Diffusion Kernels

Risi Imre Kondor
with John Lafferty

Center for Automated Learning and Discovery
School of Computer Science
Carnegie Mellon University

Kernels

- $K(x, x') = \langle x, x' \rangle^p$
- $K(x, x') = (1 + \langle x, x' \rangle)^p$
- $K(x, x') = e^{-\| x - x' \|^2 / (2\sigma^2)}$
- $K(x, x') = \tanh(\kappa \langle x, x' \rangle - \delta)$

Kernel-based algorithms

- Support Vector Machines (SVMs)
- Kernel Regression
- Kernel Principal Component Analysis
- Kernel Independent Component Analysis
- Gaussian Processes
- etc.

The learning framework

Data: $((x_i, y_i))_{i=1,2,\ldots,n}$ $\quad x_i \in \mathcal{X}$

Task: predict y from x

The Kernel

$K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ symmetric and positive definite, ie.

$$\sum_{i}^{n} \sum_{j}^{n} c_i \, c_j \, K(x_i, x_j) \geq 0 \qquad \begin{array}{l} \forall \, x_1, x_2, \ldots, x_n \in \mathcal{X} \\ \forall \, c_1, c_2, \ldots, c_n \in \mathbb{R} \end{array}$$

$$\int_\Omega \int_\Omega f(x_1) \, f(x_2) \, K(x_1, x_2) \, dx_1 \, dx_2 \geq 0 \qquad \forall f \in L_2(\Omega) \quad (Mercer)$$

What makes a good kernel?

1. encapsulates prior information
2. positive definite

Kernels: view 1

$\phi : \mathcal{X} \mapsto \mathcal{H}$

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$$

" K induces an isometric embedding of $\mathcal{X}$ in the Hilbert space $\mathcal{H}$"

"Enriching the hypothesis space"

example: SVMs

Kernels: view 2

Transformation to basis of independent components based on hypothetised covariance structure

For Gaussian processes

$$K(x, x') = \mathsf{Cov}(y_x, y_{x'})$$

Reminder: Gaussian Processes

Kernels: view 3

equivalence to regularization networks

minimize $\quad R_{\text{reg}}[f] = R_{\text{emp}}[f] + \Omega[f]$

$$R_{\text{emp}} = \textstyle\sum_i |\, y_i - f(x_i) \,|_\epsilon = \sum_i \max\left(\, 0 \,,\; |\, y_i - f(x_i) \,| - \epsilon \,\right) \qquad \text{(SVM)}$$

$$\Omega[f] = \int_{\mathcal{X}} \sum_{j=1}^{\infty} \frac{\sigma^{2j}}{j!\, 2^j} \left(\Delta^j f(x)\right)^2 dx \qquad \text{(Gaussian kernel)}$$

What makes a good kernel?

1. encapsulates prior information
2. positive definite

How can we construct kernels on discrete structures?

Exponential kernels

$$K_\beta \;=\; e^{\beta H} \;=\; \lim_{n\to\infty}\left(1+\tfrac{\beta}{n}\,H\right)^n$$

$$=\; 1+\beta H+\tfrac{\beta^2}{2!}\,H^2+\tfrac{\beta^3}{3!}\,H^3+\ldots$$

Exponential kernels

$$K_\beta \;=\; e^{\beta H} \;=\; \lim_{n\to\infty}\left(1+\tfrac{\beta}{n}\,H\right)^n$$

$$=\; 1+\beta H+\tfrac{\beta^2}{2!}\,H^2+\tfrac{\beta^3}{3!}\,H^3+\ldots$$

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

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

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

conversely,

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

for any infinitely divisible K.

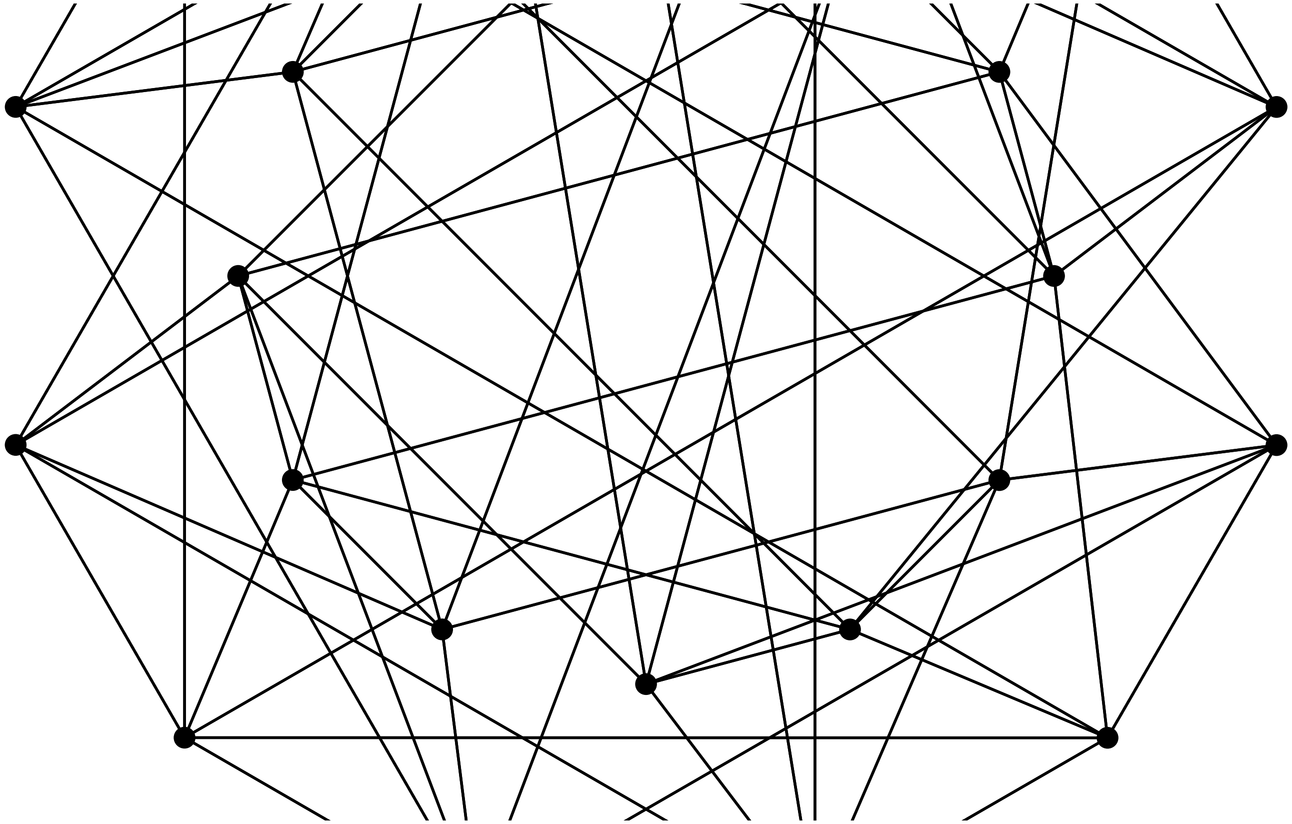
Exponential families

$$\{\ K_\beta = e^{\beta H}\ :\ \beta \in \mathbb{R}^+\ \}$$

$$\frac{d}{d\beta}\, K_\beta = H\, K_\beta \qquad\qquad K_1 = I$$

How shall we choose H?

Graphs



Diffusion kernels

$$K_\beta = e^{\beta H}$$

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

(Laplacian)

Diffusion kernels: view 1

Random walk with transition matrix $I + \beta H/n$

After n steps

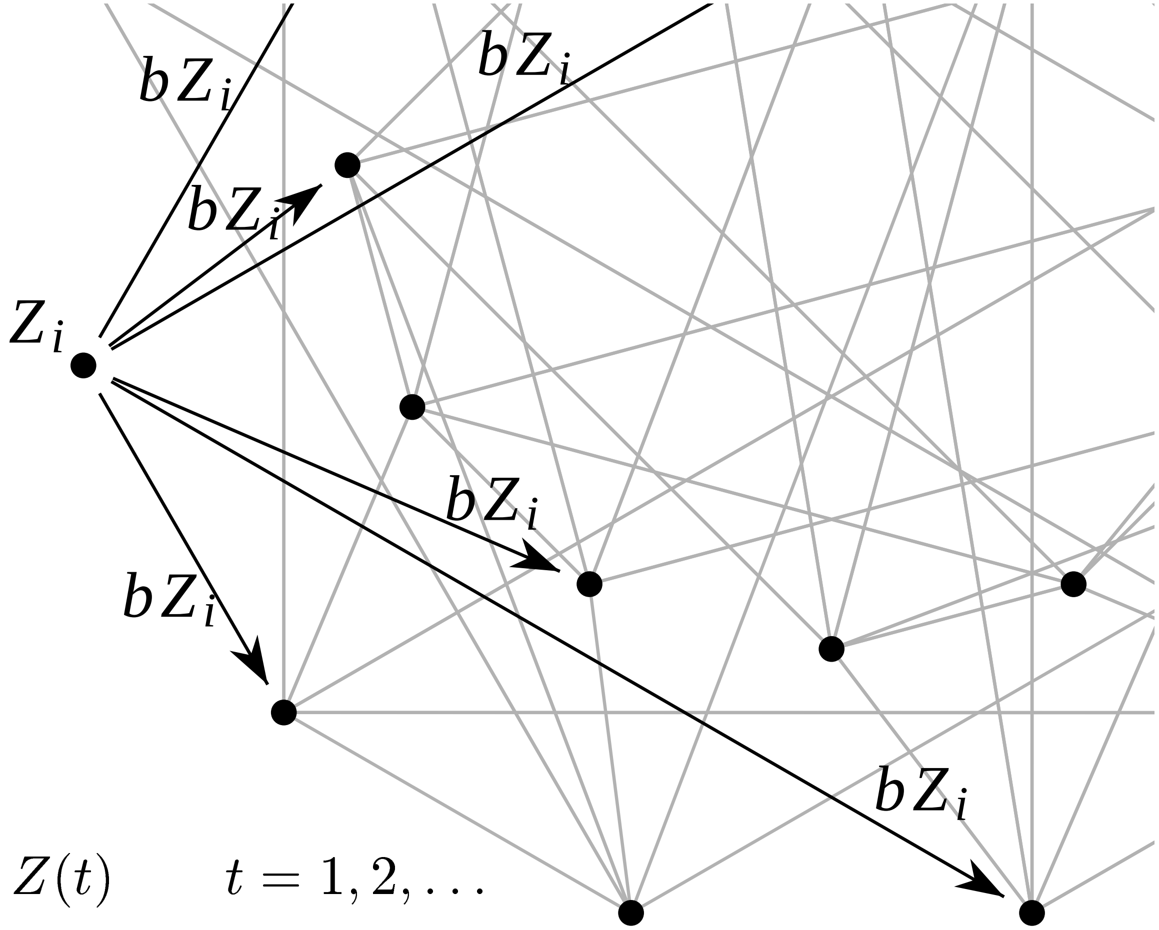
$$p_n = \left(I + \frac{\beta}{n} H \right)^n p_0$$

"lazy" random walk

Diffusion kernels: view 2

Stochastic field

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



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

evolution operator $T(t) = (1 + bH)^t$

$$Z(t) = T(t) Z(0)$$

covariance

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

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

$$T(t) = \lim_{n \to \infty} \left(1 + \frac{bH}{n}\right)^{nt} = e^{btH}$$

$$\mathbf{Cov}(t) = \sigma^2 \, e^{\,2btH}$$

Diffusion kernels: view 3

electronic analogy

$$Z(t) = e^{btH} Z(0) \qquad\qquad \frac{d}{dt}\, e^{btH} = bH$$

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

Diffusion kernels: view 4

generalized Gaussian kernel

on square grid in $\mathbb{R}^m$, $\quad \lim H = \Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \ldots + \frac{\partial^2}{\partial x_m^2}$

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

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

Diffusion kernels: view 5

Spectral graph theory

Eigenvectors of H correspond to “normal modes” of graph

approximate min-cut, max. distance type results

How to compute the diffusion kernel

Diagonalization

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

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

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 $\mathcal{X}_1 \otimes \mathcal{X}_2$

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

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

Hypercubes, etc.

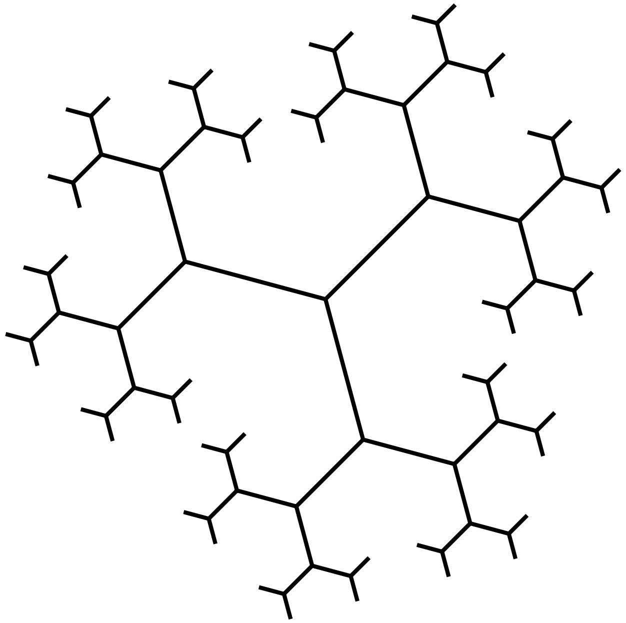
Hypercube:

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

Alphabet $\mathcal{A}$:

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

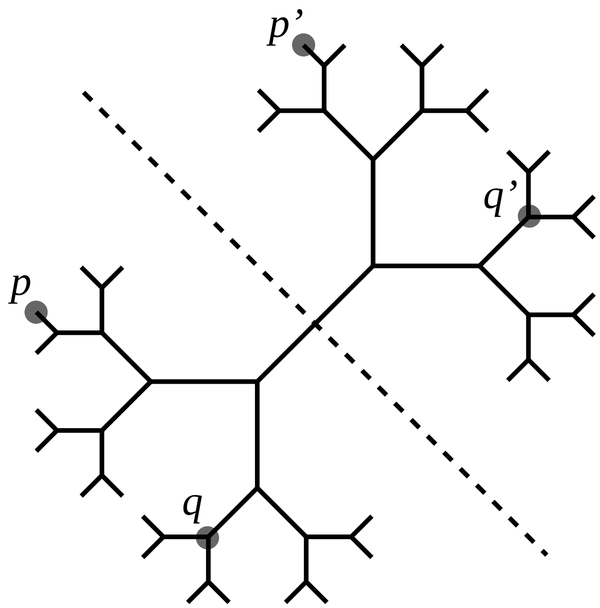
k-regular-trees



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

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

Binary trees



$$K(x, x') = \frac{1}{2} K_{\mathsf{R}}^{(3)}(d(x, x')) + \frac{1}{2} K_{\mathsf{R}}^{(3)}(d(x, r) + d(r, x') + 1)$$

Applications

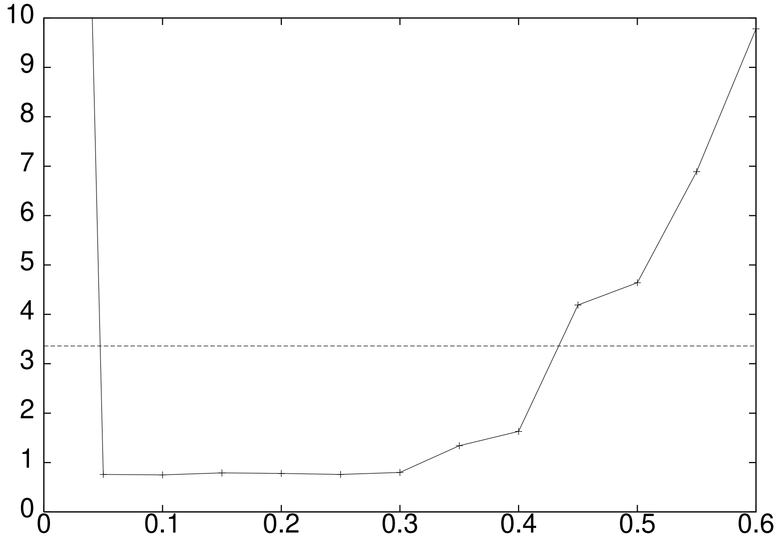
Learning on graphs

- 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
- Incorporating unlabelled data

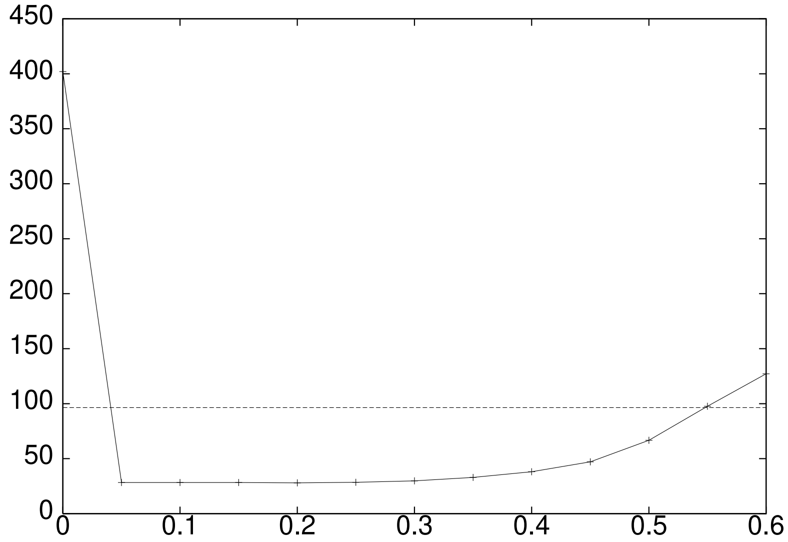
Experiments

Results

			Hamming Distance		Diffusion Kernel		
	#Attr	$\max \vert \mathcal{A} \vert$	error	$\vert SV \vert$	error	$\vert SV \vert$	β
`Breast Cancer`	9	10	7.64%	387.0	3.64%	62.9	0.30
`Hepatitis`	13	2	17.98%	750.0	17.66%	314.9	1.50
`Income`	11	42	19.19%	1149.5	18.50%	1033.4	0.40
`Mushroom`	22	10	3.36%	96.3	0.75%	28.2	0.10
`Votes`	16	2	4.69%	286.0	3.91%	252.9	2.00



Error vs. β



$|$SV$|$ vs. β

Conclusions

$$\text{Data} \quad \longrightarrow \quad \mathbb{R}^m \quad \longrightarrow^{\phi} \quad \mathcal{H}$$

References