The Role of Laplace-Beltrami Operator in Manifold Learning

Mikhail Belkin **Partha Niyogi**

The University of Chicago
Departments of Mathematics and Computer Science,
Hyde Park, Chicago, IL 60637

Manifold Model for the Data

Data space $\mathcal{M}^n$ is a Riemannian manifold, i.e. a manifold with an intrinsic notion of distance.

Observation map is an isometric embedding

$$\pi : \mathcal{M}^n \hookrightarrow \mathbb{R}^N$$

$\pi(\mathcal{M}^n)$ is a (low-dimensional) submanifold of $\mathbb{R}^N$.

Observed data $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \pi(\mathcal{M}) \subset \mathbb{R}^N$

What can we do within this framework?

Problems of Machine Learning

1. Regression/Classification
2. Data Representation
3. Clustering

$\mathbb{R}^N \to \mathbb{R}$	$\times$
$\mathcal{M} \to \mathbb{R}$	$\surd$

Laplacian provides a unifying framework.

Laplace-Beltrami Operator

The Laplace-Beltrami operator:

$$\mathcal{L} f \stackrel{def}{=} \operatorname{div} \nabla(f)$$

Gradient ∇ and $-\operatorname{div}$ are formally adjoint operators, i.e. if $\mathbf{X}$ is a vector field on manifold $\mathcal{M}$ then

$$\int_{\mathcal{M}} \langle \mathbf{X}, \nabla f \rangle \, d\nu = \int_{\mathcal{M}} \operatorname{div}(\mathbf{X}) \, f d\nu$$

If $\mathbf{X} = \nabla f$

$$\int_{\mathcal{M}} \|\nabla f\|^2 d\nu = \int_{\mathcal{M}} \mathcal{L} f \cdot f d\nu$$

$\mathcal{L}$ is a self-adjoint positive semidefinite operator. If $\mathcal{M}$ is compact, the eigenvalues of $\mathcal{L}$ are discrete.

$$0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_n \leq \ldots$$

Corresponding eigenfunctions $e_0, e_2, \ldots$ provide an **orthonormal basis** for $\mathcal{L}^2(\mathcal{M})$.

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

$$f(\mathbf{x}) = \sum_{i=0}^{\infty} a_i \, e_i(\mathbf{x})$$

$$a_i = \int_{\mathcal{M}} f(\mathbf{x}) \, e_i(\mathbf{x}) \; d\mu$$

Generalization of Fourier series.

An Example: Circle

$$\mathcal{L}f = -\frac{d^2 f}{d\phi^2}$$

Eigenfunctions: $e_n = e^{in\phi}$.
Eigenvalues: $\lambda_n = n^2$.

Fourier series:

$$f(\phi) = \sum_{n=0}^{\infty} a_n e^{in\phi}$$

$$a_n = \int_{S^1} f(\phi)\, e^{in\phi}\, d\phi$$

Laplacian as a Smoothness Functional

On a circle

$$\mathcal{S}(f) = \int_{S^1} |f(\phi)'|^2 d\nu =$$

$$-\int_{S^1} f(\phi)\, f(\phi)'' d\nu = \langle f, \mathcal{L}(f) \rangle_{S^1}$$

In general

$$\mathcal{S}(f) = \int_{\mathcal{M}} \|\nabla f(\mathbf{x})\|^2 d\nu =$$

$$\int_{\mathcal{M}} \mathcal{L}(f)\, f\, d\nu = \langle f, \mathcal{L}(f) \rangle_{\mathcal{M}}$$

Eigenfunctions of the Laplacian are ordered by smoothness:

$$\mathcal{S}(e_i) = \lambda_i e_i$$

Regression/Classification

Problem: estimate $f : \mathcal{M} \to \mathbb{R}$ given

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)$$

Regularization:

$$\tilde{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \left[\frac{1}{k} \sum_i |f(\mathbf{x}_i) - y_i|^2 + \lambda \mathcal{S}(f) \right]$$

$$\mathcal{S}(f) = \langle f, \mathcal{L} f \rangle_M$$

Many other smoothness functionals are available within the framework.

Data Representation

Problem: given $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathcal{M}$
find $y_1, \ldots, y_k \in \mathbb{R}^l$ that "represent" $\mathbf{x}_i$.
Additional condition: we want a map

$$\tau : \mathcal{M} \to \mathbb{R}^l$$

If we want to preserve neighborhoods, then a natural choice is to minimize "average distortion"

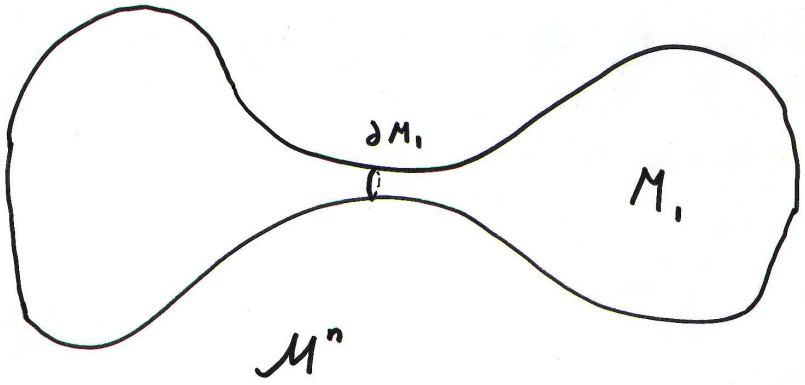
$$\int_{\mathcal{M}} \|\nabla \tau\|^2 d\mu$$

Eigenfunctions of $\mathcal{L}$ give a solution.

$$\tau(\mathbf{x}) = (e_0(\mathbf{x}), \ldots, e_l(\mathbf{x}))$$

Clustering

Isoperimetric inequalities. Cheeger's constant.



$$h = \inf \frac{\mathrm{vol}^{n-1}(\partial M_1)}{\min\left(\mathrm{vol}^n(M_1), \mathrm{vol}^n(\mathcal{M}^n - M_1)\right)}$$

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

Spectral Graph Theory

Laplace-Beltrami operator has an analogue in graph theory.

G is a graph with the weight matrix W. $L^2(G)$ is a space of functions $G \to \mathbb{R}$. The graph Laplacian

$$L = D - W$$

$$D = \begin{pmatrix} \Sigma_i\, w_{1i} & 0 & \dots & 0 \\ 0 & \Sigma_i\, w_{2i} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma_i\, w_{ki} \end{pmatrix}$$

L is a positive semidefinite operator $L^2(G)$. Eigenfunctions of the Laplacians $\mathbf{e_1}, \dots, \mathbf{e_n}$ form an orthonormal basis for $L^2(G)$.

Graph Laplacian as a Smoothness Functional on a Graph

$\mathbf{f} : G \to \mathbb{R}$ is smooth if it does not change much between the adjacent points.

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

Key fact:

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

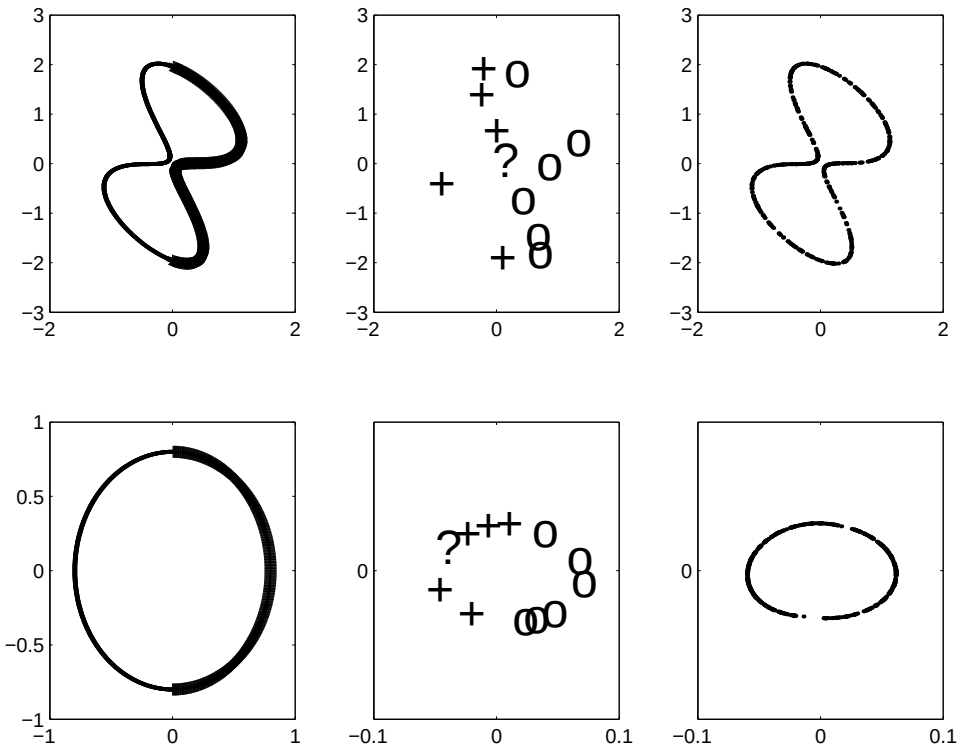
Analagous to:

$$\int_{\mathcal{M}} \|\nabla f\|^2 \, d\nu = \int_{\mathcal{M}} \mathcal{L}(f) f \, d\nu$$

Regularization:

$$\tilde{\mathbf{f}} = \underset{\mathbf{f} \in L^2(G)}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{\substack{\text{labeled} \\ \text{points}}} (f_i - y_i)^2 + \lambda S(\mathbf{f}) \right)$$

Why Manifolds Are Useful



Need only **unlabeled** points to estimate the manifold. Natural application: partially labeled classification.

Convergence of Graph Laplacian to Laplace-Beltrami operator

Gaussian kernel:

$$W_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{4t}}, & \text{if } 0 < \|x_i - x_j\| < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

If points are sampled uniformly from $\mathcal{M}$, can use heat kernel on $\mathcal{M}$ to show convergence. Heat kernel on $\mathcal{M}$ is closely related to heat kernel on $\mathbb{R}^N$.

$$\mathcal{L}(f)(\mathbf{x}_0) \approx K \sum_i e^{-\frac{\|\mathbf{x}_0 - \mathbf{x}_i\|^2}{4t}} f(\mathbf{x}_i)$$

Experimental results: optical character recognition

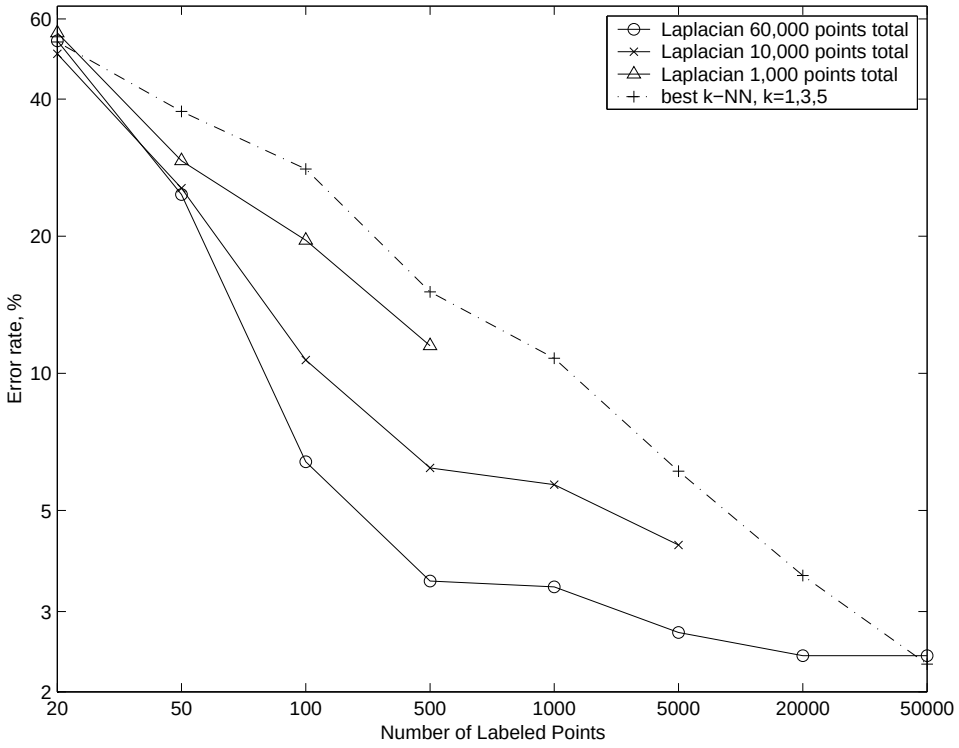


Figure 1: MNIST data set. Percentage error rates for different numbers of labeled and unlabeled points compared to best k-NN base line.

Experimental results: optical character recognition

Labeled	Number of Eigenvectors								best
points	5	10	20	50	100	200	500	1000	k-NN
20	53.7	35.8							53.4
50	48.3	24.7	12.9						37.6
100	48.6	22.0	6.4	14.4					28.1
500	49.1	22.7	5.6	3.6	3.5	7.0			15.1
1000	51.0	24.1	5.5	3.4	3.2	3.4	8.1		10.8
5000	47.5	25	5.6	3.4	3.1	2.9	2.7	2.7	6.0
20000	47.7	24.8	5.4	3.3	3.1	2.9	2.7	2.4	3.6
50000	47.3	24.7	5.5	3.4	3.1	3.0	2.7	2.4	2.3

Table 1: Percentage error rates for different numbers of labeled points for the 60000 point MNIST dataset. The error rate is calculated on the unlabeled part of the dataset, each number is an average over 20 random splits.The rightmost two columns contain the nearest neighbor base line.

Experimental results: text classification

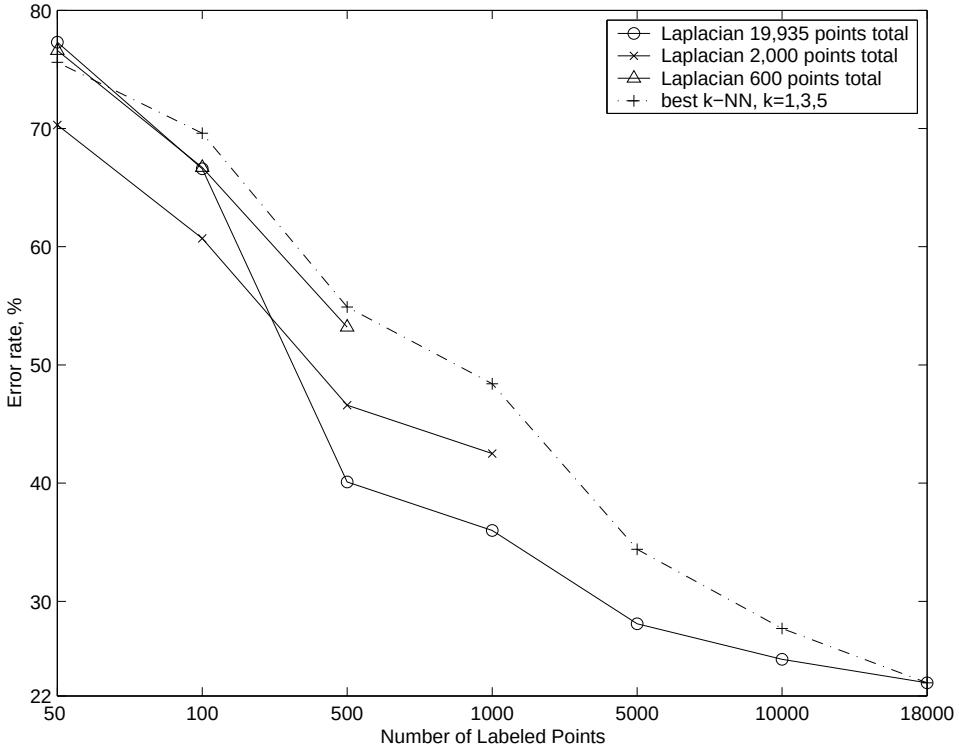


Figure 2: 20 Newsgroups data set. Error rates for different numbers of labeled and unlabeled points compared to best k-NN baseline.

Experimental results: text classification

Lab.	Number of Eigenvectors									best
pts	5	10	20	50	100	200	500	1000	2000	k-NN
50	83.4	77.3	72.1							75.6
100	81.7	74.3	66.6	60.2						69.6
500	83.1	75.8	65.5	46.4	40.1	42.4				54.9
1000	84.6	77.6	67.1	47.0	37.7	36.0	42.3			48.4
5000	85.2	79.7	72.9	49.3	36.7	32.3	28.5	28.1	30.4	34.4
10000	83.8	79.8	73.8	49.8	36.9	31.9	27.9	25.9	25.1	27.7
18000	82.9	79.8	73.8	50.1	36.9	31.9	27.5	25.5	23.1	23.1

Table 2: Percentage error rates for various numbers of labeled points and eigenvectors. The total number of points is 19935. The error is calculated on the unlabeled part of the dataset.