Learning with Kernels on Graphs, Groups and Manifolds

Risi Kondor
Columbia University, New York, USA.

Collaborators

John Lafferty

Guy Lebanon

Mikhail Belkin

Alex Smola

Tony Jebara

Count Laplace (1749-1827)

Batch Learning

Input space: $\mathcal{X}$ e.g. $\mathcal{X} = \mathbb{R}^d$

Output space: $\mathcal{Y}$

$\mathcal{Y} = \mathbb{R}$ regression

$\mathcal{Y} = \{-1, 1\}$ classi£cation

Learn $f : \mathcal{X} \mapsto \mathcal{Y}$ from examples $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$

A Naive Approach

Look for f minimizing

$$R_{\text{reg}}[f] = \frac{1}{m}\sum_{i=1}^{m} \underbrace{(f(x_i) - y_i)^2}_{\text{Loss function}} + \underbrace{\Omega[f]}_{\text{Complexity penalty}}$$

Harmonic example: $\Omega[f] = \int e^{\|\,\omega\,\|^2/2} \|\,\hat{f}(\omega)\,\|^2 \, d\omega$

$$f(x) = \frac{1}{\sqrt{2\pi}^{\,d}} \int \hat{f}(\omega) e^{i\omega\cdot x}\, d\omega \qquad \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}^{\,d}} \int f(x) e^{-i\omega\cdot x}\, dx$$

A space of functions

The functions f naturally form a linear space $\mathcal{H}$

Impose inner product such that $\langle f, f \rangle = \Omega[f]$

e.g. $\langle f, f' \rangle = \int e^{\|\,\omega\,\|^2/2} \overline{\hat{f}(\omega)} \hat{f}'(\omega)\, d\omega$

If $\mathcal{H}$ is complete, it is said to be a Hilbert space.

Looking for $f \in \mathcal{H}$ minimizing

$$R_{\mathsf{reg}}[f] = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \langle f, f \rangle$$

Looking for $f \in \mathcal{H}$ minimizing

$$R_{\mathsf{reg}}[f] = \frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - y_i\right)^2 + \langle f, f\rangle$$

Wouldn't it be neat if $\langle f, k_x\rangle = f(x)$, then

$$R_{\mathsf{reg}}[f] = \frac{1}{m}\sum_{i=1}^{m}\left(\langle f, k_{x_i}\rangle - y_i\right)^2 + \langle f, f\rangle$$

f only interacts with k_{x_i}'s $\quad \Rightarrow \quad f \in \mathsf{span}(k_{x_1}, k_{x_2}, \ldots, k_{x_m})$

Now plug in $f(x) = \sum_i \alpha_i k_{x_i}(x)$:

$$R_{\mathsf{reg}}[f] = \frac{1}{m} \sum_{i=1}^{m} (\langle f, k_{x_i} \rangle - y_i)^2 + \langle f, f \rangle$$

$$\downarrow$$

$$R_{\mathsf{reg}}[f] = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\alpha_j \left\langle k_{x_j}, k_{x_i} \right\rangle - y_i\right)^2 + \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \left\langle k_{x_i}, k_{x_j} \right\rangle$$

Letting $K_{ij} = \langle k_i, k_j \rangle$:

$$R_{\mathsf{reg}} = (K\alpha - y)^\top (K\alpha - y) + \alpha^\top K \alpha$$

To £nd $f(x) = \sum_i \alpha_i k_{x_i}(x)$ minimizing

$$R_{\text{reg}} = (K\alpha - y)^\top (K\alpha - y) + \alpha^\top K\alpha$$

set $\dfrac{\partial R}{\partial \alpha} = 0$:

$$2K(K\alpha - y) + 2K\alpha = 0$$

$$\boxed{\alpha = (K + I)^{-1} y}$$

So what are k_x and K?

Recall $\langle f, k_x \rangle = f(x)$. It is easy to show that for our example

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

and

$$K_{i,j} = \left\langle k_{x_i}, k_{x_j} \right\rangle = k_{x_i}(x_j) = \frac{1}{\sqrt{2\pi}^{\,d}} \, e^{-\| \, x_i - x_j \, \|^2/2}$$

K is the kernel!!!

Kernel Methods

General regularization network form:

$$f = \arg\min_{f \in \mathcal{H}} \left[\underbrace{\frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i)}_{\text{Empirical risk}} + \underbrace{\langle f, f \rangle}_{\text{Regularizer}} \right]$$

SVM classi£cation: $L = \max(0, 1 - y_i f(x_i))$

SVM regression: $L = |\, f(x_i) - y_i \,|_\epsilon$

Gaussian Process MAP $L = \frac{1}{\sigma_0^2}(f(x_i) - y)^2$

Conventional Explanation

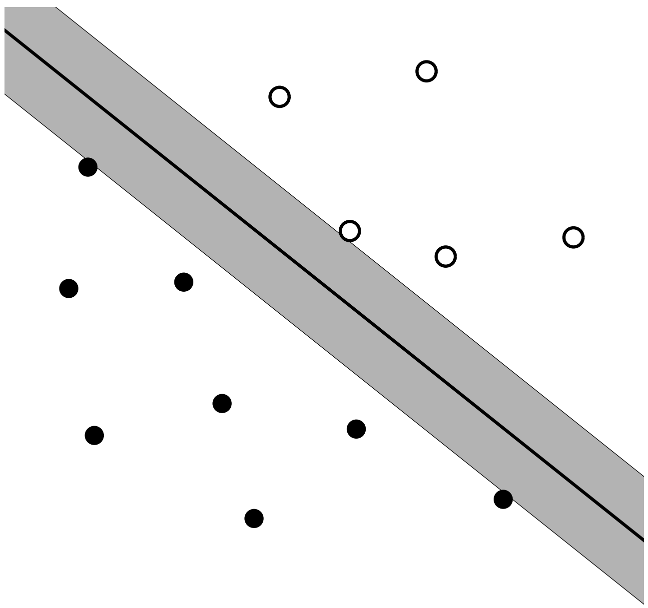
$K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ positive de£nite function: similarity measure

There exists some mapping $\Phi : \mathcal{X} \mapsto \mathcal{H}$ such that

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

Find some geometric criterion to optimize in $\mathcal{H}$, eg. maximum margin.

Nonlinear SVM's



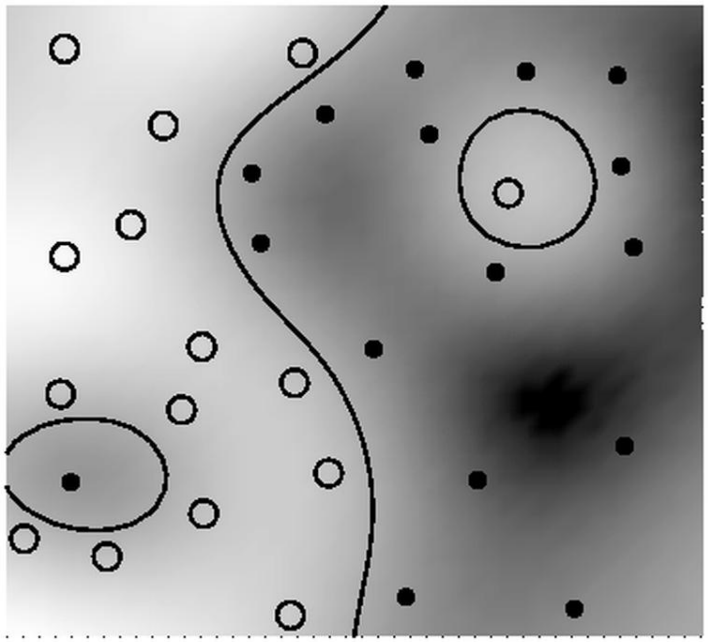


Figure courtesy of B. Schölkopf and A. Smola, ©MIT Press

Connection between two views

- One-to-one correspondence between kernel and regularizer
- Kernel is algorithmic
- Form of regularizer really explains what's going on, e.g.

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

$$\updownarrow$$

$$\langle f, f \rangle = \frac{1}{\sqrt{2\pi}^d} \int e^{\| \omega \|^2 \sigma^2 / 2} \| \hat{f}(\omega) \|^2 \, d\omega \quad \text{roughening}$$

Connection in general

Regularization operator $\overline{P} : \mathcal{H} \mapsto \mathcal{H}$ (self-adjoint)

$$\Omega[f] = \langle f, f \rangle = \int (\overline{P} f)(x) \cdot (\overline{P} f)(x) \, dx$$

Kernel operator $\overline{K} : \mathcal{H} \mapsto \mathcal{H}$

$$(\overline{K} g)(x) = \int K(x, x') g(x') \, dx' \qquad \Rightarrow \qquad (\overline{K} \delta_x)(x') = K(x, x')$$

$$\langle f, k_x \rangle = \langle f, K(x, \cdot) \rangle = \int f(x') (\overline{P}^2 \overline{K} \delta_x)(x') \, dx' = f(x)$$

hence $\boxed{\overline{K} = \left(\overline{P}\right)^{-2}}$

References so far

Girosi, Jones & Poggio Regularization Theory and Neural Network Architectures (Neural Computation, 1995)

Smola and Schölkopf From Regularization Operators to Support Vector Kernels (NIPS 1998)

Aronszajn Theory of Reproducing Kernels (1950)

Kimeldorf & Wahba Some Results on Tchebychef£an Spline Functions (1971)

⋮

Link to Diffusion

$$K_\beta(x,x') = \frac{1}{\sqrt{2\pi}\beta}\, e^{-\|\, x-x'\, \|/(2\beta)}$$

is solution to Diffusion equation

$$\frac{\partial}{\partial\beta} K_\beta(x,x_0) = \Delta\, K_\beta(x,x_0)$$

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

Generally, may de£ne Laplacian operator $\overline{\Delta} : \mathcal{H} \mapsto \mathcal{H}$ and kernel operator by

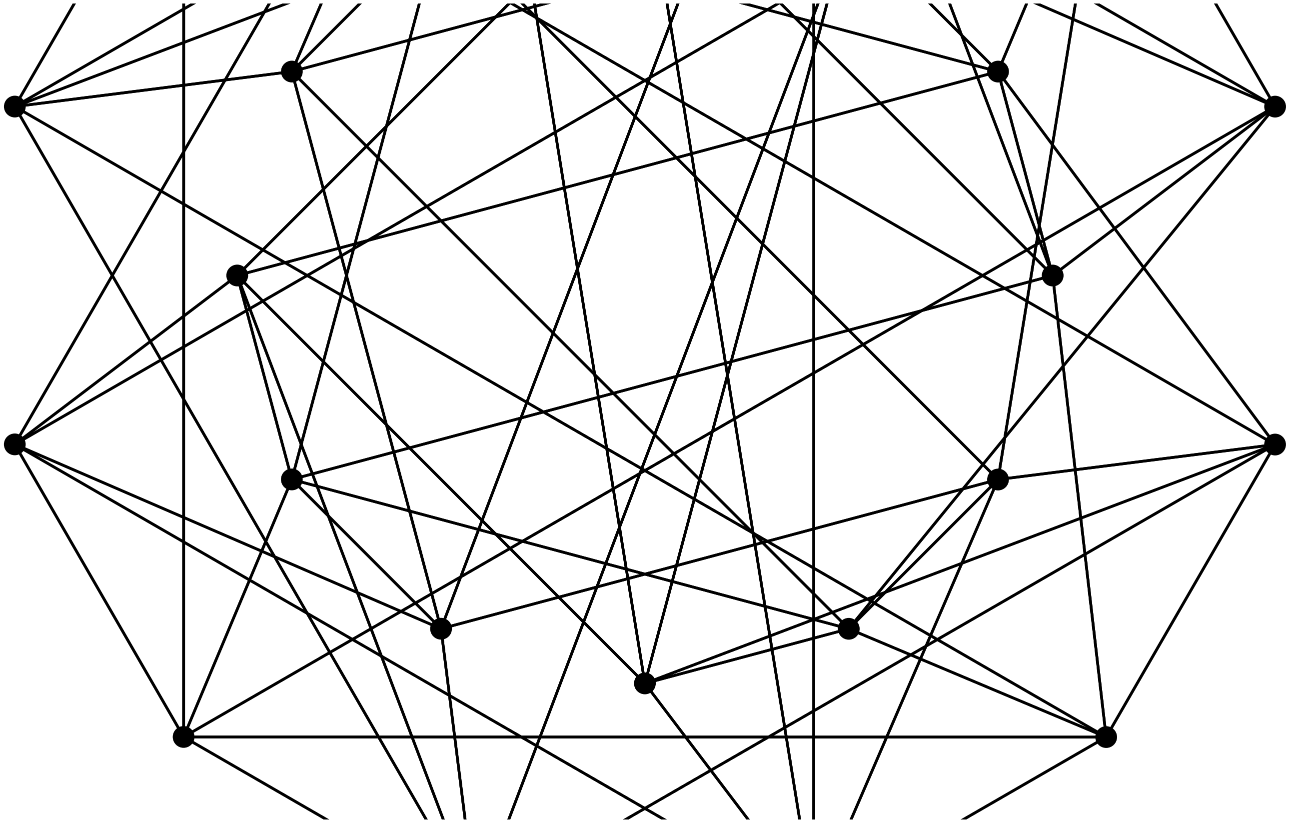
$$\frac{\partial}{\partial\beta}\, \overline{K}_\beta = \overline{\Delta K}_\beta.$$

Formally $\overline{K} = e^{\beta\overline{\Delta}}$ and $\overline{P} = e^{-\beta\overline{\Delta}/2}$.

How do we generalize to discrete spaces?

1. Graphs

Graphs



Looking for positive de£nite $K : V \times V \mapsto \mathbb{R}$, 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

In£nite number of ininitesimal steps:

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

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

(Laplacian)

Diffusion

In£nite number of ininitesimal steps:

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

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

Exponential Kernels

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

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

Exponential Kernels

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

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

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

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

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

conversely,

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

for any in£nitely divisible (or £nite) K.

Properties of Diffusion Kernels

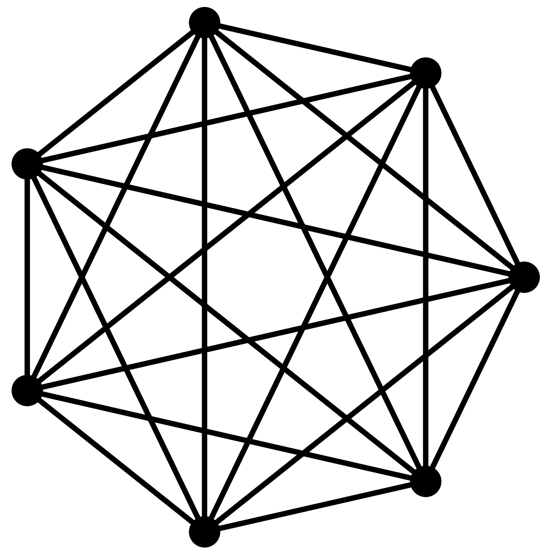
- Positive de£nite
- Analogy with continuous case
- Local relationships L induce global relationships K_β by

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

- Works for undirected weighted graphs with weights $w_{ij} = w_{ji}$

Complete graphs

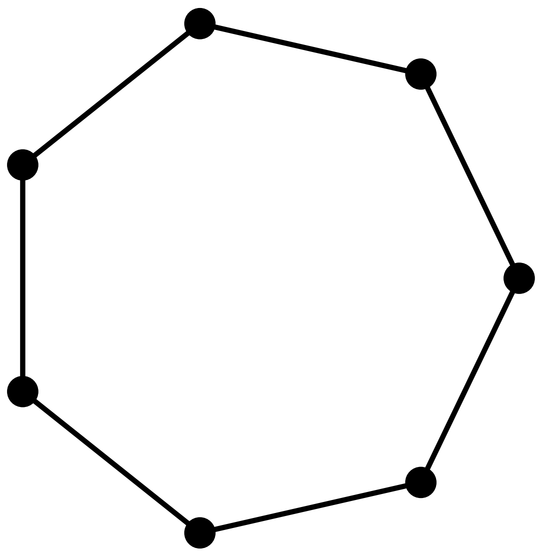
$$K(i,j) = \begin{cases} \dfrac{1 + (n-1)\, e^{-n\beta}}{n} & \text{for } i = j \\[2ex] \dfrac{1 - e^{-n\beta}}{n} & \text{for } i \neq j\ , \end{cases}$$



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

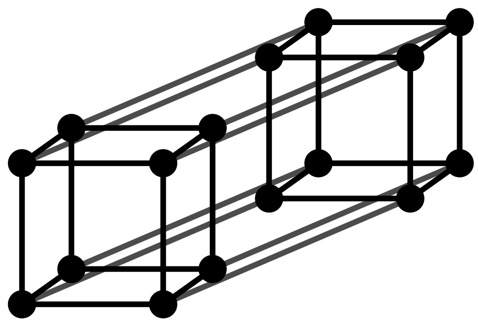
$$K^{(1,2)}\left((x_1, x_2), (x_1', x_2')\right) = K^{(1)}(x_1, x_1')\, K^{(2)}(x_2, x_2')$$

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

Hypercubes, etc.

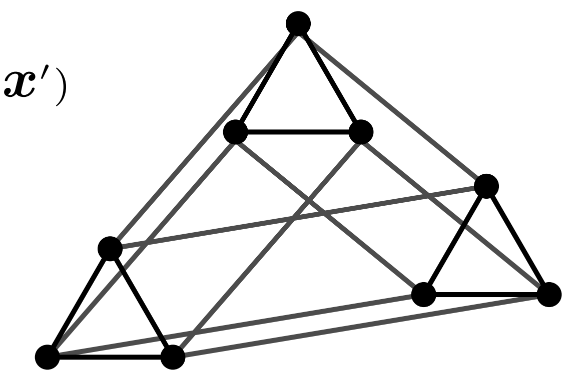
Hypercube:

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

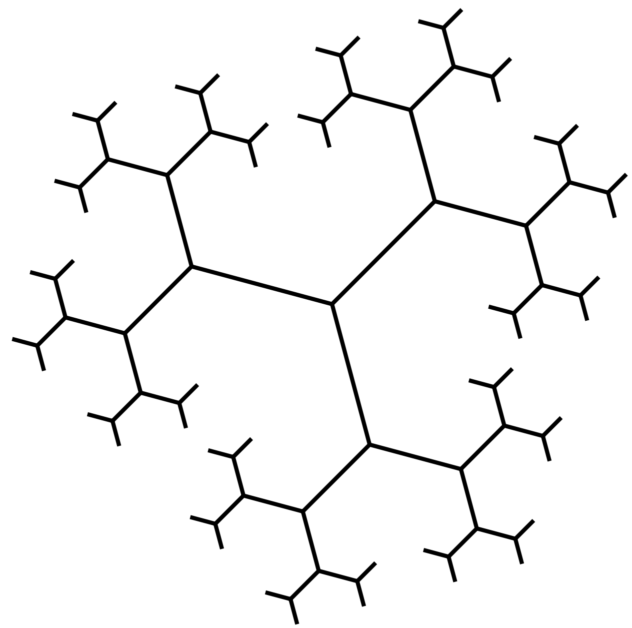


Alphabet $\mathcal{A}$:

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



k-regular-trees



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

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

Combinatorial view

$f : V \mapsto \mathbb{R}$ function on graph, or vector $(f_1, f_2, \dots, f_n)^\top$

$$f^\top L f = -\sum_{i \sim j} (f_i - f_j)^2$$

total weight of “edge violations”

Spectral view

Eigenvalues $0 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and corresponding eigenvectos $v_0, v_1, v_2, \ldots, v_n$.

$v_0 =$ const
v_1: “Fiedler-vector” smoothest function on graph orthogonal to v_0
In general v_i minimizes

$$\frac{v^\top L\, v}{v^\top v} \quad \text{for} \quad v \perp \text{span}(v_1, v_2, \ldots, v_{i-1})$$

$$L = \sum_i v_i\, \lambda_i\, v_i^\top \qquad\qquad K = e^{\beta L} = \sum_i v_i\, e^{\beta\, \lambda_i}\, v_i^\top$$

Regularization operator

Recall $\overline{K} = \left(\overline{P}\right)^{-2}$.

Now simply $\langle f, f \rangle = (Pf)^\top (Pf) = f^\top P^2 f$ and $P = K^{-1/2}$.

$$K = e^{\beta L} = \sum_i v_i \, e^{\beta \, \lambda_i} \, v_i^\top \qquad P = e^{-\beta L/2} = \sum_i v_i \, e^{-\beta \, \lambda_i / 2} \, v_i^\top$$

Generalization

[Smola & Kondor COLT 2003]

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

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

Applications

- Natural graphs: internet, web, social contacts, citations, scienti£c collaborations, etc.
- Objects with graph-like structure: strings, etc.
- Objects with unknown global structure: set of organic molecules,
- Bioinformatics: network of molecular pahways in cells [J-P Vert and Kanehisa NIPS 2002]
- Incorporating unlabelled data

2. Groups

Finite Groups

Finite set G with operation $G \times G \mapsto G$

- $x_1 x_2 \in G$ for any $x_1, x_2 \in G$ (closure)
- $x_1(x_2 x_3) = (x_1 x_2)x_3$ (associativity)
- $xe = ex = e$ for any $x \in G$ (identity)
- $x^{-1}x = xx^{-1} = e$ (inverses)

Symmetric groups S_n

$x_1 = (12)(3)(4)(5)$ $\quad$ x_1(ABCDE) = BACDE

$x_2 = (1)(324)(5)$ $\quad$ x_2(ABCDE) = ACDBE

$x_3 = x_1 x_2$ $\quad$ x_3(ABCDE) = x_1(x_2(ABCDE))

rankings, orderings, allocation, etc.
natural sense of distance

Stationary kernels

$$K(x_1, x_2) = f(x_2\, x_1^{-1}) \quad \sim \quad K(x_1, x_2) = f(x_2 - x_1)$$

eg. $K \sim e^{-(x_1 - x_2)^2/2\sigma^2}$

f positive de£nite function on G

Bochner's Theorem

f is positive de£nite and symmetrical on $\mathbb{R}^d$ iff $\widehat{f}(\boldsymbol{\omega}) > 0$

$$\widehat{f}(\boldsymbol{\omega}) = \frac{1}{\sqrt{2\pi}^{\,d}} \int e^{-i\,\boldsymbol{\omega}\cdot\boldsymbol{x}} f(\boldsymbol{x})\, d\boldsymbol{x}$$

is there an analog for £nite groups?

Representation theory

$\rho : G \to \mathbb{C}^{\,d \times d}$

$$\rho(x_1 x_2) = \rho(x_1)\rho(x_2)$$

Equivalence:

$$\rho_1(x) = t^{-1}\rho_2(x)\,t \qquad \forall\, x \in G$$

Reducibility:

$$t^{-1}\rho(x)\,t = \left(\begin{array}{c|c} \rho_1(x) & 0 \\ \hline 0 & \rho_2(x) \end{array}\right) \qquad \forall\, x \in G$$

Irreducible representations of S_5

$$\rho_{\text{trivial}}(x) \equiv (1) \qquad \longrightarrow \qquad \rho^{(5)}$$

$$\rho_{\text{sign}}(x) \equiv (\text{sgn}(x)) \qquad \longrightarrow \qquad \rho^{(1,1,1,1,1)}$$

$$\rho_{\text{def.}}(x) \in \mathbb{C}^{5\times 5} \qquad \begin{pmatrix} \boldsymbol{e}_{x(1)} \\ \boldsymbol{e}_{x(2)} \\ \boldsymbol{e}_{x(3)} \\ \boldsymbol{e}_{x(4)} \\ \boldsymbol{e}_{x(5)} \end{pmatrix} = \rho_{\text{def.}}(x) \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \boldsymbol{e}_3 \\ \boldsymbol{e}_4 \\ \boldsymbol{e}_5 \end{pmatrix}$$

Irreducible representations of S_5

$$t^{-1}\,\rho_{\mathrm{def.}}(x)\,t = \left(\begin{array}{c|c} \rho_{\mathrm{tr.}}(x) & \\ \hline & \\ & \rho^{(4,1)}(x) \\ & \end{array}\right)$$

$$\rho_{\mathrm{reg.}} = \rho^{(5)} \oplus 4\rho^{(4,1)} \oplus 5\rho^{(3,2)} \oplus 6\rho^{(3,1,1)} \oplus$$
$$5\rho^{(2,2,1)} \oplus 4\rho^{(2,1,1,1)} \oplus \rho^{(1,1,1,1,1)}$$

Fourier transforms on Groups

$$\widehat{f}(\rho) = \sum_{x \in G} \rho(x)\, f(x) \qquad\qquad \rho \in \mathcal{R}$$

$$\mathcal{F} :\ \mathbb{C}G \mapsto \bigoplus_{\rho \in \mathcal{R}} \mathbb{C}^{d_\rho \times d_\rho}$$

inversion:

$$f(x) = \frac{1}{|\,G\,|} \sum_{\rho \in \mathcal{R}} d_\rho \,\mathrm{trace}[\widehat{f}(\rho)\rho(x^{-1})]$$

Bochner on Groups

The function f is positive de£nite on G if and only if the matrices $\widehat{f}(\rho)$ are all positive de£nite.

Conjugacy classes

conjugacy classes: $x_1 \cong_{\text{Conj.}} x_2$ iff $x_1 = t^{-1}\, x_2\, t$ for some t

eg. on S_n $(\cdot)(\cdot)(\cdot)(\cdot)\ldots$

$(\cdot\;\cdot)(\cdot)(\cdot)(\cdot)\ldots$

$(\cdot\;\;\cdot\;\;\cdot)(\cdot)(\cdot)\ldots$

$(\cdot\;\cdot)(\cdot\;\cdot)(\cdot)(\cdot)\ldots$

$\vdots$

Corollary to Bochner

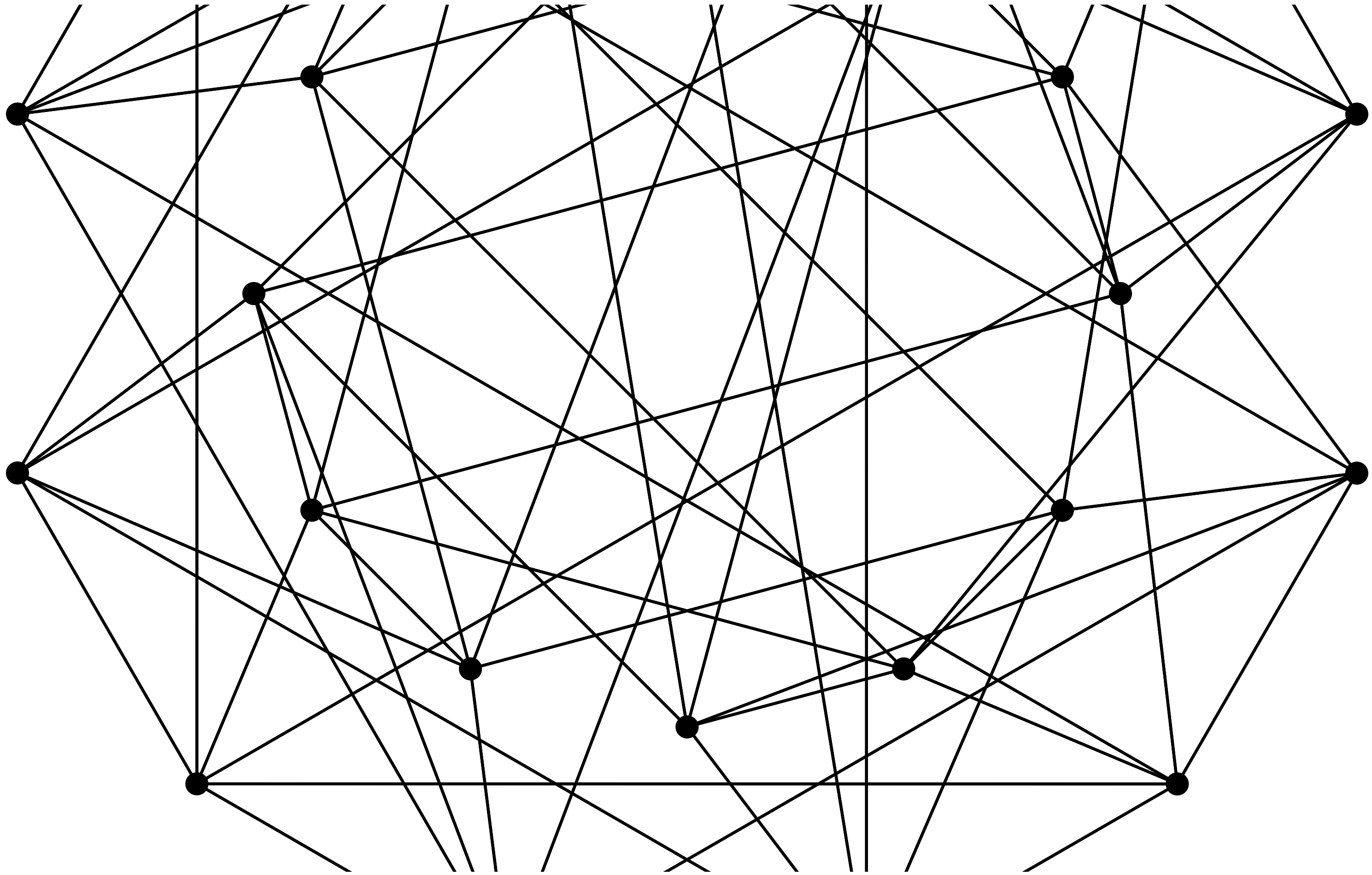
The function f is positive de£nite on G and constant on conjugacy classes if and only if

$$f(x) = \sum_{\rho \in \mathcal{R}} c_\rho \, \chi_\rho \qquad\qquad c_\rho > 0 \; .$$

characters:

$$\chi_\rho(x) = \mathrm{trace}[\rho(x)]$$

Diffusion on Cayley graph of S_4



3. Manifolds

Riemannian Manifolds

Surface $\mathcal{M}$ with sense of distance that locally looks like $\mathbb{R}^d$

In neigborhood of x, points can be represented as vector $\boldsymbol{x} + \boldsymbol{\delta x}$,

$$d(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\delta x})^2 = (\boldsymbol{\delta x})^\top G(x)\, \boldsymbol{\delta x} + O(\| \boldsymbol{\delta x} \|^4)$$

Metric tensor: $G(x) \in \mathbb{R}^{d \times d}$ positive de£nite for all $x \in \mathcal{M}$

For path $p : [0,1] \mapsto \mathbb{R}^d$,

$$\ell(p) = \int_0^1 \left(\sum_{i=1}^{d} \sum_{j=1}^{d} \partial_i p(\gamma)\, G(p(\gamma))_{ij}\, \partial_j p(\gamma) \right)^{1/2} d\gamma$$

Laplacian on a Riemannian Manifold

Flat space: $\Delta = \partial_1^2 + \partial_2^2 + \ldots + \partial_d^2$

Manifold:

$$\Delta = \frac{1}{\sqrt{\det G}} \sum_{ij} \partial_i \sqrt{\det G}\,(G^{-1})_{ij}\,\partial_j$$

gives rise to operator $\overline{\Delta} : \mathcal{L}_2(\mathcal{M}) \mapsto \mathcal{L}_2(\mathcal{M})$ as before

Diffusion Kernel on $\mathcal{M}$

Solution of

$$\partial_t \overline{K}_t = \overline{\Delta K}_t \quad \text{with} \quad K_0 = I$$

1. $K_t(x, x') = K_t(x', x)$
2. $\lim_{t \to 0} K_t(x, x') = \delta_x(x')$
3. $\left(\Delta - \frac{\partial}{\partial t}\right) K = 0$
4. $K_t(x, x') = \int_M K_{t-s}(x, x'') K_s(x'', x') \, dz$
5. $K_t(x, x') = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(x) \, \phi_i(x')$

Manifold Structures in Data

- Even when data at £rst seems very high dimensional, it is often constrained to a low dimensional manifold i.e. it only has a few internal degrees of freedom
- Constraining the kernel to the manifold is expected to help
- The graph Laplacian sampled from $\mathcal{M}$ approximates the Laplacian of $\mathcal{M}$ [Belkin & Niyogi NIPS 2001]
- A natural use for unlabeled data points is to help construct the kernel [Belkin & Niyogi NIPS 2002]

The Statistical Manifold

[Lafferty & Lebanon NIPS 2002]

For a family $\{\, p(x|\theta) \; : \; \theta \in \mathbb{R}^d \,\}$ of statistical models the Fisher metric is

$$G_{ij}(\theta) = \mathsf{E}(\partial_i \ell_\theta \; \partial_j \ell_\theta) = \int (\partial_i \log p(x|\theta)) \; (\partial_j \log p(x|\theta)) \; p(x|\theta) \, dx$$

or equivalently

$$G_{ij} = 4 \int \left(\partial_i \sqrt{p(x|\theta)} \right) \left(\partial_j \sqrt{p(x|\theta)} \right) \, dx \, .$$

Locally approximated by Kullback-Leibler divergence

The Multinomial

$$p(x|\theta) = \frac{(n+1)!}{x_1!x_2!\dots x_{n+1}!}\,\theta_1^{x_1}\theta_2^{x_2}\dots\theta_{n+1}^{x_{n+1}} \qquad \sum_{i=1}^{n+1}\theta_i = 1$$

$$G_{ij}(\theta) = \sum_{k=1}^{n+1}\frac{1}{\theta_i}(\partial_i\theta_k)\,(\partial_j\theta_k) \qquad \theta \in \mathcal{P}_d$$

Consider the map $T : \mathcal{P}_d \mapsto \mathcal{S}_d$ via $T : \theta \mapsto (\sqrt{\theta_1}, \sqrt{\theta_2}, \dots, \sqrt{\theta_{n+1}})$ on $\mathcal{S}_d$ the metric becomes the natural metric of the sphere, hence

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

Diffusion on the Sphere

$$K_t(x, x') = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{d^2(x, x')}{4t}\right) \sum_{i=0}^{N} \psi_i(x, x')\, t^i + O(t^N)$$

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

Proposed and applied to text data in [Lafferty & Lebanon NIPS 2002]

Conclusions

“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.”
(Nelson 1968)

Conclusions

- Kernel methods: algorithm + kernel
- Kernel alone encapsulates all knowledge about $\mathcal{X}$
- Laplacian is a unifying concept in Mathematics
- Connection with diffusion intuitively appealing, but real justi£cation for exponential kernels lies deep in operator-land
- Exponentially induced kernels lift knowledge of local structure to global level
- Opens new links to Abstract Algebra and Information Geometry

References

Belkin & Niyogi Laplacian Eigenmaps for Dimensionality Reduction (NIPS 2001)

Kondor & Lafferty Diffusion Kernels on Graphs and Other Discrete Structures (ICML 2002)

Lafferty & Lebanon Information Diffusion Kernels (NIPS 2002)

Belkin & Niyogi Using Manifold Structure for Partially Labelled Classi£cation (NIPS 2002)

Vert & Kanehisa Graph-driven Feature Abstraction from Microarray Data Using Diffusion Kernels and Kernel CCA (NIPS 2002)

Smola & Kondor Kernels and Regularization on Graphs (COLT 2003)