Random Walks, Random Fields, and Graph Kernels

John Lafferty

School of Computer Science
Carnegie Mellon University

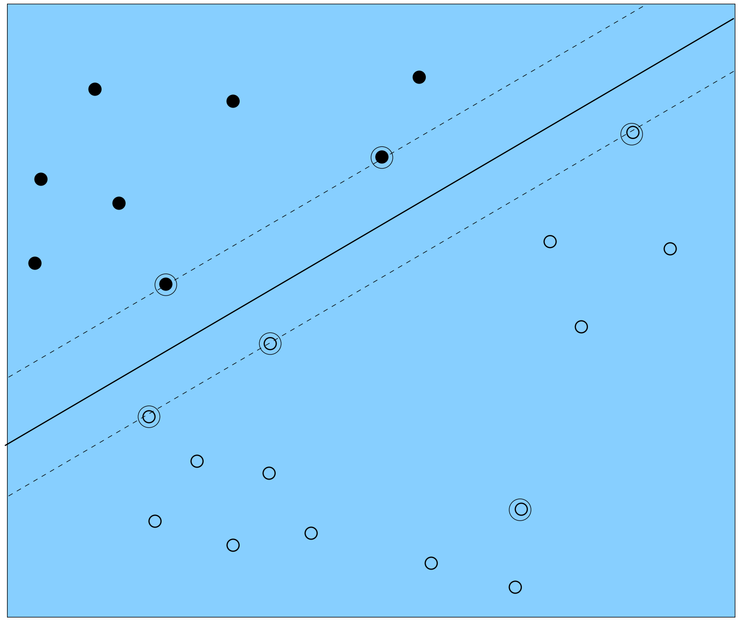
Based on work with

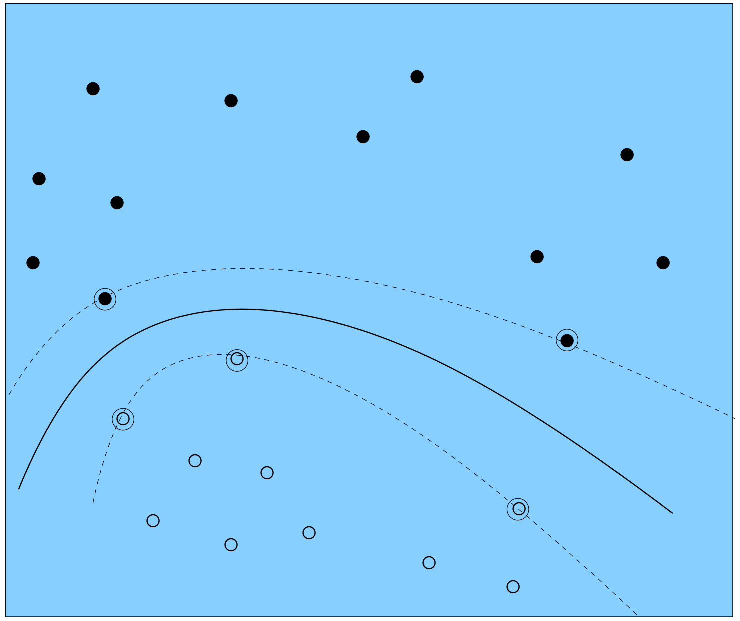
Avrim Blum, Zoubin Ghahramani, Risi Kondor
Mugizi Rwebangira, Jerry Zhu

Outline

$$\begin{tikzcd} \text{Graph Kernels} \arrow[r] & \text{Random Fields} \arrow[d] \\ \text{Random Walks} \arrow[u] & \text{Continuous Fields} \arrow[l] \end{tikzcd}$$

Using a Kernel





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

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

The Kernel Trick

$K(x, x')$ positive semidefinite:

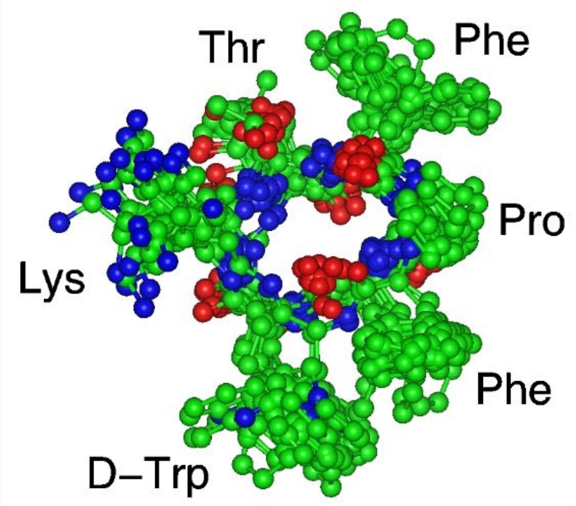
$$\int_{\mathcal{X}} \int_{\mathcal{X}} f(x) f(x') K(x, x')\, dx' dx \geq 0$$

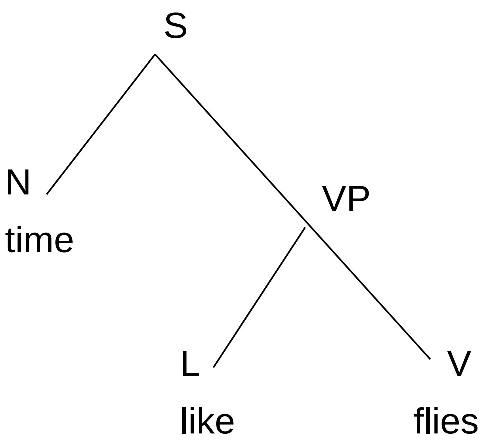
Taking feature space of functions $\mathcal{F} = \{\Phi(x) = K(\cdot, x), x \in \mathcal{X}\}$, has "reproducing property" $g(x) = \langle K(\cdot, x), g \rangle$.

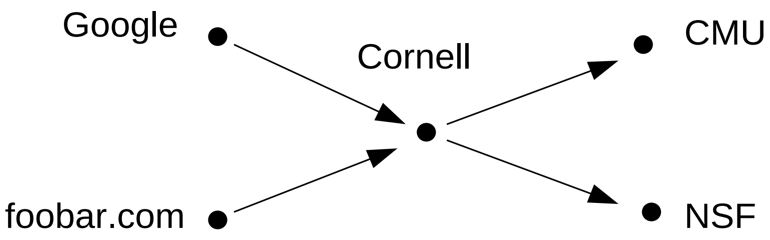
$$\langle \Phi(x), \Phi(x') \rangle = \langle K(\cdot, x), K(\cdot, x') \rangle = K(x, x')$$

Structured Data

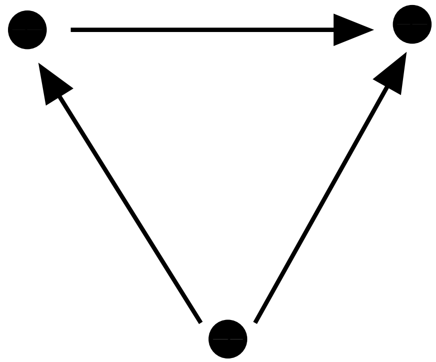
What if data lies on a graph or other data structure?







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

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

Then $\Delta = d^*d$ (as matrix) is discrete analogue of $\operatorname{div} \circ \nabla$

Combinatorial Laplacian

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 $\Delta f = 0$.

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

Diffusion Kernels on Graphs

(Kondor and L., 2002)

If Δ is the graph Laplacian, in analogy with the continuous setting,

$$\frac{\partial}{\partial t} K_t = \Delta K_t$$

is the *heat equation* on a graph. Solution

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

is the *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

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

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:

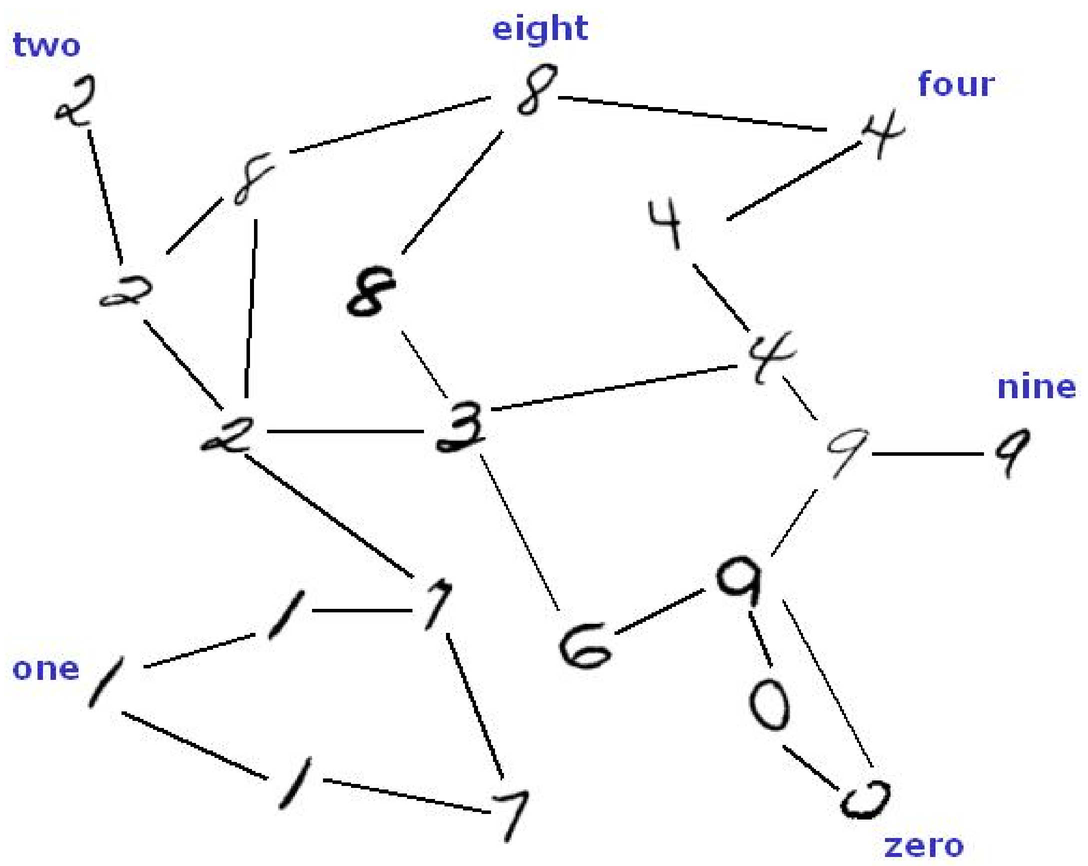
- Infinite trees (Chung & Yau, 1999)
- Rooted trees
- Cycles
- Strings with wildcards

Results on UCI Datasets

	Hamming		*Diffusion Kernel*			*Improv.*	
Data Set	error	$\|SV\|$	error	$\|SV\|$	β	Δerr	$\Delta\|SV\|$
`Breast Cancer`	7.64%	387.0	3.64%	62.9	0.30	62%	83%
`Hepatitis`	17.98%	750.0	17.66%	314.9	1.50	2%	58%
`Income`	19.19%	1149.5	18.50%	1033.4	0.40	4%	8%
`Mushroom`	3.36%	96.3	0.75%	28.2	0.10	77%	70%
`Votes`	4.69%	286.0	3.91%	252.9	2.00	17%	12%

Recent application to protein classification by Vert and Kanehisa (NIPS 2002).

Random Fields View of Combining Labeled/Unlabeled Data



Random Fields View

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

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

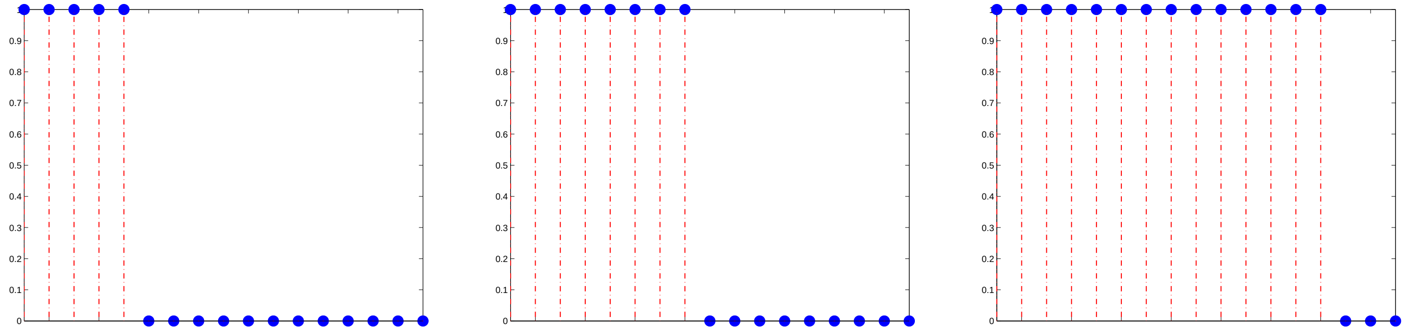
$$\begin{aligned}
\text{Energy} \quad H(f) &= \frac{1}{2} \sum_{x \sim y} w_{xy} \left(f(x) - f(y) \right)^2 \\
&\equiv -\sum_{x \sim y} w_{xy} f(x) f(y)
\end{aligned}$$

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

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

Graph Mincuts

Graph mincuts can be very unbalanced

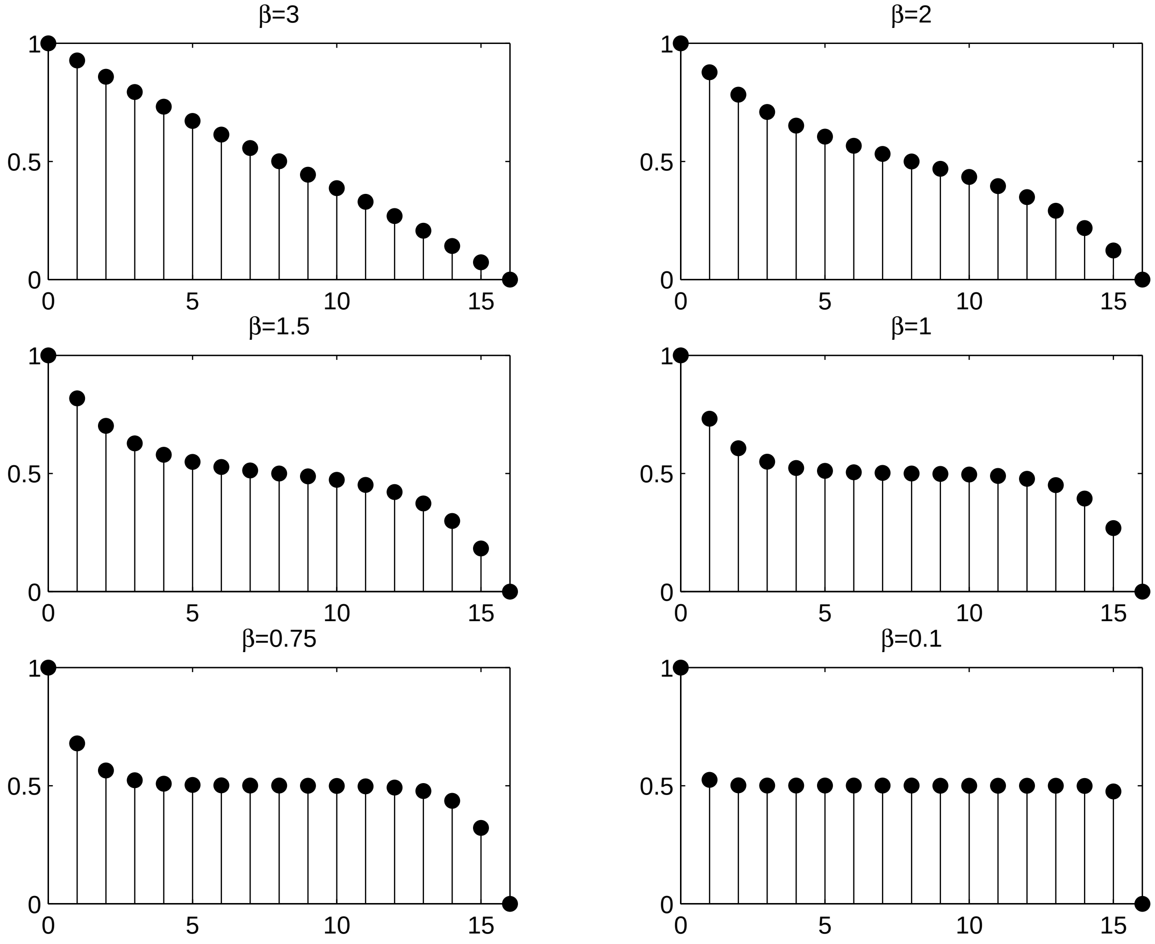


Graph mincuts don't exploit probabilistic properties of random fields

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 (Provably) Efficient to Approximate

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

Question: Can we compute averages using graph algorithms in the zero temperature limit?

Idea: “Relax” to Statistical Field Theory

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

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

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

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

Physical Interpretation: analytic continuation to imaginary time, $t \mapsto it$ Poincaré group $\mapsto$ Euclidean group.

View from Statistical Field Theory (cont.)

Most probable field is harmonic

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

Subgraph S with boundary ∂S.

Dirichlet Problem: unique solution

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

Random Walk Solution

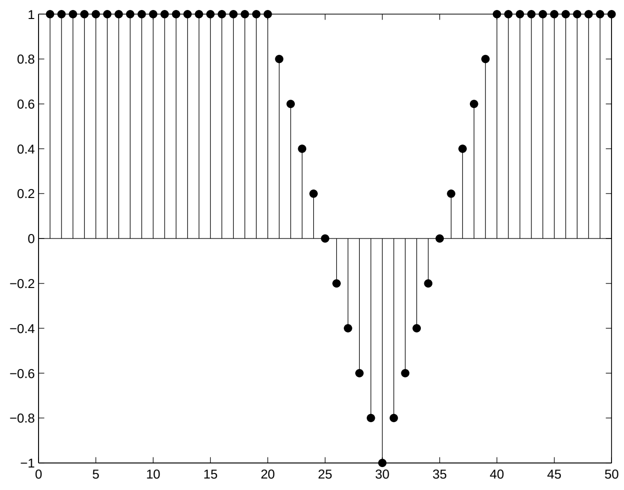
Perform random walk on unlabeled data, stop when hit a labeled point.

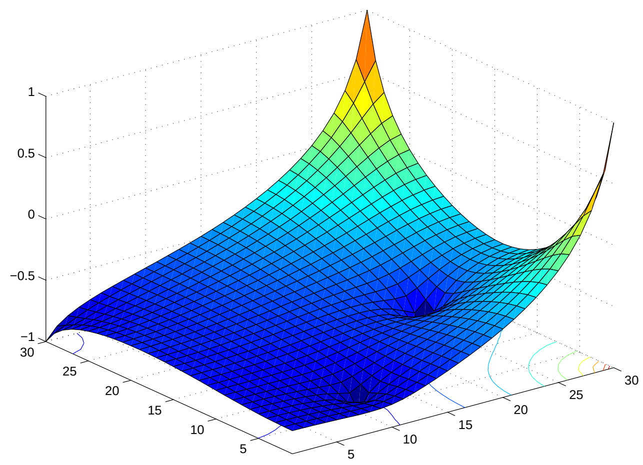
What is the probability of hitting a positive labeled point before a negative labeled point?

Precisely the same as minimum energy (continuous) random field. *Label Propagation*.

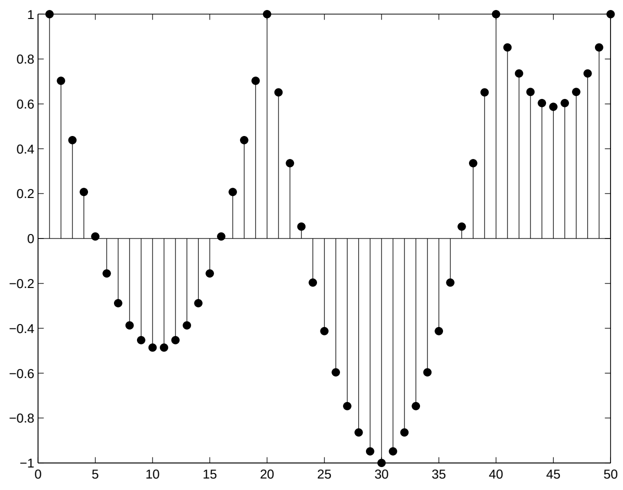
Related work by Szummer and Jaakkola (NIPS 2001)

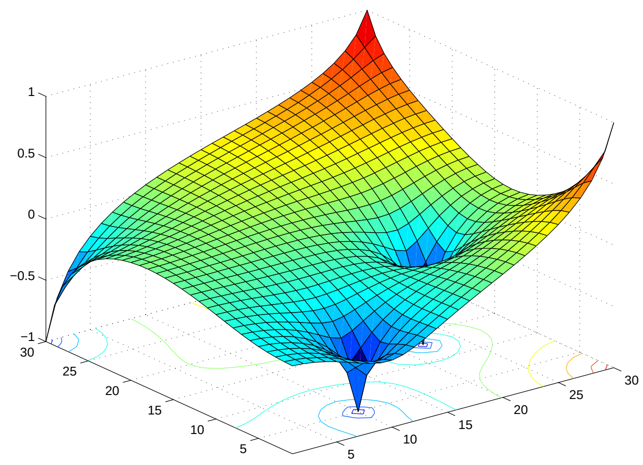
Unconstrained





Constrained





View from Statistical Field Theory

In one-dimensional case: low temperature limit of average Ising model is the same is minimum energy Euclidean field. (Landau)

Intuition: average over graph s-t mincuts; harmonic solution is linear.

Not true in general...

Computing the Partition Function

Let λ_i be spectrum of Δ, Dirichlet boundary conditions:

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

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

$$\det\Delta \;=\; \frac{\#\text{ rooted spanning forests}}{\prod_i \mathsf{deg}(i)}$$

Connection with Diffusion Kernels

Again take Δ, combinatorial Laplacian with Dirichlet boundary conditions (zero on labeled data)

For $K_t = e^{t\Delta}$ diffusion kernel let $\overline{K} = \int_0^\infty K_t \, dt$

Solution to the Dirichlet problem (label prop, minimum energy continuous field):

$$f^*(x) = \sum_{z \in \text{“fringe”}} \overline{K}(x, z)\, f_{\mathcal{D}}(z)$$

Connection with Diffusion Kernels (cont.)

Want to solve Laplace's equation: $\Delta f = g$. Solution given in terms of Δ^{-1}.

Quick way to see connection using spectral representation:

$$
\begin{aligned}
\Delta_{x,x'} &= \sum_i \mu_i \, \phi_i(x) \, \phi_i(x') \\
K_t(x,x') &= \sum_i e^{-t\mu_i} \, \phi_i(x) \, \phi_i(x') \\
\Delta^{-1}_{x,x'} &= \sum_i \frac{1}{\mu_i} \phi_i(x) \, \phi_i(x') = \int_0^\infty K_t(x,x') \, dt
\end{aligned}
$$

Used by Chung and Yau (2000).

Bounds on Covering Numbers and Generalization Error, Continuous Case

Eigenvalue bounds from differential geometry (Li and Yau):

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

Give bounds on SVM hypothesis class covering numbers

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

Bounds on Generalization Error

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

Question: Can diffusion kernel connection be exploited to get transductive generalization error bounds for random walks approach?

Summary

Random fields with discrete class labels—intractable, unstable

Continuous fields—tractable, more desirable behavior for segentation and labeling

Intimate connections with random walks, electric networks, graph flows, and diffusion kernels

Advantages/disadvantages?