \text{grad} f, \text{grad} f \rangle \approx \sum_{i,j} W_{ij} (f(x_i) - f(x_j))^2$$

$$\int f(\Delta f) \approx \mathbf{f}^T L \mathbf{f}$$

Manifold Regularization

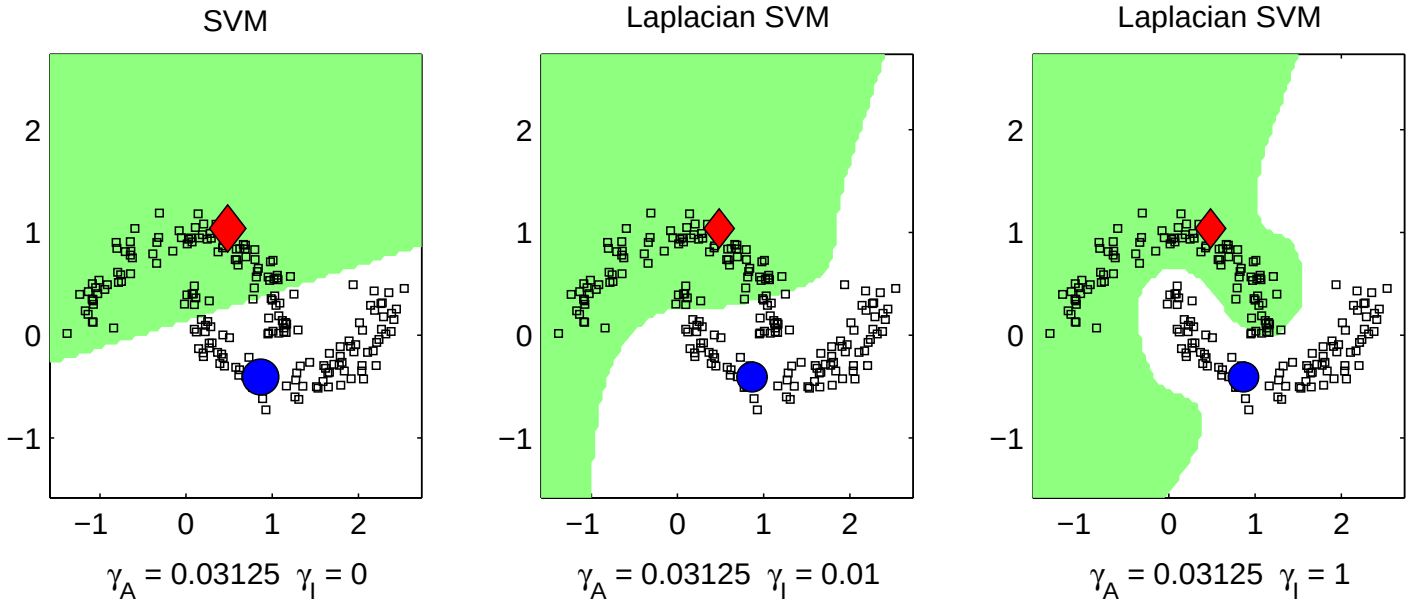
$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \gamma_A \|f\|_K^2 + \gamma_I \sum_{i\sim j} W_{ij}(f(x_i) - f(x_j))^2$$

Representer Theorem: $f_{opt} = \sum_{i=1}^{n+m} \alpha_i K(x, x_i)$

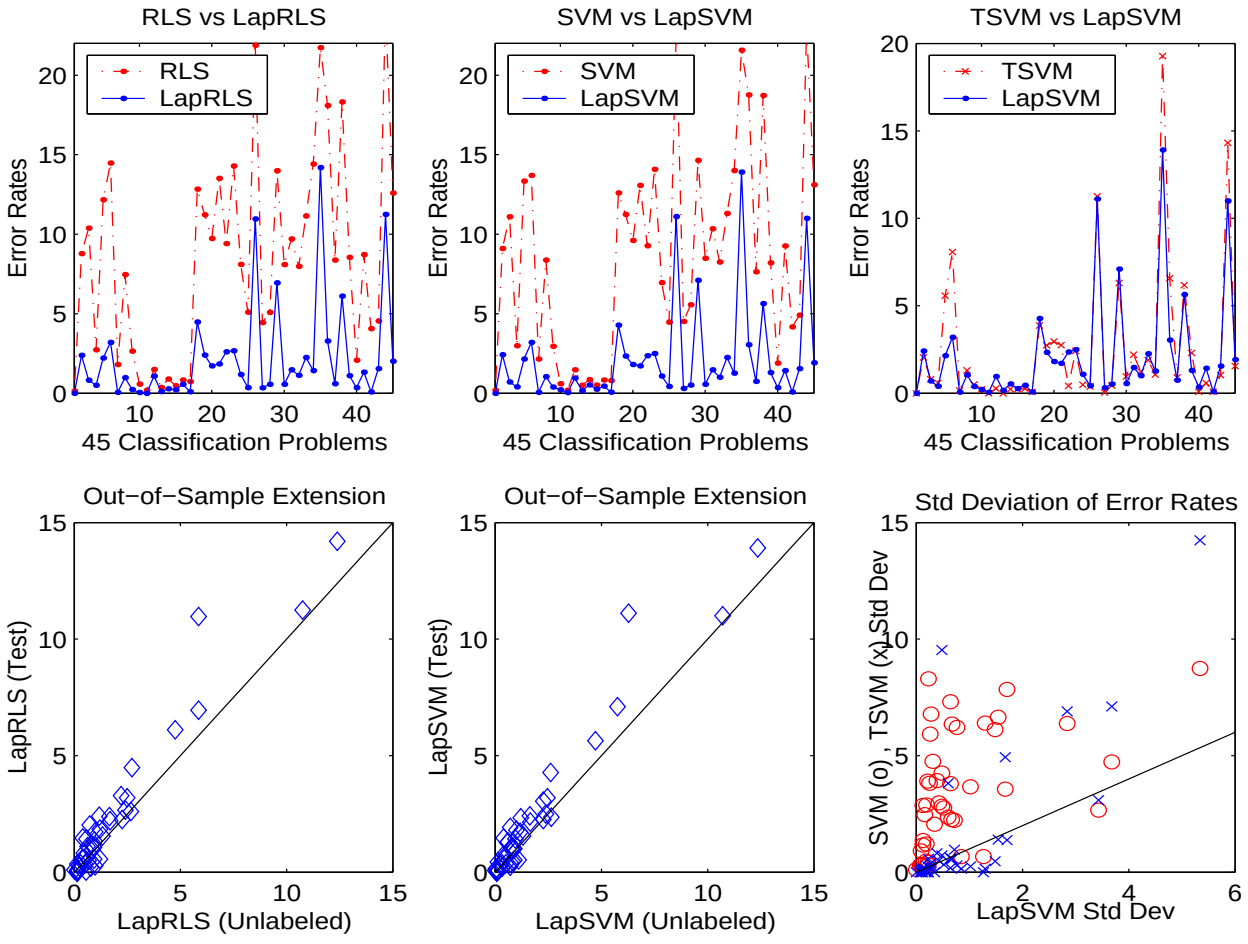
$V(f(x), y) = (f(x) - y)^2$: Least squares

$V(f(x), y) = (1 - yf(x))_+$: Hinge loss (Support Vector Machines)

Ambient and Intrinsic Regularization



Experimental Results: USPS



Experimental comparisons

Dataset → Algorithm ↓	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
RLS (full labels)	3.82	0.0	2.49	2.21	5.6	6.0	2.2
SVM (l labels)	8.32	24.64	23.18	18.87	25.6	22.2	15.6
RLS (l labels)	8.28	25.39	22.90	18.81	28.0	28.4	21.7
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**
LapRLS	**5.18**	**3.36**	**12.69**	10.01	19.2	11.0	6.9
LapSVM$_{joint}$	-	-	-	-	**5.7**	**6.7**	**6.4**
LapRLS$_{joint}$	-	-	-	-	**5.6**	**8.0**	**5.8**

Convergence Theorem

with prob. $> 1 - \delta$

$$|E_{\gamma_A,\gamma_I,n} - E_{opt}| \leq C + \gamma_A \|f_{opt}\|_K^2 + \gamma_I \int_{\mathcal{M}} f_{opt}(\Delta^l f_{opt})$$

$$C = \frac{4}{\beta^{3/2}} \sqrt{\frac{1}{n} \log(\frac{2}{\delta})}$$

$$\beta^2 = \frac{\gamma_A}{\kappa^2} + \frac{\gamma_I}{\mu^2}$$

$$\kappa^2 = \sup_{x \in X} K(x,x) \qquad \mu^2 = \sup_{x \in \mathcal{M}} \sum_i (\frac{1}{\lambda_i})^l \phi_i^2(x)$$

Graph and Manifold Laplacian

Fix $f : X \to \mathbb{R}$.
Fix $x \in \mathcal{M}$

$$(L_n f) = \sum_j (f(x) - f(x_j)) e^{-\frac{\|x - x_j\|^2}{4t_n}}$$

Put $t_n = n^{-k-2-\alpha}$, where $\alpha > 0$

with prob. 1, $$\lim_{n \to \infty} \frac{(4\pi t_n)^{-\frac{k+1}{2}}}{n} (L_n f)|_x = \Delta_{\mathcal{M}} f|_x$$

Belkin (2003), Belkin and Niyogi (2004,2005)

also Lafon (2004), Coifman et al,Hein, Gine and Koltchinski

Remarks on Noise

1. Arbitrary probability distribution on the manifold: convergence to weighted Laplacian.
2. Noise off the manifold:

$$\mu = \mu_{\mathcal{M}} + \mu_{\mathbb{R}^N}$$

3. Noise off the manifold:

$$z = x + \eta \ (\sim N(0, \sigma^2 I))$$

We have

$$\lim_{t \to 0} \lim_{\sigma \to 0} L^{t,\sigma} f(x) = \Delta f(x)$$

Important Issues

- How to handle noise theoretically and practically?
- How to choose the graph neighborhood correctly?
- How often do manifolds arise in natural data? What is the right metric on these manifolds?
- What are other ways in which one might utilize the geometry of natural distributions?
- Identify real problems where this approach can make a difference.
- Complexity estimates and provably correct algorithms rather than heuristics.

Connections and Implications

- Clustering

  sparse cuts, Cheeger constants, Laplacians

  (Narayanan, Belkin, Niyogi, 2006)

- Homology

  combinatorial Laplacian, simplicial complex

  (Niyogi, Smale, Weinberger, 2004)

- Volume estimation

  heat flow based algorithms

  (Belkin, Narayanan, Niyogi, 2006)

- Solving PDE's

  random meshes in contrast to triangulations

- Random Matrices and Graphs

  results on spectra

Random Graphs and Matrices

Given $x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$

$$W_{ij} = \frac{1}{t(4\pi t)^{d/2}} e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$Eig[W_{ij}] = Eig[L_n^{t_n}] \rightarrow Eig[\mathcal{L}_{\mathcal{M}}]$$

Belkin Niyogi 06

(Patodi, Dodziuk: triangulated manifolds)

Semi-supervised Learning

p on $X \times \mathbb{R}$

$m_p(x) = E[y|x]$ Regression Fn.

$S = \{(x_i, y_i)|i = 1, \ldots, n\}$

$A(S)$ Learner's output

Rate

$$E_{S\sim p}\|A(S) - m_p\|^2_{L^2(p_X)}$$

For a class $\mathcal{P}$

$$R(n) = \inf_A \sup_{p\in\mathcal{P}} E_{S\sim p}\|A(S) - m_p\|^2_{L^2(p_X)}$$

Additional Structure

$$\mathcal{P}_{\mathcal{M}} = \{p \in \mathcal{P} | p_X \text{ supported on } \mathcal{M}\}$$

Then

$$\mathcal{P} = \cup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}}$$

Comparison Theorem

$$R(n) = \inf_{A} \sup_{\mathcal{M}} \sup_{p \in \mathcal{P}_{\mathcal{M}}} E_{S \sim p} \| A(S) - m_p \|^2_{L^2(p_X)}$$

$$Q(n) = \sup_{\mathcal{M}} \inf_{A} \sup_{p \in \mathcal{P}_{\mathcal{M}}} E_{S \sim p} \| A(S) - m_p \|^2_{L^2(p_X)}$$

[Theorem]

$$Q(n) = O(\frac{1}{\sqrt{n}}) \text{ but } R(n) = \Omega(1)$$

Learning Homology

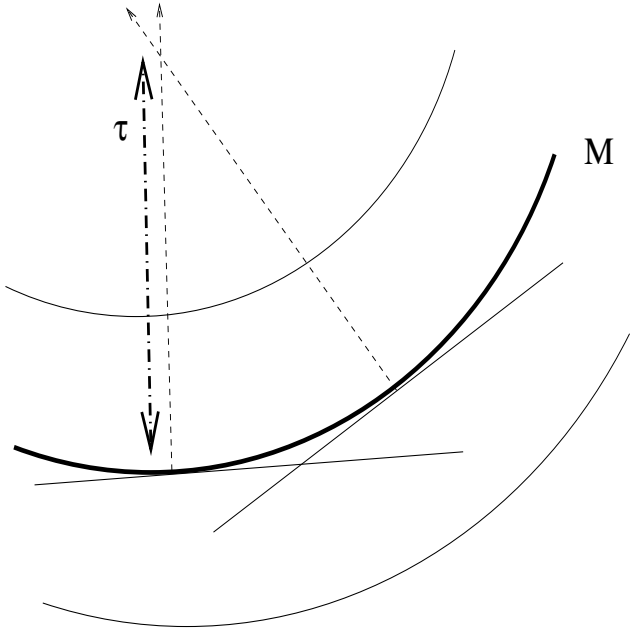
$$x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$$

Can you learn qualitative features of $\mathcal{M}$?

- Can you tell a torus from a sphere?
- Can you tell how many connected components?
- Can you tell the dimension of $\mathcal{M}$?

(e.g. Carlsson, Zamorodian, Edelsbrunner, Guibas, Oudot, Lieutier, Chazal, Dey, Amenta,Choi, Cohen-Steiner, de Silva etc.)

Well Conditioned Submanifolds



Tubular Neighborhood

Condition No. $\frac{1}{\tau}$

Min. distance to *medial axis*

Euclidean and Geodesic distance

$\mathcal{M} \subset \mathbb{R}^N$ condition $\sim \tau$

$p, q \in \mathcal{M}$ where $||p - q||_{\mathbb{R}^N} = d$.

For all $d \leq \frac{\tau}{2}$,

$$d_{\mathcal{M}}(p, q) \leq \tau - \tau\sqrt{1 - \frac{2d}{\tau}}$$

In fact, Second Fundamental Form Bounded by $\frac{1}{\tau}$

Homology

$$x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$$

$$U = \cup_{i=1}^{n} B_\epsilon(x_i)$$

If ϵ well chosen, then U deformation retracts to $\mathcal{M}$.

Homology of U is constructed using the *nerve* of U and agrees with the homology of $\mathcal{M}$.

Theorem

$\mathcal{M} \subset \mathbb{R}^d$ with cond. no. τ

$\bar{x} = \{x_1, \ldots, x_n\} \sim$ uniformly sampled i.i.d.

$0 < \epsilon < \frac{\tau}{2}$ $\qquad\qquad$ $\beta = \frac{vol(\mathcal{M})}{(\sin^{-1}(\epsilon/2\tau))^k vol(B_{\epsilon/8})}$

Let $U = \cup_{x \in \bar{x}} B_\epsilon(x)$

$$n > \beta(\log(\beta) + \log(\frac{1}{\delta}))$$

with prob. $> 1 - \delta$,
homology of U equals the homology of $\mathcal{M}$

(Niyogi, Smale, Weinberger, 2004)

A Data-derived complex

$$x_1, \ldots, x_n \in \mathbb{R}^d$$

Pick $\epsilon > 0$ and balls $B_\epsilon(x_i)$

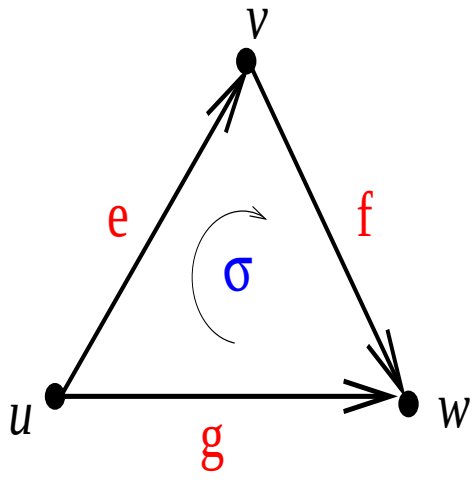
Put j-face for every $(i_0, \ldots, i_j)$ such that

$$\cap_{m=0}^{j} B_\epsilon(x_{i_m}) \neq \phi$$

Chains and the Combinatorial Laplacian

j chain is a formal sum $\sum_\sigma \alpha_\sigma \sigma$

C_j is the vector space of j-chains



$$\partial_j : C_j \to C_{j-1}$$

$$\partial_j^* : C_{j-1} \to C_j$$

$$\Delta_j = \partial_j^* \partial_j + \partial_{j+1} \partial_{j+1}^*$$

Noise

P on $\mathbb{R}^D$

such that

$P(x, y) = P(x)P(y|x)$ where $x \in \mathcal{M}, y \in N_x$

$$a \leq P(x)$$

$$P(y|x) = \sigma^2 I_{D-d}$$

Small Noise

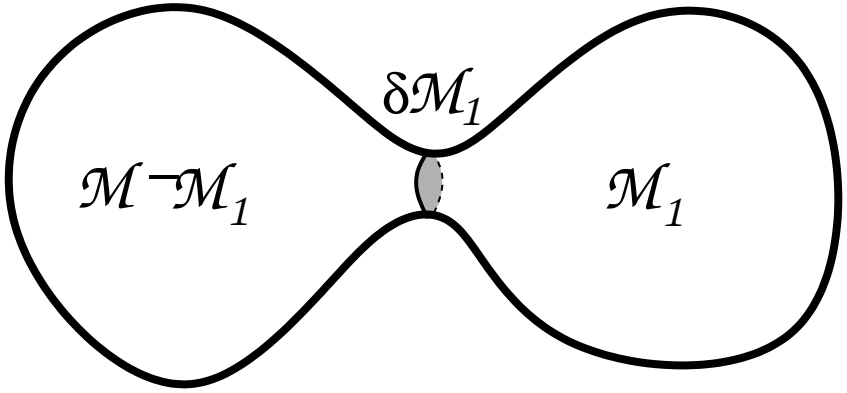
$$\sqrt{D-d}\sigma \leq c\tau$$

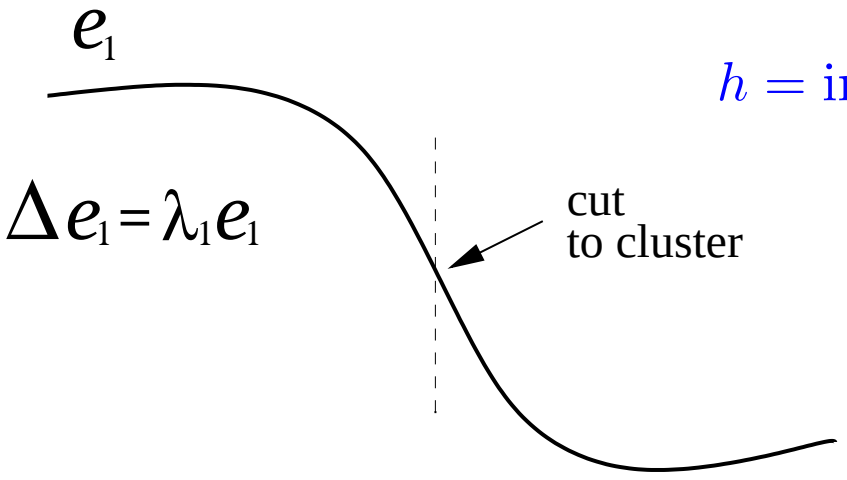
[Theorem]
There exists an algorithm that recovers homology that is polynomial in D.

Niyogi, Smale, Weinberger; to appear

Spectral Clustering

Isoperimetric inequalities. Cheeger constant.





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

$$h \leq \frac{\sqrt{\lambda_1}}{2}$$ [Cheeger]

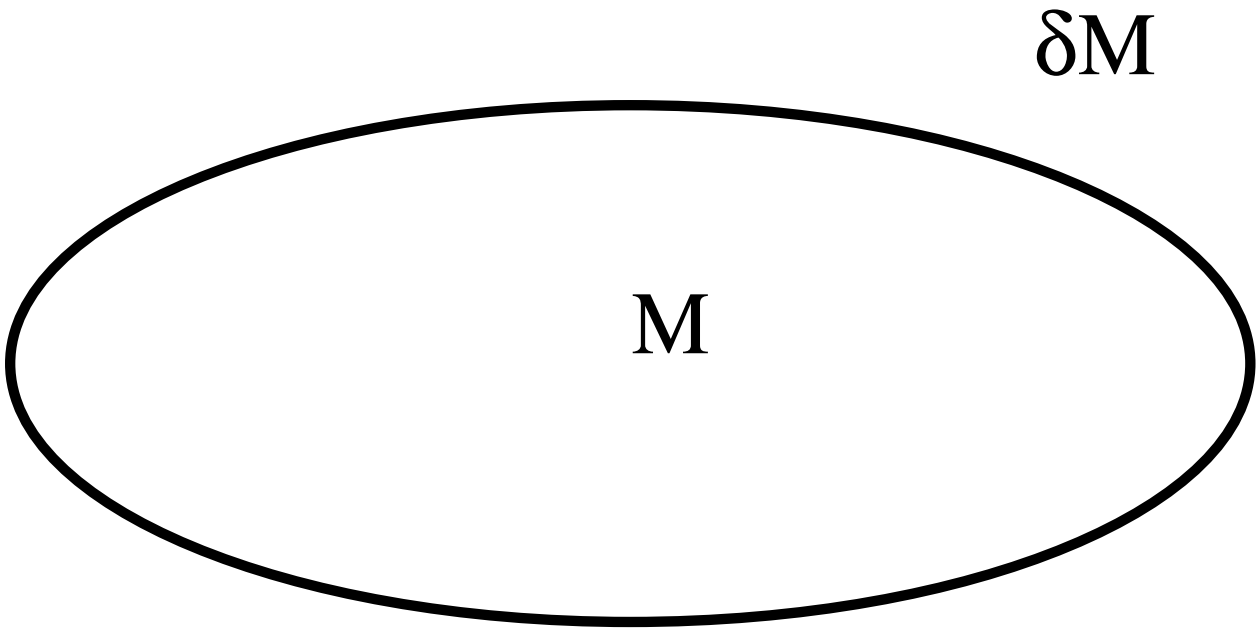
Volume Computation

- Volume of a convex body is deterministically hard to compute. (Bárány and Füredi, 1987).
- Polynomial by randomized algorithm. (Dyer, Kannan, Frieze, 1991).
- Current best $O^*(d^4)$ by Lovász and Vempala.

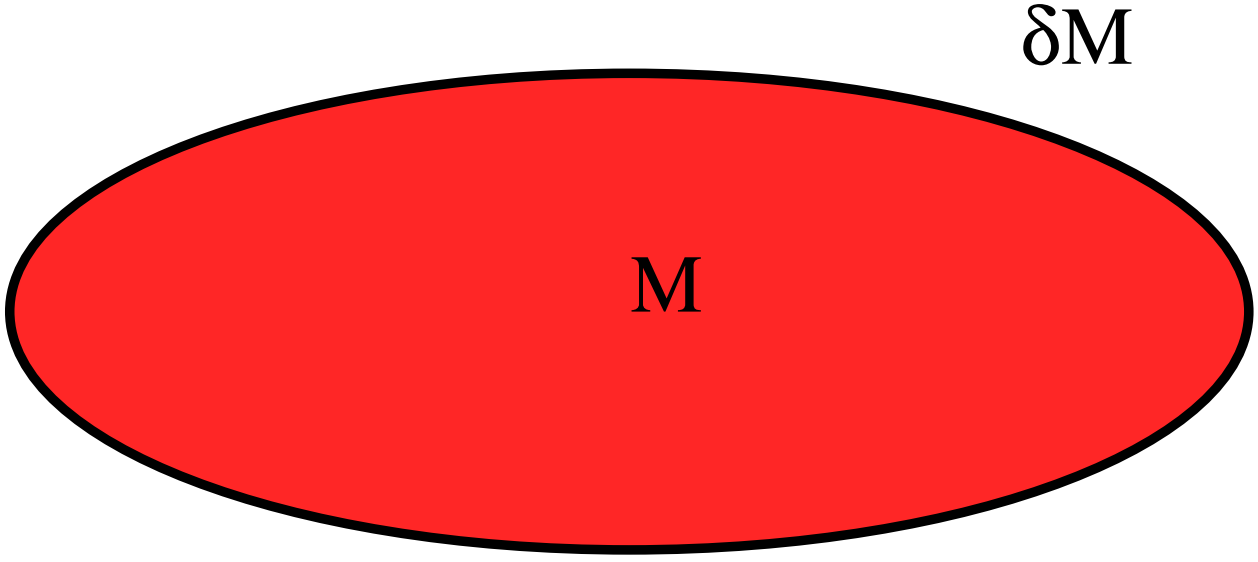
Volume Computation

- Volume of a convex body is deterministically hard to compute. (Bárány and Füredi, 1987).
- Polynomial by randomized algorithm. (Dyer, Kannan, Frieze, 1991).
- Current best $O^*(d^4)$ by Lovász and Vempala.
- Surface volume at least as hard as volume. Grötschel, Lovász and Schrijver (1987) list as open problem.
- Dyer, Gritzmann and Hufnagel (1998) compute surface volume in randomized polynomial time. Complexity seems high.

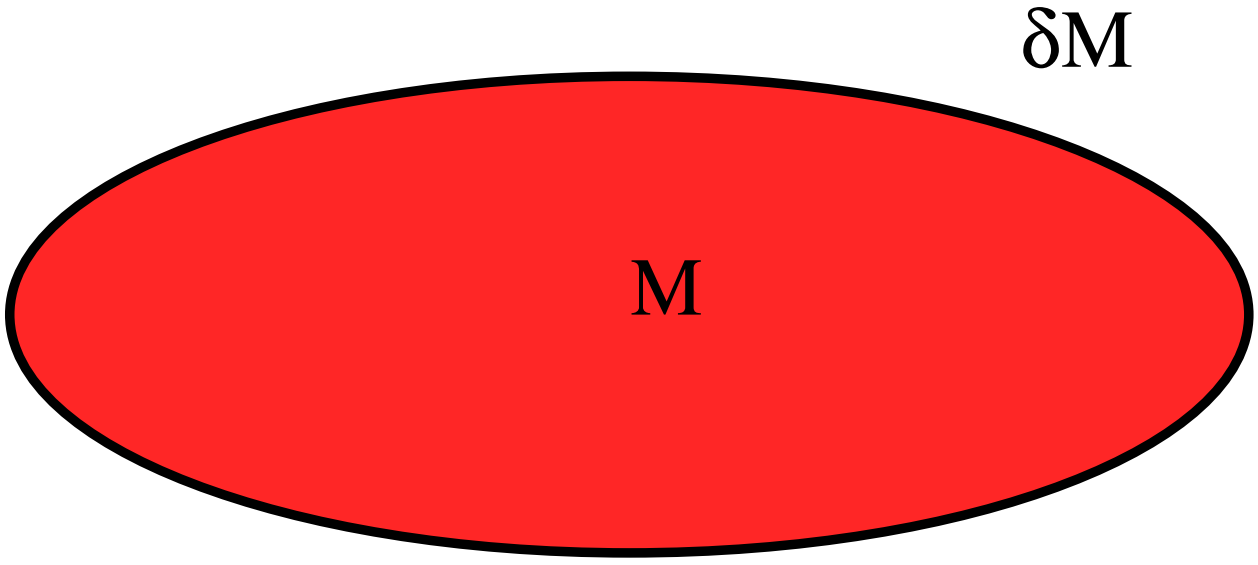
Surface Volume



Surface Volume



Surface Volume



Heat Equation $\Delta u = u_t$ with $u(x,0) = f(x)$

Solution $u(x,t) = f(x) * K_t(x,y)$

$$F_t(M) = \sqrt{\frac{\pi}{t}} \int_{\mathbb{R}^d \setminus M} u(x,t)dx$$

Theorems

[Theorem 1]
$\lim_{t \to 0} F_t(M) = |\partial M|$

[Theorem 2]
Let M be a convex body in $\mathbb{R}^d$ such that
(i) $B_r \subset M \subset B_R$
(ii) ∂M has condition $1/\tau$
Then, it is possible to find the surface area of M in time

$$O^* \left(\frac{d^4}{\epsilon^2} + \frac{d^{3.5} R^3}{r^2 \tau \epsilon^3} \right),$$

with error of ϵ with prob. $> 3/4$.

Belkin, Narayanan, Niyogi (2005)

Future Directions

- Algorithmic Nash embedding
- Random Hodge Theory
- Machine Learning
- Partial Differential Equations
- Graphics
- Algorithms