Laplacians, Spectra, Kernels And All That

John Lafferty

School of Computer Science
Carnegie Mellon University

Based on joint work/discussions with

Risi Imre Kondor, Guy Lebanon, Avrim Blum, Mugizi Rwebangira, Jerry Zhu

An Elliptical Sun

The Laplace operator in its various manifestations is the most beautiful and central object in all of mathematics. Probability theory, mathematical physics, Fourier analysis, partial differential equations, the theory of Lie groups, and differential geometry all revolve around this sun, and its light even penetrates such obscure regions as number theory and algebraic geometry.

—Edward Nelson, *Tensor Analysis*

There is nothing new under the sun.

—Ecclesiastes, 1:9

Outline

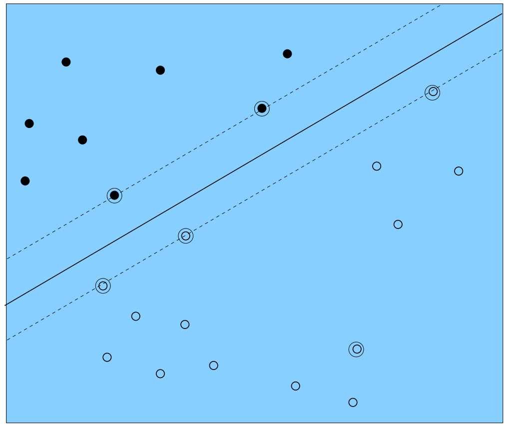
- Kernels
- Laplacians – continuous and discrete (differential geometry)
- Diffusion kernels on graphs
- Diffusion kernels on statistical models
- Learning from labeled and unlabeled data

Kernels

(Aizerman, Braverman & Rozonoér, 1964) (Boser, Guyon & Vapnik, 1992)

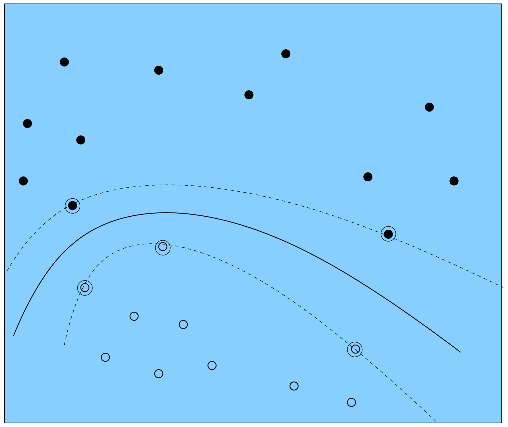
- A powerful & elegant tool—dramatic improvements in accuracy (and sometimes computational efficiency)
- Turn linear classifiers into non-linear classifiers
- Implicit representation of high dimensional feature
- Popular indoor sport—"Kernel X", where X is PCA, ICA, CCA Clustering, ... 20+ papers at upcoming NIPS
- *Primarily a tool for Euclidean space*

Linear Classifiers



$$\hat{f}(x) \;=\; \sum_{i=1}^{N} \alpha_i \, y_i \, \langle x, x_i \rangle$$

Using a Kernel



$$\hat{f}(x) \;=\; \sum_{i=1}^{N} \alpha_i \, y_i \, K(x, x_i)$$

Popular Kernels

Gaussian: $\quad K_\sigma(x, x') \;=\; \exp\left(- \|x - x'\|^2 / 2\sigma^2\right)$

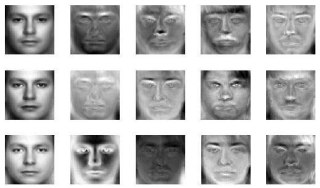
Polynomial: $\quad K_d(x, x') \;=\; (1 + \langle x, x' \rangle)^d$

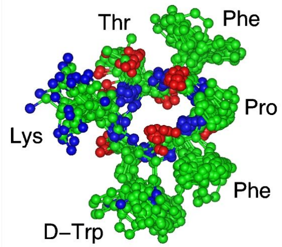
Sigmoid: $\quad K_\alpha(x, x') \;=\; \tanh(\langle x, x' \rangle + \alpha)$

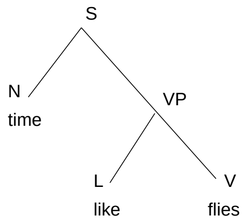
However, these only apply in Euclidean spaces. Usually means data has already been "compromised"

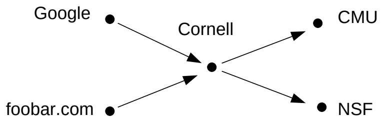
Structured Data

What if data lies on a graph or other data structure?







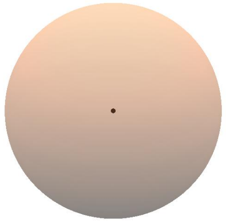


The Laplacian

Recall the usual Laplacian from calculus:

$$\Delta f \;=\; \sum_i \frac{\partial^2 f}{\partial x_i^2} \;=\; I^{ij}\,\partial_i\,\partial_j f \;=\; \mathrm{div}\circ\nabla f$$

Laplacian: Invariant Definition



$$\Delta f(p) \;=\; -\lim_{r\to 0}\frac{2n}{r^2}\left(f(p)-\frac{1}{\mathrm{vol}(S_r)}\int_{S_r(p)} f\right)$$

f is *harmonic* if $\Delta f = 0$

The Laplacian on a Riemannian Manifold

On a Riemannian manifold with metric g, divergence is

$$\mathrm{div}\,X \;=\; \frac{1}{\sqrt{\det g}}\,\partial_i\left(X^i\sqrt{\det g}\right)$$

The geometric Laplacian is then $\mathrm{div}\circ\nabla$ or

$$\Delta f \;=\; \frac{1}{\sqrt{\det g}}\,\partial_i\left(g^{ij}\sqrt{\det g}\,\partial_i f\right)$$
$$\;=\; g^{ij}\partial_j\partial_i f + \text{(lower order terms)}$$

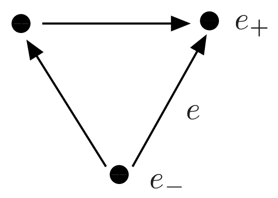
More generally, $\Delta = d^*d + d\,d^*$.

Laplacian Represents the Geometry

The Laplacian is determined by the metric, but also *determines* the metric, by evaluating on functions locally given by $x^i x^j$.

Spectral theory of Δ is intimately connected with geometry of the manifold

Combinatorial Laplacian



Think of edge e as "tangent vector" at e_-.

For $f : V \longrightarrow \mathbb{R}$, $df : E \longrightarrow \mathbb{R}$ is the 1-form

$$df(e) = f(e_+) - f(e_-)$$

Combinatorial Laplacian

Representing "exterior derivative" d as matrix

$$D_{e,v} = \begin{cases} 1 & \text{if } v = e_+ \\ -1 & \text{if } v = e_- \\ 0 & \text{otherwise} \end{cases}$$

Then $\Delta = D^* D$ is discrete analogue of $\operatorname{div} \circ \nabla = d^* d$

Combinatorial Laplacian

Since $\langle f, \Delta g \rangle = \langle df, dg \rangle$, Δ is self-adjoint and positive.

It is an *averaging operator*

$$\begin{aligned} \Delta f(x) &= \sum_{y \sim x} w_{xy}(f(x) - f(y)) \\ &= d(x)\, f(x) - \sum_{x \sim y} w_{xy} f(y) \end{aligned}$$

We say f is *harmonic* if $\Delta f = 0$.

Boundary Conditions

If the manifold has a boundary, need appropriate boundary conditions so that Δ is self-adjoint.

Dirichlet boundary conditions: $f|_{\partial M} = 0$

Neumann boundary conditions: $\frac{\partial f}{\partial \nu}|_{\partial M} = 0$, for the outer normal direction ν.

The Heat Kernel

Properties of the heat kernel on a Riemannian manifold:

1. $K_t(x,y) = K_t(y,x)$
2. $\lim_{t\to 0} K_t(x,y) = \delta_x(y)$
3. $\left(\Delta - \frac{\partial}{\partial t}\right) K = 0$
4. $K_t(x,y) = \int_M K_{t-s}(x,z) K_s(z,y)\, dz$
5. $K_t(x,y) = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(x)\, \phi_i(y)$
6. $e^{t\Delta} f(x) = \int_M K_t(x,y)\, f(y)\, dy$

Parametrices

Have an asymptotic expansion for the heat kernel:

$$K_t(x,y) = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{d^2(x,y)}{4t}\right) \sum_{i=0}^{N} \psi_i(x,y) t^i + O(t^N)$$

in terms of the *parametrices* ψ_i.

Much of modern differential geometry is based on Laplacians, heat kernels and related constructions.

Outline

- Kernels
- Laplacians – continuous and discrete
- Diffusion kernels on graphs
- Diffusion kernels on statistical models
- Learning from labeled and unlabeled data

Diffusion Kernels on Graphs

(Kondor and Lafferty, ICML 2002)

Kernels are "*generalized covariances*." Crucial positive semidefinite property required by RKHS theory is in general difficult to obtain.

Let Δ be the graph Laplacian. In analogy with the continuous setting,

$$\frac{d}{dt} K_t \;=\; \Delta K_t$$

is the *heat equation* on a graph. Solution

$$K_t \;=\; e^{t\Delta}$$

is called the *heat kernel* or *diffusion kernel*.

Physical Interpretation

- $\left(\Delta - \frac{\partial}{\partial t}\right) K = 0$, initial condition $\delta_x(y)$:
- $e^{t\Delta} f(x) = \int_M K_t(x,y)\, f(y)\, dy$

For a kernel-based classifier

$$\hat{y}(\boldsymbol{x}) = \sum_i \alpha_i\, y_i\, K_t(\boldsymbol{x}_i, \boldsymbol{x})$$

decision function is given by heat flow with initial condition

$$f(\boldsymbol{x}) = \begin{cases} \alpha_i & \boldsymbol{x} = \boldsymbol{x}_i \in \text{positive labeled data} \\ -\alpha_i & \boldsymbol{x} = \boldsymbol{x}_i \in \text{negative labeled data} \\ 0 & \text{otherwise} \end{cases}$$

RKHS Representation

General spectral representation of a kernel as $K(x,y) = \sum_{i=1}^n \lambda_i \phi_i(x)\phi_i(y)$ leads to *reproducing kernel Hilbert space*

$$\left\langle \sum_i a_i \phi_i, \sum_i b_i \phi_i \right\rangle_{\mathcal{H}_K} = \sum_i \frac{a_i\, b_i}{\lambda_i}$$

For the diffusion kernel, RKHS inner product is

$$\langle f, g \rangle_{\mathcal{H}_K} = \sum_i e^{t\lambda_i} \widehat{f}_i\, \widehat{g}_i$$

Interpretation: Functions with small norm don't "oscillate" rapidly on the graph.

Building Up Kernels

If $K_t^{(i)}$ are kernels on $\mathcal{X}_i$
$K_t = \otimes_{i=1}^n K_t^{(i)}$ is a kernel on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_n$.

For the hypercube:

$$K_t(x, x') \propto (\tanh t)^{\overbrace{d(x,x')}^{\text{Hamming distance}}}$$

Similar kernels apply to standard categorical data. Other graphs with explicit diffusion kernels:

- Infinite trees (Chung & Yau, 1999)
- Rooted trees
- Cycles
- Strings with wildcards

Recent Application of Diffusion Kernels

(Vert and Kanehisa, upcoming NIPS)

- Task: Gene function prediction from microarray data. Question: Can knowledge of metabolic pathways and catalyzing enzymes improve the performance?
- Idea: Documented network of biochemical reactions forms a graph where genes are vertices. Relevant features are likely to exhibit correlation with respect to the topology of the graph. Combine diffusion kernel with expression profiles using kernel CCA (Bach and Jordan, 2002).
- Experiments on genes of yeast *S. Cerevisiae* yield consistent improvement in predicting genes' functional class.

Recent Application of Diffusion Kernels

(Kandola, Shawe-Taylor, & Cristianini, upcoming NIPS**)**

Taking $G = XX^\top$, where X is the document/term matrix, they consider the kernel

$$K_t = \exp(tG)$$

and show dramatic improvements in text classification performance over bag-of-words methods.

Outline

- Kernels
- Laplacians – continuous and discrete
- Diffusion kernels on graphs
- Diffusion kernels on statistical models
- Learning from labeled and unlabeled data

Diffusion on the Information Manifold

(Lafferty and Lebanon, NIPS'02)

- "Similarity" between x and x' may be difficult to quantify–even for domain experts.
- Generative models are fairly easy to derive. A good generative model will use domain knowledge.
- Fisher information on the family "codifies" this into a distance. Gives statistical family structure of a Riemannian manifold.
- Associate each data point with a model: $x \mapsto \theta(x)$, and consider diffusion kernel on information manifold.

Information Geometry

$\mathfrak{F} = \{p(\cdot \mid \theta),\ \theta \in \Theta \subset \mathbb{R}^d\}$ d-dimensional statistical family.

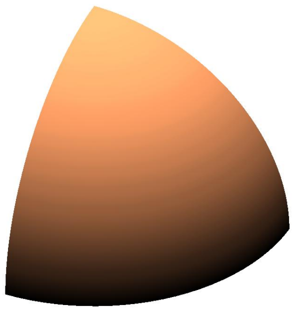
Fisher information matrix $[g_{ij}(\theta)]$

$$\begin{aligned} g_{ij}(\theta) &= E_\theta[\partial_i \ell_\theta \, \partial_j \ell_\theta] \\ &= \int_{\mathcal{X}} \partial_i \log p(x \mid \theta) \, \partial_j \log p(x \mid \theta) \, p(x \mid \theta) \, dx \end{aligned}$$

Defines a Riemannian metric on Θ, which is invariant under reparameterization.

See (Amari, '00; Kass & Voss, '97)

Special Case: Multinomial



$\sqrt{\cdot}$ maps the simplex to the sphere

Special Case: Multinomial

$$\theta \in d\text{-simplex} \;\Rightarrow\; \left(\sqrt{\theta_1}, \ldots, \sqrt{\theta_{d+1}}\right) \in d\text{-sphere}$$

Multinomial geometry:

$$d(\theta, \theta') = 2\arccos\left(\sum_{i=1}^{d+1} \sqrt{\theta_i \theta'_i}\right)$$

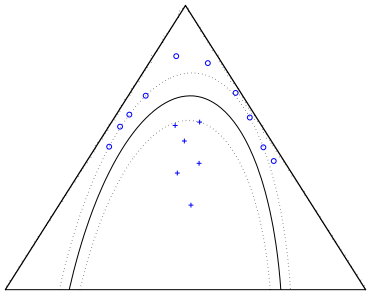
Use maximum likelihood mapping $x \mapsto \widehat{\theta}(x)$.

Special Case: Multinomial

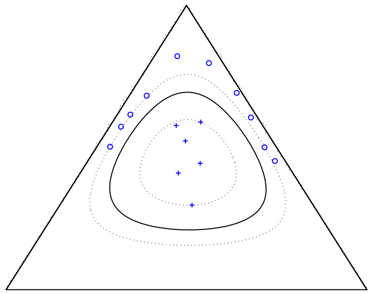
Parametrix expansion leads to approximation

$$K_t(x, x') \approx \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{1}{t}\arccos^2\left(\sum_{i=1}^{d+1} \sqrt{\theta_i(x)\theta_i(x')}\right)\right)$$

SVM Decision Boundaries: Trinomial



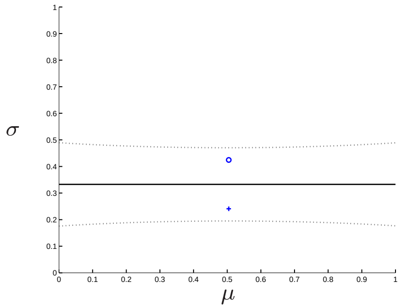
Gaussian kernel



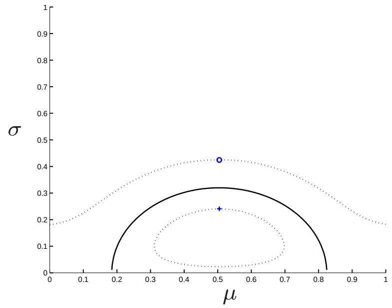
Information diffusion kernel

Special Case: Spherical Normal

$p(x \mid \theta) = \mathcal{N}(\boldsymbol{\mu}, \sigma I)$. Information metric: hyperbolic space.



Gaussian kernel



Information diffusion kernel

Decision boundary for the diffusion kernel makes intuitive sense, since as $\sigma \searrow 0$, mean is known with increasing certainty.

Bounds on the Spectrum

From (Li and Yau, 1980), for compact M, dimension d, positive Ricci curvature, convex boundary, and volume V, Dirichlet eigenvalues of the Laplacian satisfy

$$c_1(d)\left(\frac{j}{V}\right)^{\frac{2}{d}} \leq \mu_j \leq c_2(d)\left(\frac{j+1}{V}\right)^{\frac{2}{d}}$$

Manifold of the multinomial satisfies these conditions.

These bounds can be used together with results of (Guo et al.,'02) that "effective dimension" of the kernel is

$$\left(\frac{\lambda_1 \cdots \lambda_{j_n^*}}{n^2}\right)^{\frac{1}{j_n^*}}$$

Bounds on Covering Numbers and Generalization Error

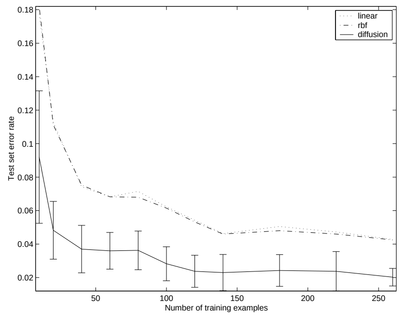
Theorem. *Suppose the information manifold M is compact, with positive Ricci curvature, dimension d and volume V. Then the SVM hypothesis class covering numbers $\mathcal{N}(\epsilon, \mathcal{F}_R(\boldsymbol{x}))$ for the diffusion kernel K_t on M satisfy*

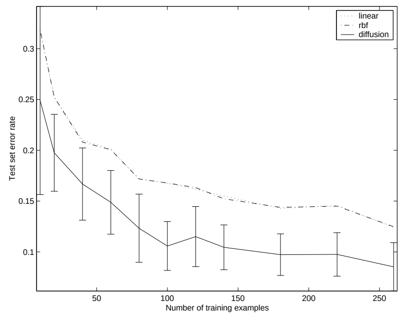
$$\log \mathcal{N}(\epsilon, \mathcal{F}_R(\boldsymbol{x})) = O\left(\left(\frac{V}{t^{\frac{d}{2}}}\right) \log^{\frac{d+2}{2}}\left(\frac{1}{\epsilon}\right)\right)$$

More Recent Bounds

Better bounds on generalization error are now available based on *Rademacher averages* (Bartlett, Bousquet, & Mendelson, preprint).

WebKB Text Classification Experiments





Intuition: improvement comes from $\sqrt{\ }$ in metric, which places more emphasis on values close to the boundary of the simplex.

Experiments in progress: $\mathfrak{F}$ taken as family of smoothed models

Related Work

Haussler and Jaakkola proposed the *Fisher Kernel*

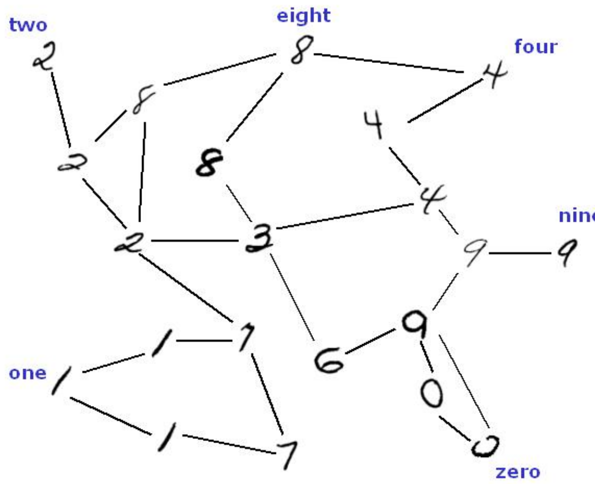
$$K_F(x, x') \;=\; \langle \nabla \log p(x \,|\, \theta), \nabla \log p(x' \,|\, \theta) \rangle$$

- Invariant, but depends on the geometry of the information manifold only at a single point.
- Heuristically exponentiate: $\exp(-\alpha\, K_F(x, x'))$
- A *single* model is used, in practice need more fine-grained distinctions

Outline

- Kernels
- Laplacians – continuous and discrete
- Diffusion kernels on graphs
- Diffusion kernels on statistical models
- Learning from labeled and unlabeled data

Labeled and Unlabeled Data as Graph



View from Statistical Physics

View each vertex x as having label $f(x) \in \{+1, -1\}$.

Ising model on graph/lattice, spins $f : V \longrightarrow \{+1, -1\}$

$$\text{Energy} \quad H(f) = \frac{1}{2}\sum_{x \sim y} w_{xy}\,(f(x) - f(y))^2$$

$$\equiv -\sum_{x \sim y} w_{xy} f(x)\, f(y)$$

$$\text{Gibbs distribution} \quad P(f) = \frac{1}{Z(\beta)} e^{-\beta H(f)} \qquad \beta = \frac{1}{T}$$

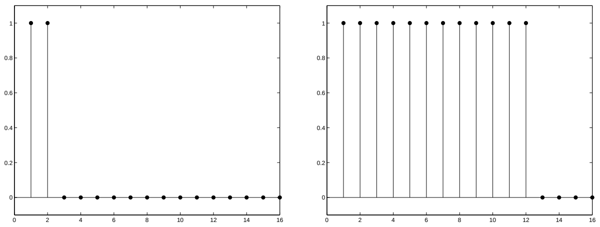
$$\text{Partition function} \quad Z(\beta) = \sum_f e^{-\beta H(f)}$$

Using Graph Cuts

- Fix values on labeled data (boundary values)
- *What is the most probable (lowest energy) configuration?*
- Solution can be computed as a *graph mincut* (Blum and Chawla, ICML'01)
- Non-binary case: 2-approximation of optimal labeling (Kleinberg and Tardos, FOCS'99)

Graph Mincuts

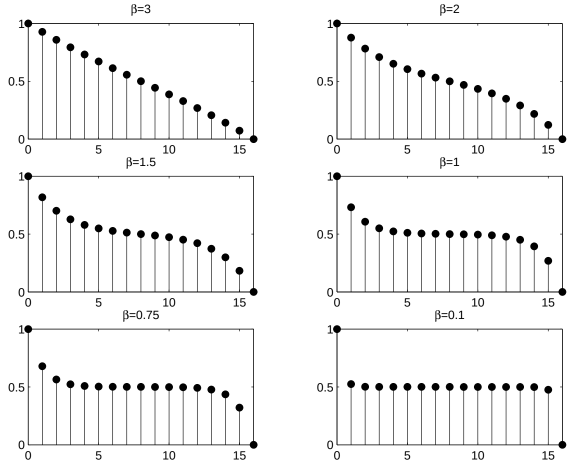
Graph mincuts can be very unbalanced



Idea: Replace by *averages* under Ising model

$$E_\beta[f(x)] = \sum_{f|_{\partial S} = f_B} f(x)\,\frac{e^{-\beta H(f)}}{Z(\beta)}$$

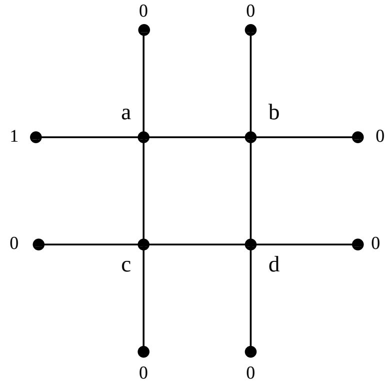
Pinned Ising Model



Not (Yet?) Provably Efficient to Approximate

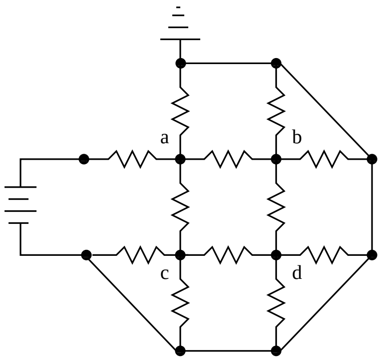
Unfortunately, analogue of rapid mixing result of Jerrum & Sinclair for *ferromagnetic* Ising model not known for mixed boundary conditions

Random Walk Perspective



Starting from $x \in \{a, b, c, d\}$, what is the probability of hitting the vertex labeled 1 before a vertex labeled 0?

Electric Network Perspective



What is the voltage at $x \in \{a, b, c, d\}$?

Solution is Harmonic

Weighted graph $G = (V, E)$, edge weights w_{xy}, combinatorial Laplacian Δ.

Subgraph S with boundary ∂S.

Dirichlet Problem:

$$\begin{aligned} \Delta f &= 0 \quad \text{on } S \\ f|_{\partial S} &= f_B \end{aligned}$$

See (Doyle and Snell,'84) or (Lyons and Peres,'02)

Random Walk Solution

Transmission matrix P written as

$$P = \begin{pmatrix} I & 0 \\ R & Q \end{pmatrix}$$

R : Transition from S to ∂S

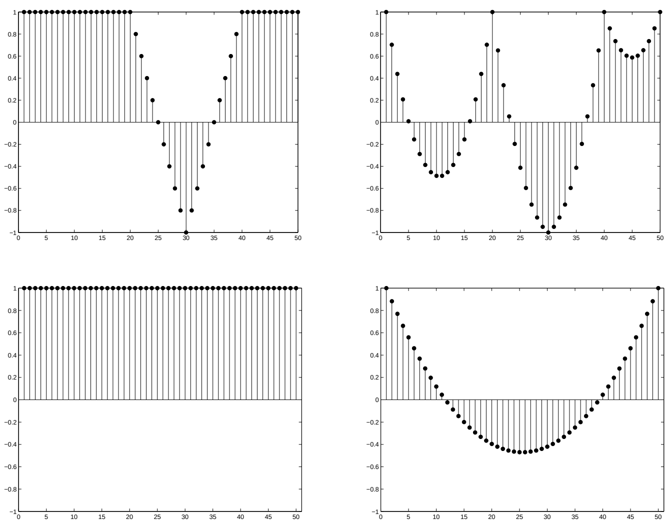
Q : Transition within S

I : Stop when hit ∂S

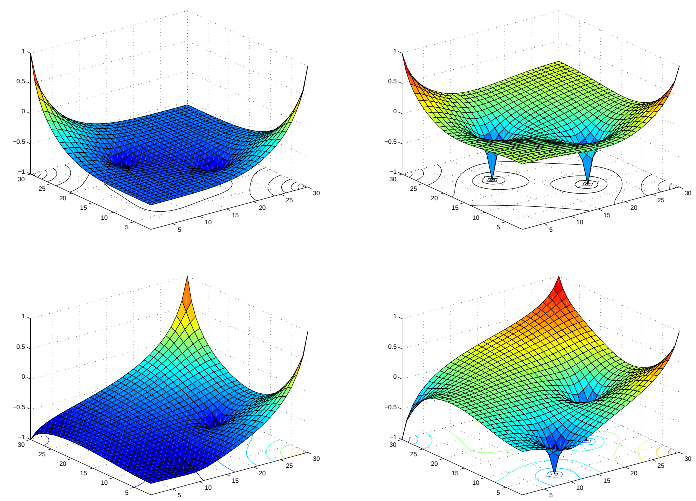
Solution: $f = (I - Q)^{-1} R f_B$

Label Propagation algorithm of Zhu and Gharhamani.

Unconstrained Constrained



Unconstrained Constrained



View from Statistical Physics

Euclidean field theory on graph/lattice, fields $f : V \longrightarrow \mathbb{R}$

$$\text{Energy} \quad H(f) = \frac{1}{2} \sum_{x \sim y} w_{xy} \left(f(x) - f(y)\right)^2$$

$$\text{Gibbs distribution} \quad P(f) = \frac{1}{Z(\beta)} e^{-\beta H(f)} \qquad \beta = \frac{1}{T}$$

$$\text{Partition function} \quad Z(\beta) = \int_f e^{-\beta H(f)}\, df$$

Arises from analytic continuation to imaginary time, $t \mapsto it$
Poincaré group $\mapsto$ Euclidean group.

View from Statistical Physics

What is the most probable field?

$$\begin{aligned} f^* &= \arg\min_{f|_{\partial f} = f_B} H(f) \\ \Delta f^* &= 0 \\ f^*|_{\partial f} &= f_B \end{aligned}$$

Exactly the same as Label Propagation.

Computing the Partition Function

Let λ_i be the Dirichlet eigenvalues of Δ,

$$Z(\beta) = \frac{e^{-\beta H(f^*)}}{(\beta\pi)^{n/2}\sqrt{\det \Delta}} \qquad \det \Delta = \prod_{i=1}^{n} \lambda_i$$

By generalization of matrix-tree (Chung & Langlands,'96)

$$\det \Delta = \frac{\#\text{ rooted spanning forests}}{\prod_i d_i}$$

Connection with Heat Kernel

What is the connection between random walk on unlabeled data and heat kernel?

Let Δ be the combinatorial Laplacian with Dirichlet boundary conditions.

If K_t is the Dirichlet heat kernel, let

$$\overline{K} = \int_0^\infty K_t \, dt$$

Solution to the Dirichlet problem (= Label Propagation):

$$f^*(x) = \sum_{y \in \partial S} \sum_{z \sim y} \overline{K}(x, z)\, p_{zy}\, f_B(y)$$

Summary

- Diffusion kernels are broad generalization of standard Gaussian kernels—apply to discrete data
- Promising applications (protein/text classification)
- Applying to Fisher information combines generative models and discriminative learning methods
- Rich connections between some emerging methods for learning from labeled and unlabeled data
- Many interesting directions to explore, much to discover...