Diffusion Kernels and Friends

Inducing kernels from the local structure of input space

Risi Kondor
Columbia University, New York, USA.

Collaborators

John Lafferty

Mikhail Belkin

Alex Smola

Tony Jebara

Count Laplace (1749-1827)

Regularization Networks

Learning from examples $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$

Looking for $f : \mathcal{X} \mapsto \mathcal{Y}$ minimizing

$$R_{\text{reg}}[f] = \underbrace{\frac{1}{m} \sum_{i=1}^{m} \overbrace{L(y_i, f(x_i))}^{\text{Loss function}}}_{\text{Empirical Risk}} + \underbrace{\left\langle \overline{Q} f, \overline{Q} f \right\rangle_{\mathcal{L}_2}}_{\text{Complexity penalty}}$$

Regularization operator: $\overline{Q}$

Inner product here: $\langle f, g \rangle_{\mathcal{L}_2} = \int f(x)\, g(x)\, dx$

The Kernel — A Similarity Measure

Feature map: $\Phi : \mathcal{X} \mapsto \mathcal{F}$

Hypothesis f linear in $\mathcal{F}$

Kernel: $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$ — similarity measure

Kernel K and regularization operator $\overline{Q}$ are related

Kernel must be positive definite, i.e.

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, K(x_i, x_j) \geq 0 \qquad \forall x_1, x_2, \ldots, x_n \in \mathcal{X}, \quad \forall \alpha_i \in \mathbb{R}$$

Correspondence between kernel and operator

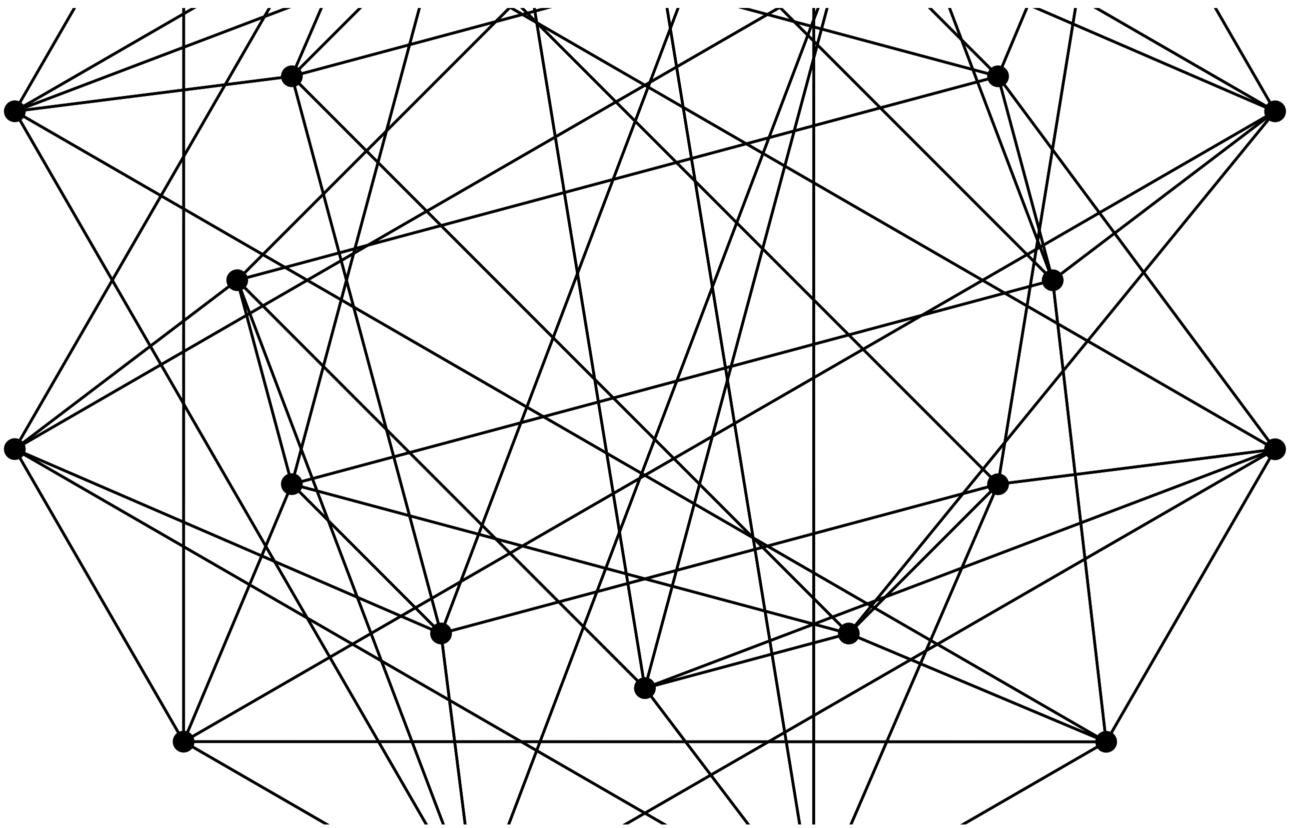
A typical kernel on $\mathcal{X} = \mathbb{R}^d$ is the Gaussian RBF

$$K(x, x') = \frac{1}{(2\pi\sigma^2)^{d/2}} \, e^{-\| x - x' \| / (2\sigma^2)}.$$

Regularization operator: $(\widehat{\overline{Q} f})(\omega) = e^{\sigma^2 \| \omega \|^2} \, \widehat{f}(\omega)$

Graphs

Graphs



Graphs

- Natural graphs: internet, web, social contacts, citations, scientific collaborations, etc.
- Objects with graph-like structure: groups (permutations), strings, etc.
- Objects with unknown global structure: set of organic molecules, interacting proteins in a cell
- Incorporating unlabelled data

Looking for positive definite $K : V \times V \mapsto \mathbb{R}$, now just a matrix

Try Random Walks

$$A_{ij} = \begin{cases} 1 & i \sim j \\ 0 & \text{otherwise} \end{cases}$$

A symmetric $\Rightarrow$ even powers pos. def.

$$K = A^2 \ \ ? \quad A^4 \ \ ? \quad A^\infty \ \ ?$$

$$K = \alpha_1 A^2 + \alpha_2 A^4 + \ldots \ \ ?$$

Diffusion Kernels

Infinite number of ininitesimal steps:

$$K = e^{\beta L} = \lim_{n \to \infty} \left(1 + \frac{\beta}{n} L \right)^n$$

$$L_{ij} = \begin{cases} 1 & i \sim j \\ -d_i & i = j \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Laplacian)}$$

For any symmetric L, $K = e^{\beta L}$ is positive definite.

$$e^{\beta L} = \lim_{n\to\infty} \left(I + \frac{\beta}{n} L\right)^n = \lim_{n\to\infty} \left(I + \frac{\beta}{2n} L\right)^{2n}$$

conversely,

$$K = \left(K^{1/n}\right)^n = \lim_{n\to\infty} \left(I + \frac{1}{n} L\right)^n = e^L$$

for any infinitely divisible (or finite) K.

Exponential Kernels

$$
\begin{aligned}
K_\beta \;&=\; e^{\beta L} = \lim_{n\to\infty} \left(I + \frac{\beta}{n}\,L\right)^n \\
&=\; I + \beta L + \frac{\beta^2}{2!}\,L^2 + \frac{\beta^3}{3!}\,L^3 + \ldots
\end{aligned}
$$

Exponential Kernels

$$K_\beta \;=\; e^{\beta L} = \;\lim_{n\to\infty}\left(I + \tfrac{\beta}{n}\,L\right)^n$$

$$=\; I + \beta L + \tfrac{\beta^2}{2!}\,L^2 + \tfrac{\beta^3}{3!}\,L^3 + \ldots$$

$$\frac{d}{d\beta}\,K_\beta = L\,K_\beta \qquad\qquad K_0 = I$$

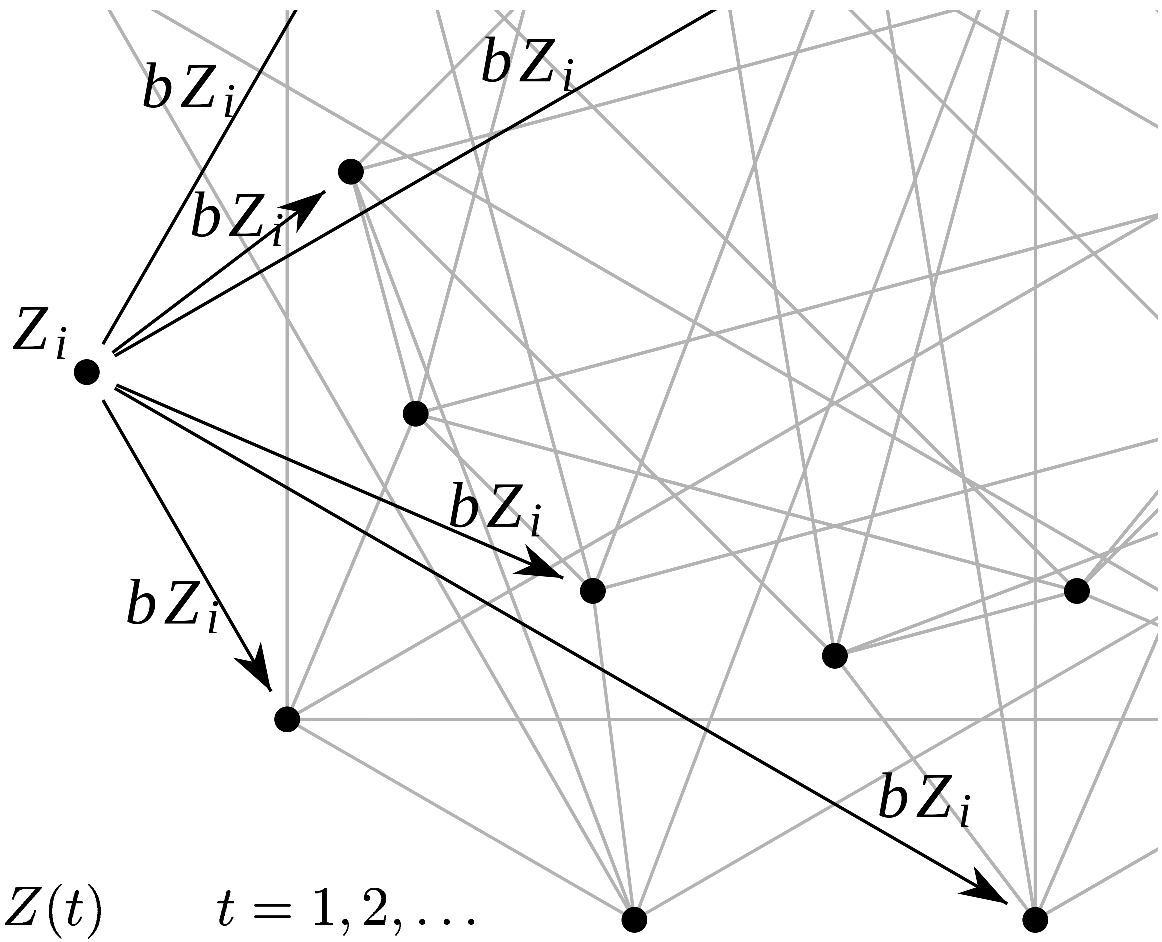
Diffusion kernels: view 2

$K = e^{\beta L}$

Stochastic field

$\mathsf{E}[Z_i(0)] = 0$
$\mathsf{Var}[Z_i(0)] = \sigma^2$
$Z_i(0)$ indep.

redistribution:
$Z(t+1) = (I + bL)\, Z(t) \qquad t = 1, 2, \ldots$



evolution operator $T(t) = (I + bL)^t \qquad Z(t) = T(t)Z(0)$

covariance

$$\mathrm{Cov}_{ij}(t) = \overline{Z_i(t)\, Z_j(t)} = \overline{\Big(\sum_{i'} T_{ii'}\, Z_{i'}(0)\Big)\Big(\sum_{j'} T_{jj'}\, Z_{j'}(0)\Big)}$$

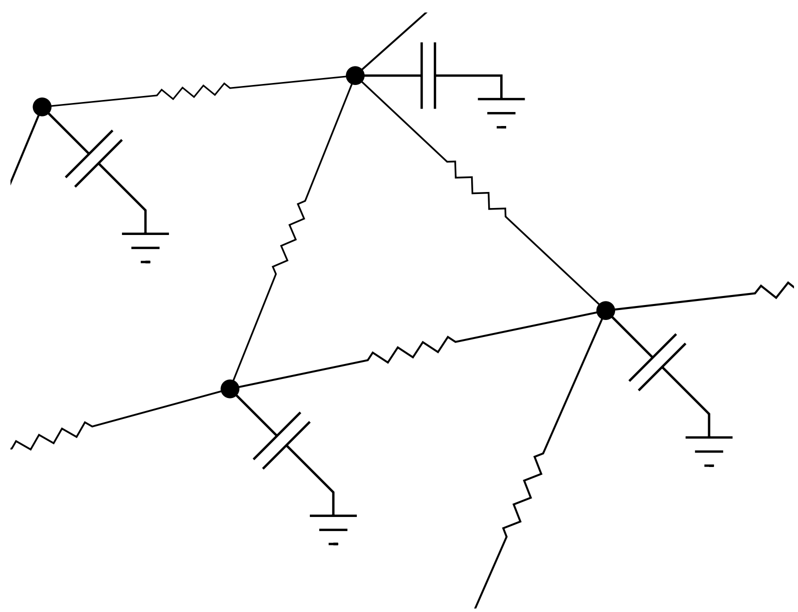
$$\mathbf{Cov}(t) = \sigma^2\, T(t)^T\, T(t) = \sigma^2\, T(t)^2 = \sigma^2\, (I + bL)^t$$

continuous limit: $\mathbf{Cov}(t) = \lim_{n\to\infty} \left(I + \frac{bL}{n}\right)^{2nt} = \sigma^2\, e^{\,2btL}$

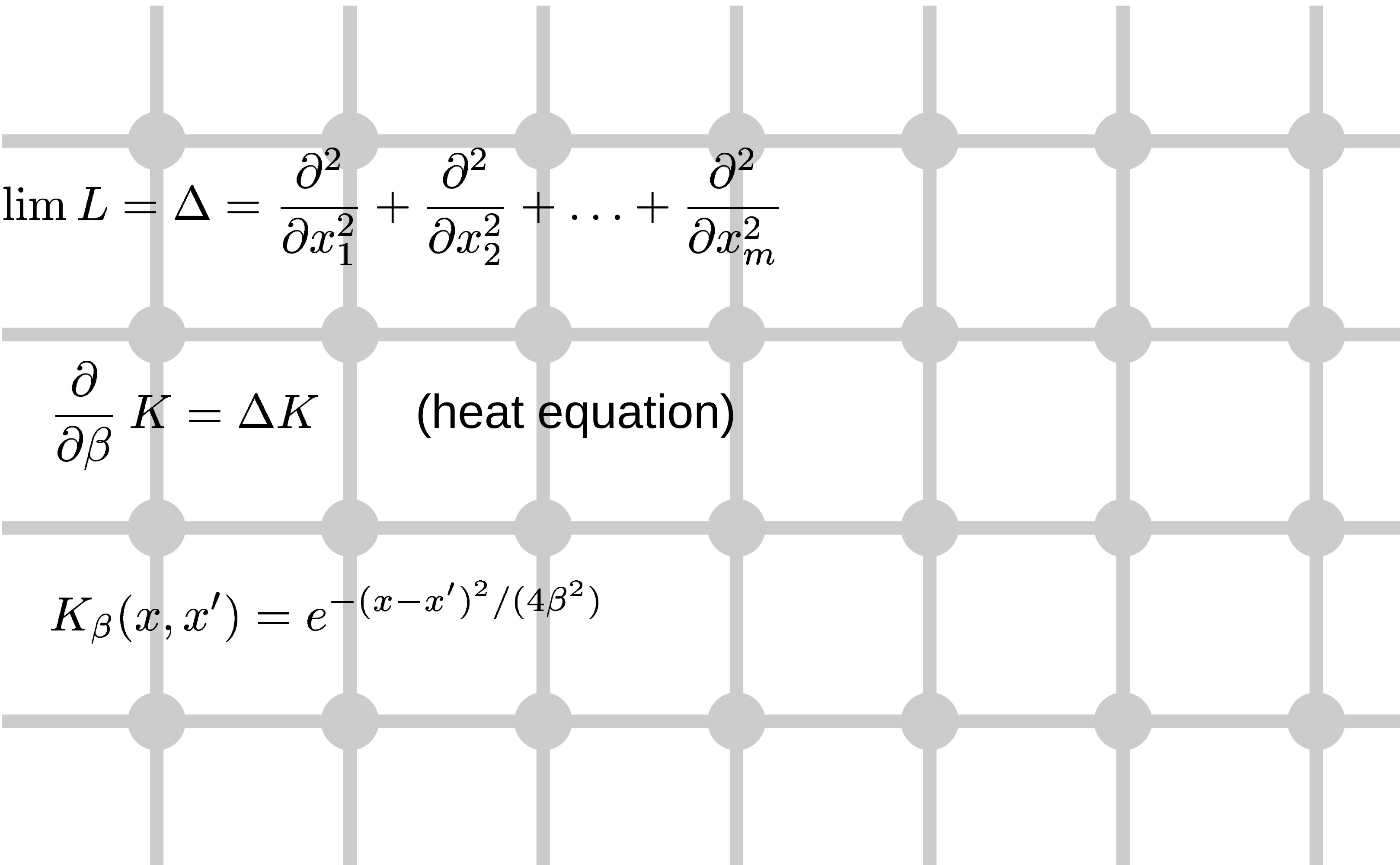
electrical analogy

$$\frac{d}{dt}\,Z_i(t) = b\,\sum_{i\sim j}(Z_j(t) - Z_i(t))$$

$$\frac{d}{dt}\,U_i(t) = \frac{1}{RC}\,\sum_{i\sim j}(U_j(t) - U_i(t))$$



Diffusion kernels: view 3



$$\lim L = \Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \ldots + \frac{\partial^2}{\partial x_m^2}$$

$$\frac{\partial}{\partial \beta} K = \Delta K \qquad \text{(heat equation)}$$

$$K_\beta(x, x') = e^{-(x-x')^2/(4\beta^2)}$$

Diffusion kernels: view 4

Spectral graph theory

Eigenvectors of L correspond to “normal modes” of graph

approximate min-cut, max. distance type results

Computing the diffusion kernel

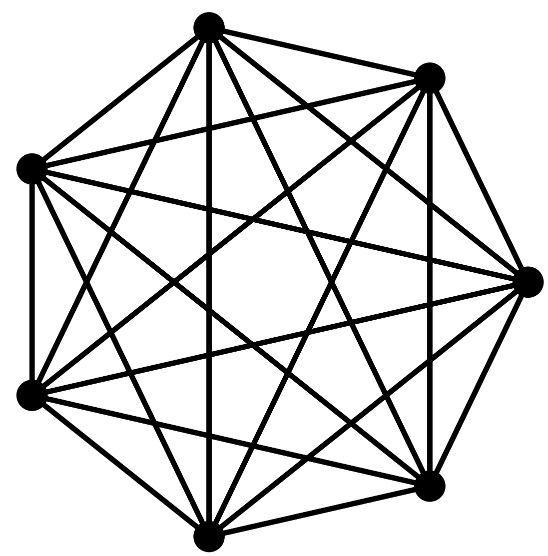
Diagonalization

$$L = T^{-1} D\, T$$

$$e^{\beta L} = T^{-1} e^{\beta D}\, T$$

Complete graphs

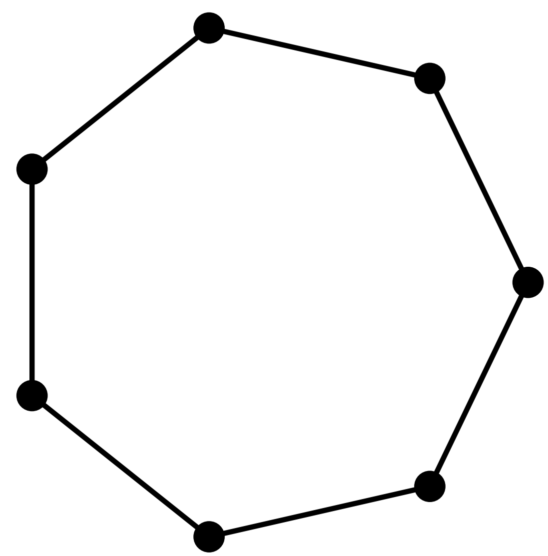
$$K(i,j) = \begin{cases} \dfrac{1+(n-1)\,e^{-n\beta}}{n} & \text{for } i = j \\[2ex] \dfrac{1-e^{-n\beta}}{n} & \text{for } i \neq j\ , \end{cases}$$



for $n = 2$ $\qquad K_\beta(i,j) \propto (\tanh\beta)^{d(i,j)}$

Closed chains

$$K(i,j) = \frac{1}{n}\sum_{\nu=0}^{n-1} e^{-\omega_\nu \beta} \cos \frac{2\pi\nu(i-j)}{n}$$



Tensor product kernels

$K^{(1)}$ kernel on $\mathcal{X}_1$
$K^{(2)}$ kernel on $\mathcal{X}_2$

$K^{(1,2)} = K^{(1)} \otimes K^{(2)}$ kernel on $\mathcal{X}_1 \otimes \mathcal{X}_2$

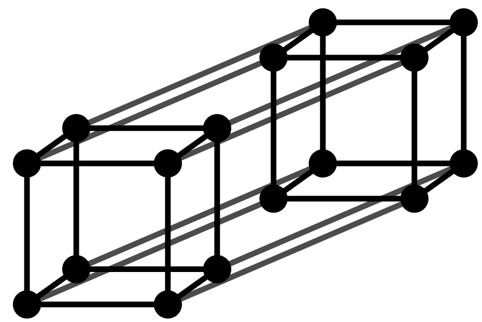
$$K^{(1,2)}\left((x_1, x_2), (x_1', x_2')\right) = K^{(1)}(x_1, x_1')\, K^{(2)}(x_2, x_2')$$

$$L^{(1,2)} = L^{(1)} \otimes I^{(2)} + L^{(2)} \otimes I^{(1)}$$

Hypercubes, etc.

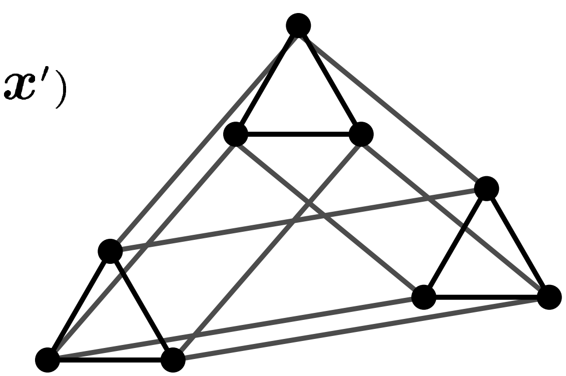
Hypercube:

$$K(x, x') = (\tanh \beta)^{d(x,x')}$$

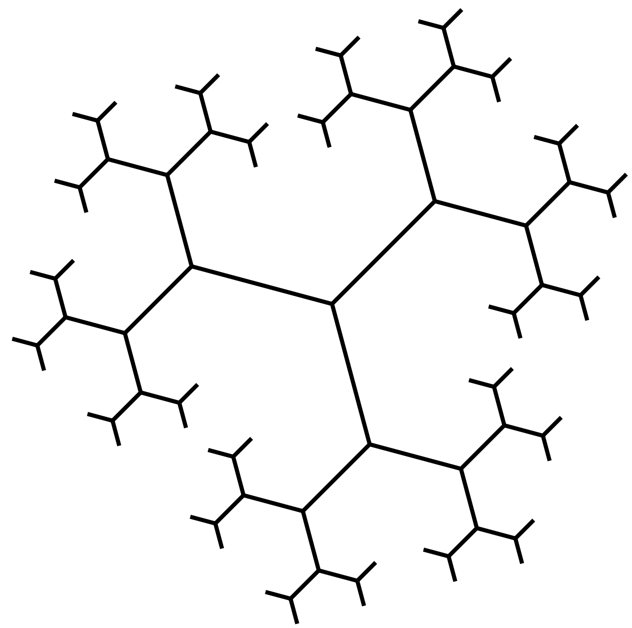


Alphabet $\mathcal{A}$:

$$K(x, x') = \left(\frac{1 - e^{-|\mathcal{A}|\beta}}{1 + (|\mathcal{A}| - 1)\, e^{-|\mathcal{A}|\beta}} \right)^{d(\boldsymbol{x},\boldsymbol{x}')}$$



k-regular-trees



$$K(x,x') = K(d(x,x')) =$$

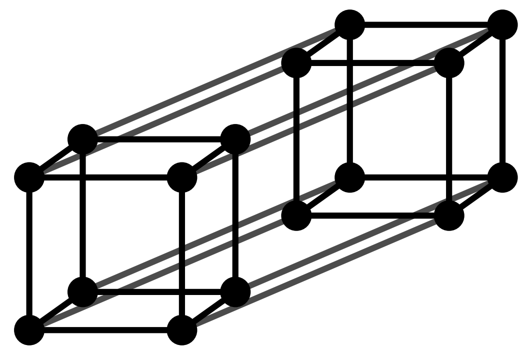
$$\frac{2}{\pi(k-1)} \int_0^{\pi} \frac{e^{-\beta\left(1-\frac{2\sqrt{k-1}}{k}\cos x\right)} \sin x \left[\,(k-1)\sin(d+1)\,x - \sin(d-1)\,x\,\right]}{k^2 - 4\,(k-1)\,\cos^2 x}\,dx$$

Applications

Experiments

Voted Perceptron algorithm (Freund and Schapire, 1999)

5 standard categorical datasets



Diffusion kernels in Bioinformatics

Graph-driven feature extraction from microarray data using diffusion kernels and kernel CCA (J.-P. Vert and M. Kanehisa, NIPS 2002)

Diffusion kernel on network of chemical pathways, genes catalyzing them.

A more general framework

Looking for $f : \mathcal{X} \mapsto \mathcal{Y}$ minimizing

$$R_{\text{reg}}[f] = \frac{1}{m} \sum_{i=1}^{m} L(y_i, f(x_i)) + \langle \overline{Q} f, \overline{Q} f \rangle$$

Linear Operators on $\mathcal{L}_2(\mathcal{X})$

$$f(x) = \sum_i \alpha_i \, K(x, x_i) \quad \rightarrow \quad f(x) = (\overline{K} g)(x) = \int_{\mathcal{X}} K(x, x') \, g(x') \, dx'$$

For K symmetric, positive definite, $\overline{K}$ is self-adjoint ($\langle g, \overline{K} g \rangle = \langle \overline{K} g, g \rangle$) and positive ($\langle g, \overline{K} g \rangle \geq 0$).

$$\langle w, w \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j)$$

$$\downarrow$$

$$\langle \overline{Q} f, \overline{Q} f \rangle = \int_{\mathcal{X}} \int_{\mathcal{X}} g(x) \, K(x, x') \, g(x') \, dx' \, dx = \langle g, \overline{K} g \rangle$$

Linear Operators on $\mathcal{L}_2(\mathcal{X})$

$$\langle \overline{Q}\, f, \overline{Q}\, f \rangle = \langle g, \overline{K} g \rangle$$

$$\langle \overline{Q}\, \overline{K} g, \overline{Q}\, \overline{K} g \rangle = \langle g, \overline{K} g \rangle$$

$$\overline{Q} = \overline{K}^{\,-1/2}$$

$\overline{K}$ and $\overline{Q}$ share an eigensystem $v_1, v_2, \ldots$ and their eigenvalues are related by $\lambda_i^{(Q)} = \lambda_i^{(K)\,-1/2}$ (Girosi, Jones and Poggio, 1995) .

Locality and invariance

$\overline{Q}$ is supposed to capture roughness of f . Want to make it a local differential operator.

Essentially unique second order differential operator on a d dimensional linear space or manifold $\mathcal{X}$ invariant under isometries is the Laplacian $\Delta : \mathcal{L}_2(\mathcal{X}) \mapsto \mathcal{L}_2(\mathcal{X})$

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \ldots + \frac{\partial^2}{\partial x_d^2}.$$

Natural building block for $\overline{Q}$ and hence $\overline{K}$.

The Gaussian Kernel

$$K_\beta(x,x') = \tfrac{1}{\sqrt{2\pi\beta}}\, e^{-\|\, x-x'\,\|^2/(2\beta)} \quad \rightarrow \quad \tfrac{\partial}{\partial\beta} K_\beta(x,x_0) = \Delta\, K_\beta(x,x_0).$$

$$(\overline{K}_\beta\, g)(x) = \int_{\mathcal{X}} K_\beta(x,x')\, g(x')\, dx'$$

$$\frac{\partial}{\partial\beta}(\overline{K}_\beta\, g)(x) = \int_{\mathcal{X}} \frac{\partial}{\partial\beta} K_\beta(x,x')\, g(x')\, dx'$$

$$\frac{\partial}{\partial\beta}(\overline{K}_\beta\, g)(x) = \int_{\mathcal{X}} \Delta K_\beta(x,x')\, g(x')\, dx'$$

$$\frac{d}{d\beta}\overline{K}_\beta = \overline{\Delta}\,\overline{K}_\beta$$

$\overline{K}_\beta = e^{\beta\Delta}$ $\qquad$ $\overline{Q}_\beta = e^{-\beta\Delta/2}$ $\qquad$ Just a diffusion kernel!

Diffusion Kernels on Manifolds

Belkin and Niyogi (NIPS 2001, 2002)

$$\overline{K} = e^{\beta \Delta}$$

on a Riemannian manifold found **from the data**

Natual way of incorporating unlabelled data.

Diffusion Kernels on the Statistical Manifold

Information Diffusion Kernels (Lafferty and Lebanon NIPS 2002)

parametric family $p(x|\theta)$ gives rise to manifold with Fisher metric

$$[G_\theta]_{i,j} = \mathsf{E}[\partial_i \ell_\theta \, \partial_j \ell_\theta] = \int_{\mathcal{X}} \partial_i \log p(x|\theta) \, \partial_j \log p(x|\theta) \, p(x|\theta) \, dx$$

For $x_i \in \mathcal{X}$, compute corresponding $\hat{\theta}_i$ (say, MLE); then use diffusion kernel on manifold between $\{\hat{\theta}_i\}$

From continuous to discrete

$$\langle g, \Delta g \rangle = \int_{\mathcal{X}} g(x)\, \Delta g(x)\, dx = -\int_{\mathcal{X}} (\nabla g(x)) \cdot (\nabla g(x))\ dx$$

$$\downarrow$$

$$g\, L\, g = -\sum_{i \sim j} (g_i - g_j)^2$$

$$L_{ij} = \begin{cases} 1 & i \sim j \\ -d_i & i = j \\ 0 & \text{otherwise} \end{cases}$$

Other locally induced kernels

Smola and Kondor (under review)

$$\Delta = \sum_i \lambda_i \, v_i \, v_i^\top$$

$$K = \sum_i r(\lambda_i) \, v_i \, v_i^\top$$

- Diffusion kernel: $r(\lambda) = \exp(\beta\lambda)$
- p-step random walk kernel: $-(a+\lambda)^{-p}$
- Regularized Laplacian kernel: $\sigma^2\lambda - 1$

Conclusions

- Inducing a kernel from the local (differential) structure of $\mathcal{X}$ makes sense from a regularization theory point of view

- Inducing a kernel from the local (differential) structure of $\mathcal{X}$ makes sense from a regularization theory point of view
- A unifying concept: the Laplacian

- Inducing a kernel from the local (differential) structure of $\mathcal{X}$ makes sense from a regularization theory point of view
- A unifying concept: the Laplacian
- Diffusion kernels physically motivated, others are possible

- Inducing a kernel from the local (differential) structure of $\mathcal{X}$ makes sense from a regularization theory point of view
- A unifying concept: the Laplacian
- Diffusion kernels physically motivated, others are possible
- These kernels are not necessarily easy to compute in practice