Statistical learning on graphs and groups through embeddings in Hilbert spaces

Jean-Philippe Vert
Jean-Philippe.Vert@ensmp.fr

Centre for Computational Biology
Ecole des Mines de Paris, ParisTech

International conference on Embeddings of Graphs and Groups into Hilbert and Banach spaces with applications, Centre interfacultaire Bernoulli, EPFL, Lausanne, Jan. 22-26, 2007.

Outline

Outline

Outline

1. Positive definite kernels and statistical learning
2. Kernels on groups and semigroups
3. Kernels on graphs
4. Conclusion

Outline

1. Positive definite kernels and statistical learning
2. Kernels on groups and semigroups
3. Kernels on graphs
4. Conclusion

Outline

1. Positive definite kernels and statistical learning
2. Kernels on groups and semigroups
3. Kernels on graphs
4. Conclusion

Positive Definite (p.d.) Kernels

Definition

A positive definite (p.d.) kernel on the set $\mathcal{X}$ is a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric:

$$\forall \left(\mathbf{x}, \mathbf{x}'\right) \in \mathcal{X}^2, \quad K\left(\mathbf{x}, \mathbf{x}'\right) = K\left(\mathbf{x}', \mathbf{x}\right),$$

and which satisfies, for all $N \in \mathbb{N}$, $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \in \mathcal{X}^N$ et $(a_1, a_2, \dots, a_N) \in \mathbb{R}^N$:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq 0.$$

P.d. kernels are inner products

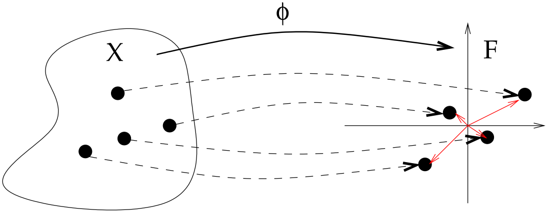
Theorem (Aronszajn, 1950)

K is a p.d. kernel on the set $\mathcal{X}$ if and only if there exists a Hilbert space $\mathcal{H}$ and a mapping

$$\Phi : \mathcal{X} \mapsto \mathcal{H} \,,$$

such that, for any $\mathbf{x}, \mathbf{x}'$ in $\mathcal{X}$:

$$K\left(\mathbf{x}, \mathbf{x}'\right) = \left\langle \Phi\left(\mathbf{x}\right), \Phi\left(\mathbf{x}'\right) \right\rangle_{\mathcal{H}} \,.$$



Reproducing kernel Hilbert space

Definition

Let $\mathcal{X}$ be a set and $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be a class of functions forming a (real) Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}}$. The function $K : \mathcal{X}^2 \mapsto \mathbb{R}$ is called a reproducing kernel (r.k.) of $\mathcal{H}$ if

1. $\mathcal{H}$ contains all functions of the form
$$\forall \mathbf{x} \in \mathcal{X}, \quad K_{\mathbf{x}} : \mathbf{t} \mapsto K(\mathbf{x}, \mathbf{t}) \; .$$
2. For every $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{H}$ the reproducing property holds:
$$f(\mathbf{x}) = \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}} \; .$$

If a r.k. exists, then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS).

Equivalence between positive definite and reproducing kernels

Theorem (Aronszajn, 1950)

K is a p.d. kernel if and only if there exists a RKHS having K as r.k.

Explicit construction of the RKHS

- If K is p.d., then the RKHS $\mathcal{H}$ is the vector subspace of $\mathbb{R}^{\mathcal{X}}$ spanned by the functions $\{K_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ (and their pointwise limits).
- For any $f, g \in \mathcal{H}_0$, given by:

$$f = \sum_i a_i K_{\mathbf{x}_i}, \quad g = \sum_j b_j K_{\mathbf{y}_j},$$

the inner product is given by:

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i,j} a_i b_j K\left(\mathbf{x}_i, \mathbf{y}_j\right).$$

Equivalence between positive definite and reproducing kernels

Theorem (Aronszajn, 1950)

K is a p.d. kernel if and only if there exists a RKHS having K as r.k.

Explicit construction of the RKHS

- If K is p.d., then the RKHS $\mathcal{H}$ is the vector subspace of $\mathbb{R}^{\mathcal{X}}$ spanned by the functions $\{K_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ (and their pointwise limits).
- For any $f, g \in \mathcal{H}_0$, given by:

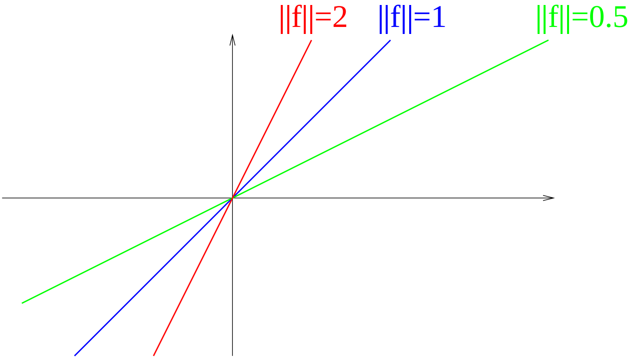
$$f = \sum_i a_i K_{\mathbf{x}_i}, \quad g = \sum_j b_j K_{\mathbf{y}_j},$$

the inner product is given by:

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i,j} a_i b_j K\left(\mathbf{x}_i, \mathbf{y}_j\right).$$

Example : RKHS of the linear kernel (cont.)

$$\begin{cases} K(\mathbf{x}, \mathbf{x}') & = \mathbf{x}^\top \mathbf{x}' \, . \\ f(\mathbf{x}) & = w^\top \mathbf{x} \, , \\ \| f \|_{\mathcal{H}} & = \| w \|_2 \, . \end{cases}$$



Smoothness functional

A simple inequality

- By Cauchy-Schwarz we have, for any function $f \in \mathcal{H}$ and any two points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$:

$$\begin{aligned} \left| f(\mathbf{x}) - f(\mathbf{x}') \right| &= \left| \langle f, K_{\mathbf{x}} - K_{\mathbf{x}'} \rangle_{\mathcal{H}} \right| \\ &\leq \| f \|_{\mathcal{H}} \times \| K_{\mathbf{x}} - K_{\mathbf{x}'} \|_{\mathcal{H}} \\ &= \| f \|_{\mathcal{H}} \times d_K(\mathbf{x}, \mathbf{x}') \ . \end{aligned}$$

- The norm of a function in the RKHS controls how fast the function varies over $\mathcal{X}$ with respect to the geometry defined by the kernel (Lipschitz with constant $\| f \|_{\mathcal{H}}$).

Important message

Small norm $\implies$ slow variations.

The representer theorem

Theorem (Kimeldorf and Wahba, 1971)

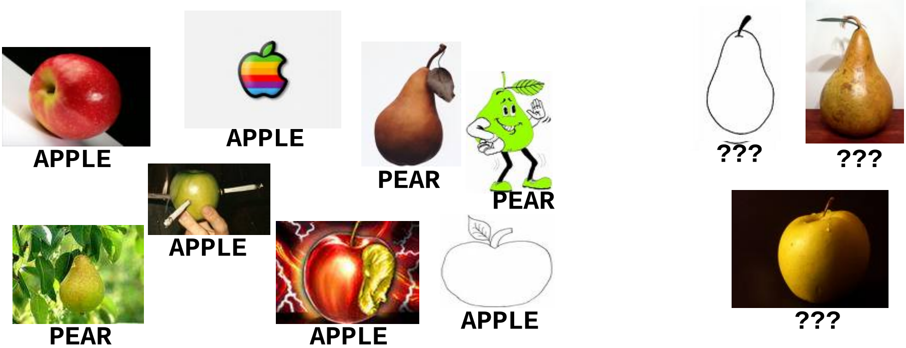
- Let $\mathcal{X}$ be a set endowed with a p.d. kernel K, $\mathcal{H}_K$ the corresponding RKHS, and $\mathcal{S} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subset \mathcal{X}$ a finite set of points in $\mathcal{X}$.
- Let $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ be a function of $n+1$ variables, strictly increasing with respect to the last variable.
- Then, any solution to the optimization problem:

$$\min_{f \in \mathcal{H}_K} \Psi\left(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n), \| f \|_{\mathcal{H}_K}\right),$$

admits a representation of the form:

$$\forall \mathbf{x} \in \mathcal{X}, \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}).$$

Pattern recognition



- Input variables $\mathbf{x} \in \mathcal{X}$
- Output $y \in \{-1, 1\}$.
- Training set $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

Empirical risk minimization (ERM)

ERM estimator

- Loss function $l(f(\mathbf{x}), \mathbf{y}) \in \mathbb{R}$ small when $f(\mathbf{x})$ is a good predictor for y
- Empirical risk:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), Y_i) .$$

- The ERM estimator on the functional class $\mathcal{F}$ is the solution of:

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} R_n(f) .$$

- Statistical learning theory : the estimator is consistent when the "complexity" of the class $\mathcal{F}$ is controlled

ERM in RKHS balls

Principle

- Suppose $\mathcal{X}$ is endowed with a p.d. kernel
- We consider the ball of radius B in the RKHS as function class for the ERM:

$$\mathcal{F}_B = \{ f \in \mathcal{H} : \| f \|_{\mathcal{H}} \leq B \} .$$

- Theoretical justifications exist (upper bounds on the "complexity" of $\mathcal{F}_B$).

ERM in practice

Reformulation as penalized minimization

- We must solve the constrained minimization problem:

$$\begin{cases} \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l\left(f\left(\mathbf{x}_i\right), \mathbf{y}_i\right) \\ \text{subject to } \| f \|_{\mathcal{H}} \leq B\,. \end{cases}$$

- To make this practical we assume that l is a convex function of f.
- The problem is then a convex problem in f for which strong duality holds. In particular f solves the problem if and only if it solves for some dual parameter λ the unconstrained problem:

$$\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} l\left(f\left(\mathbf{x}_i\right), \mathbf{y}_i\right) + \lambda \| f \|_{\mathcal{H}}^2 \right\},$$

and complimentary slackness holds ($\lambda = 0$ or $\| f \|_{\mathcal{H}} = B$).

Optimization in RKHS

- By the representer theorem, the solution of the unconstrained problem can be expanded as:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) .$$

- Plugging into the original problem we obtain the following unconstrained and convex optimization problem in $\mathbb{R}^n$:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} l\left(\sum_{j=1}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \mathbf{y}_i \right) + \lambda \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \right\} .$$

- This can be implemented using general packages for convex optimization or specific algorithms (e.g., SVM).

Example : support vector machines

The classifier is:

$$\forall \mathbf{x} \in \mathcal{X}, \quad f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \ ,$$

where α is the solution of the following QP:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^d} 2 \sum_{i=1}^{n} \alpha_i y_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \ ,$$

subject to:

$$0 \leq y_i \alpha_i \leq C, \quad \text{for } i = 1, \ldots, n \, .$$

Example

Summary

- 3 ways to map $\mathcal{X}$ to a Hilbert space
  1. Explicitly define and compute $\Phi : \mathcal{X} \to \mathcal{H}$
  2. Define a p.d. kernel over $\mathcal{X}$
  3. Define a RKHS over $\mathcal{X}$
- The p.d. kernel is sufficient for a variety of applications in data analysis and machine learning

Outline

1. Positive definite kernels and statistical learning
2. Kernels on groups and semigroups
3. Kernels on graphs
4. Conclusion

Semigroups

Definition

- A semigroup $(S, \circ)$ is a nonempty set S equipped with an associative composition $\circ$ and a neutral element e.
- A semigroup with involution $(S, \circ, *)$ is a semigroup $(S, \circ)$ together with a mapping $* : S \to S$ called involution satisfying:
  1. $(s \circ t)^* = t^* \circ s^*$, for $s, t \in S$.
  2. $(s^*)^* = s$ for $s \in S$.

Examples

- Any group $(G, \circ)$ is a semigroup with involution when we define $s^* = s^{-1}$.
- Any abelian semigroup $(S, +)$ is a semigroup with involution when we define $s^* = s$, the identical involution.

Positive definite functions on semigroups

Definition

Let $(S, \circ, *)$ be a semigroup with involution. A function $\phi : S \to \mathbb{R}$ is called positive definite if the function:

$$\forall s, t \in S, \quad K(s,t) = \phi\left(s^* \circ t\right)$$

is a p.d. kernel on S.

Example: translation invariant kernels

$\left(\mathbb{R}^d, +, -\right)$ is an abelian group with involution. A function $\phi : \mathbb{R}^d \to \mathbb{R}$ is p.d. if the function

$$K(x,y) = \phi(x-y)$$

is p.d. on $\mathbb{R}^d$ (translation invariant kernels).

Semicharacters

Definition

A funtion $\rho : S \to \mathbb{C}$ on an abelian semigroup with involution $(S, +, *)$ is called a semicharacter if

1. $\rho(0) = 1$,
2. $\rho(s + t) = \rho(s)\rho(t)$ for $s, t \in S$,
3. $\rho(s^*) = \overline{\rho(s)}$ for $s \in S$.

The set of semicharacters on S is denoted by S^*.

Integral representation of p.d. functions

Definition

- An function $\alpha : S \rightarrow \mathbb{R}$ on a semigroup with involution is called an absolute value if (i) $\alpha(e) = 1$, (ii)$\alpha(s \circ t) \leq \alpha(s)\alpha(t)$, and (iii) $\alpha\left(s^*\right) = \alpha(s)$.
- A function $f : S \rightarrow \mathbb{R}$ is called exponentially bounded if there exists an absolute value α and a constant $C > 0$ s.t. $\mid f(s) \mid \leq C\alpha(s)$ for $s \in S$.

Theorem

Let $(S, +, *)$ an abelian semigroup with involution. A function $\phi : S \rightarrow \mathbb{R}$ is p.d. and exponentially bounded (resp. bounded) if and only if it has a representation of the form:

$$\phi(s) = \int_{S^*} \rho(s) d\mu(\rho) \,.$$

where μ is a Radon measure with compact support on S^* (resp. on $\hat{S}$, the set of bounded semicharacters).

Integral representation of p.d. functions

Definition

- An function $\alpha : S \to \mathbb{R}$ on a semigroup with involution is called an absolute value if (i) $\alpha(e) = 1$, (ii)$\alpha(s \circ t) \leq \alpha(s)\alpha(t)$, and (iii) $\alpha\left(s^*\right) = \alpha(s)$.
- A function $f : S \to \mathbb{R}$ is called exponentially bounded if there exists an absolute value α and a constant $C > 0$ s.t. $\mid f(s) \mid \leq C\alpha(s)$ for $s \in S$.

Theorem

Let $(S, +, *)$ an abelian semigroup with involution. A function $\phi : S \to \mathbb{R}$ is p.d. and exponentially bounded (resp. bounded) if and only if it has a representation of the form:

$$\phi(s) = \int_{S^*} \rho(s) d\mu(\rho) .$$

where μ is a Radon measure with compact support on S^* (resp. on $\hat{S}$, the set of bounded semicharacters).

Example 1: $(R_+, +, Id)$

P.d. functions

- $S = (\mathbb{R}_+, +, Id)$ is an abelian semigroup.
- The set of bounded semicharacters is exactly the set of functions:

$$s \in \mathbb{R}_+ \mapsto \rho_a(s) = e^{-as},$$

  for $a \in [0, +\infty]$
- A function $\phi : \mathbb{R}_+ \to \mathbb{R}$ is p.d. and bounded if and only if it has the form:

$$\phi(s) = \int_0^{\infty} e^{-as} d\mu(a) + b\rho_\infty(s)$$

  where $\mu \in \mathcal{M}_+^b(\mathbb{R}_+)$ and $b \geq 0$.
- ϕ is p.d., bounded and continuous iff it is the Laplace transform of a measure in $\mathcal{M}_+^b(\mathbb{R})$.

Example 2: characterization of p.d. t.i. kernels

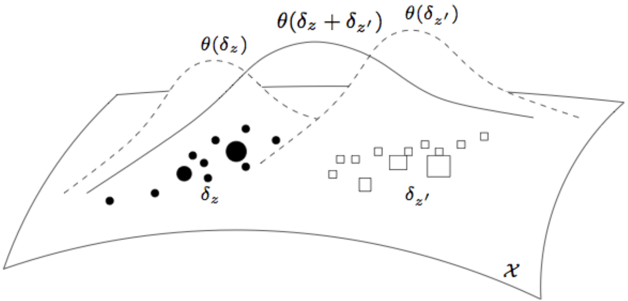
Theorem (Bochner)

A function $\kappa(x-y)$ on $\mathbb{R}^d$ is positive definite if and only if it is the Fourier transform of a function $\hat{\kappa}(\omega)$ symmetric, positive, and tending to 0 at infinity.

Examples

$$
\begin{aligned}
K_{Gaussian}(x,y) &= e^{-\frac{(x-y)^2}{2\sigma^2}}\,, \\
K_{Laplace}(x,y) &= \frac{1}{2}e^{-\gamma|x-y|}\,, \\
K_{Filter}(x,y) &= \frac{\sin\left(\Omega(x-y)\right)}{\pi(x-y)}\,.
\end{aligned}
$$

Example 3: Semigroup kernels for finite measures (1/6)



- We assume that data to be processed are “bags-of-points”, i.e., sets of points (with repeats) of a space $\mathcal{U}$.
- Example : a finite-length string as a set of k-mers.
- How to define a p.d. kernel between any two bags that only depends on the union of the bags?

Example 3: Semigroup kernels for finite measures (2/6)

Semigroup of bounded measures

- We can represent any bag-of-point $\mathbf{x}$ as a finite measure on $\mathcal{U}$:

$$\mathbf{x} = \sum_i a_i \mu_{x_i} \, ,$$

where a_i is the number of occurrences on $\mathbf{x}_i$ in the bag and μ_x is a basic measure centered on x.
- The measure that represents the union of two bags is the sum of the measures that represent each individual bag.
- This suggests to look at the semigroup $\left(\mathcal{M}_+^b\left(\mathcal{U}\right), +, Id\right)$ of bounded Radon measures on $\mathcal{U}$ and to search for p.d. functions ϕ on this semigroup.

Example 3: Semigroup kernels for finite measures (3/6)

Semicharacters

- For any Borel measurable function $f : \mathcal{U} \to \mathbb{R}$ the function $\rho_f : \mathcal{M}_+^b(\mathcal{U}) \to \mathbb{R}$ defined by:

$$\rho_f(\mu) = e^{\mu[f]}$$

  is a semicharacter on $\left(\mathcal{M}_+^b(\mathcal{U}), +\right)$.
- Conversely, ρ is continuous semicharacter (for the topology of weak convergence) if and only if there exists a continuous function $f : \mathcal{U} \to \mathbb{R}$ such that $\rho = \rho_f$.
- No such characterization for non-continuous characters, even bounded.

Example 3: Semigroup kernels for finite measures (4/6)

Corollary

Let $\mathcal{U}$ be a Hausdorff space. For any Radon measure $\mu \in \mathcal{M}_+^c\left(C\left(\mathcal{U}\right)\right)$ with compact support on the Hausdorff space of continuous real-valued functions on $\mathcal{U}$ endowed with the topology of pointwise convergence, the following function K is a continuous p.d. kernel on $\mathcal{M}_+^b\left(\mathcal{U}\right)$ (endowed with the topology of weak convergence):

$$K(\mu, \nu) = \int_{C(\mathcal{X})} e^{\mu[f]+\nu[f]} d\mu(f)\,.$$

Remarks

The converse is not true: there exist continuous p.d. kernels that do not have this integral representation (it might include non-continuous semicharacters)

Example 3: Semigroup kernels for finite measures (5/6)

Example : entropy kernel

- Let $\mathcal{X}$ be the set of probability densities (w.r.t. some reference measure) on $\mathcal{U}$ with finite entropy:

$$h(\mathbf{x}) = -\int_{\mathcal{U}} \mathbf{x} \ln \mathbf{x} \,.$$

- Then the following entropy kernel is a p.d. kernel on $\mathcal{X}$ for all $\beta > 0$:

$$K\left(\mathbf{x}, \mathbf{x}'\right) = e^{-\beta h\left(\frac{\mathbf{x}+\mathbf{x}}{2}\right)} \,.$$

- Remark: only valid for densities (e.g., for a kernel density estimator from a bag-of-parts)

Example 3: Semigroup kernels for finite measures (6/6)

Examples : inverse generalized variance kernel

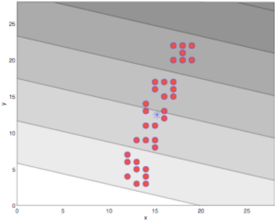
- Let $\mathcal{U} = \mathbb{R}^d$ and $\mathcal{M}_+^V(\mathcal{U})$ be the set of finite measure μ with second order moment and non-singular variance

$$\Sigma(\mu) = \mu\left[xx^\top\right] - \mu[x]\,\mu[x]^\top .$$

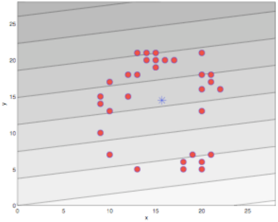
- Then the following function is a p.d. kernel on $\mathcal{M}_+^V(\mathcal{U})$, called the inverse generalized variance kernel:

$$K(\mu, \mu') = \frac{1}{\det \Sigma\left(\frac{\mu+\mu'}{2}\right)} .$$

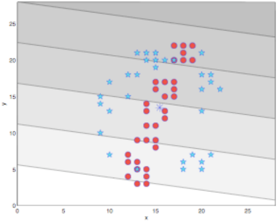
Application of semigroup kernel



$\Sigma_{1,1} = 0.0552$
$\Sigma_{2,2} = 0.0013$



$\Sigma'_{1,1} = 0.0441$
$\Sigma'_{2,2} = 0.0237$



$\Sigma''_{1,1} = 0.0497$
$\Sigma''_{2,2} = 0.0139$

Weighted linear PCA of two different measures, with the first PC shown. Variances captured by the first and second PC are shown. The generalized variance kernel is the inverse of the product of the two values.

Kernelization of the IGV kernel

Motivations

- Gaussian distributions may be poor models.
- The method fails in large dimension

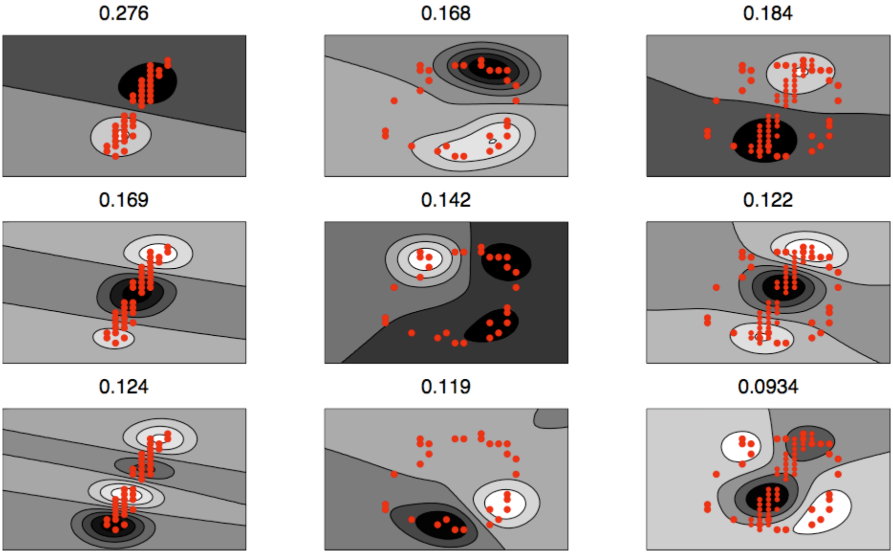
Solution

1. Regularization:

$$K_\lambda\left(\mu,\mu'\right) = \frac{1}{\det\left(\Sigma\left(\frac{\mu+\mu'}{2}\right) + \lambda I_d\right)}.$$

2. Kernel trick: the non-zero eigenvalues of $UU^\top$ and $U^\top U$ are the same $\Longrightarrow$ replace the covariance matrix by the centered Gram matrix (technical details in Cuturi et al., 2005).

Illustration of kernel IGV kernel



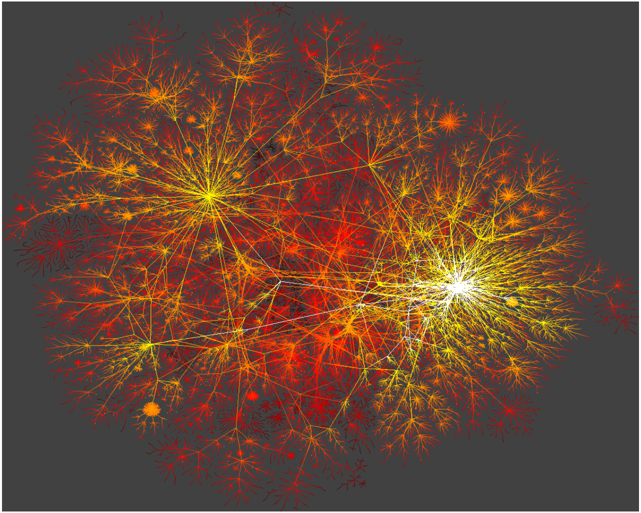
Semigroup kernel remarks

Motivations

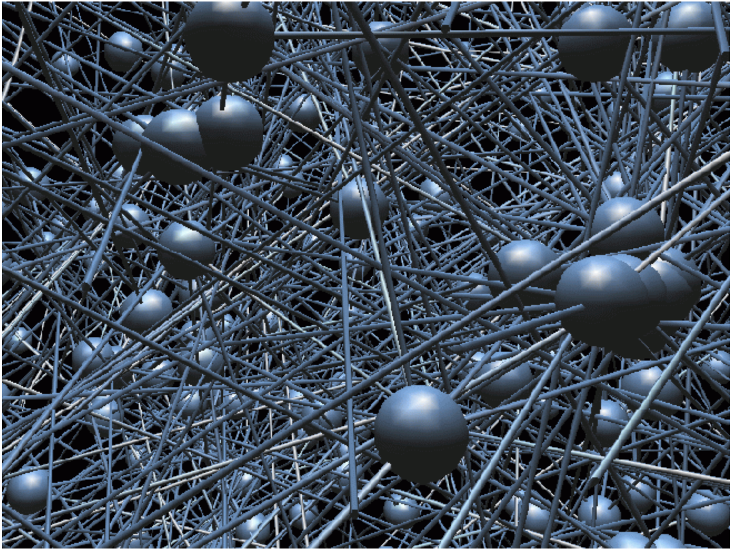
- A very general formalism to exploit an algebric structure of the data.
- Kernel IVG kernel has given good results for character recognition from a subsampled image.
- The main motivation is more generally to develop kernels for complex objects from which simple "patches" can be extracted.
- The extension to nonabelian groups (e.g., permutation in the symmetric group) might find natural applications.

Outline

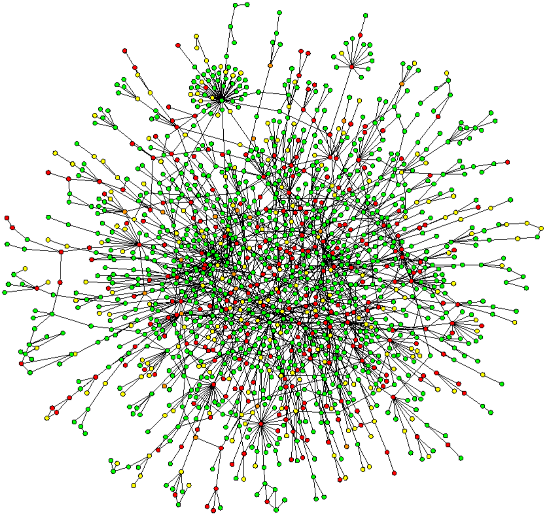
Example: web



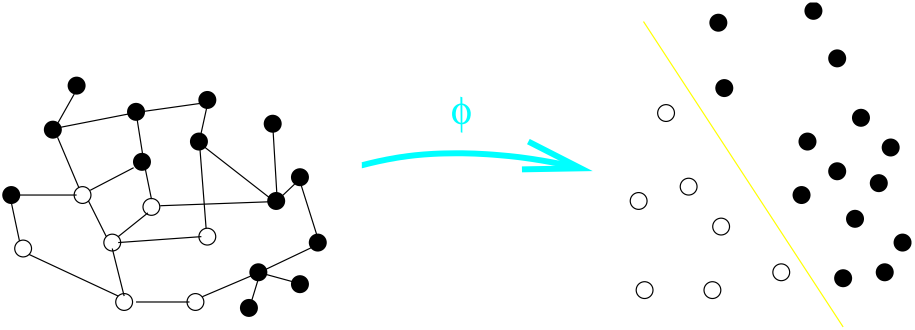
Example: social network



Example: protein-protein interaction



Kernel on a graph



- We need a kernel $K(\mathbf{x}, \mathbf{x}')$ between nodes of the graph.
- Example: predict gene protein functions from high-throughput protein-protein interaction data.

General remarks

Strategies to make a kernel on a graph

- $\mathcal{X}$ being finite, any symmetric semi-definite matrix K defines a valid p.d. kernel on $\mathcal{X}$.
- How to "translate" the graph topology into the kernel?
  - Direct geometric approach: $K_{i,j}$ should be "large" when $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" to each other on the graph?
  - Functional approach: $\| f \|_K$ should be "small" when f is "smooth" on the graph?
  - Link discrete/continuous: is there an equivalent to the continuous Gaussien kernel on the graph (e.g., limit by fine discretization)?

General remarks

Strategies to make a kernel on a graph

- $\mathcal{X}$ being finite, any symmetric semi-definite matrix K defines a valid p.d. kernel on $\mathcal{X}$.
- How to "translate" the graph topology into the kernel?
  - Direct geometric approach: $K_{i,j}$ should be "large" when $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close" to each other on the graph?
  - Functional approach: $\| f \|_K$ should be "small" when f is "smooth" on the graph?
  - Link discrete/continuous: is there an equivalent to the continuous Gaussien kernel on the graph (e.g., limit by fine discretization)?

First approach : Geometric

A direct approach

- Remember : for $\mathcal{X} = \mathbb{R}^n$, the Gaussian RBF kernel is:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-d(\mathbf{x}, \mathbf{x}')^2 / 2\sigma^2\right),$$

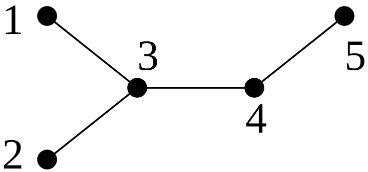
  where $d(\mathbf{x}, \mathbf{x}')$ is the Euclidean distance.
- If $\mathcal{X}$ is a graph, let $d(\mathbf{x}, \mathbf{x}')$ be the shortest-path distance between $\mathbf{x}$ and $\mathbf{x}'$.
- Problem: the shortest-path distance is not a Hilbert distance...

Second approach : Functional

Idea

- Define a priori a smoothness functional on the functions $f : \mathcal{X} \rightarrow \mathbb{R}$.
- Show that it defines a RKHS and identify the corresponding kernel

Notations

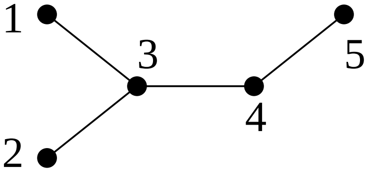


$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Graph Laplacian

Definition

The Laplacian of the graph is the matrix $L = A - D$.



$$L = A - D = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 1 & 1 & -3 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

Properties of the Laplacian

Lemma

Let $L = A - D$ be the Laplacian of the graph:

- *For any $f : \mathcal{X} \to \mathbb{R}$,*

$$\Omega(f) := \sum_{i \sim j} \left(f(\mathbf{x}_i) - f(\mathbf{x}_j) \right)^2 = -f^\top L f$$

- *$-L$ is a symmetric positive semi-definite matrix*
- *0 is an eigenvalue with multiplicity 1 associated to the constant eigenvector $\mathbf{1} = (1, \ldots, 1)$*
- *The image of L is*

$$Im(L) = \left\{ f \in \mathbb{R}^m : \sum_{i=1}^{m} f_i = 0 \right\}$$

Our first graph kernel

Theorem

The set $\mathcal{H} = \left\{ f \in \mathbb{R}^m : \sum_{i=1}^m f_i = 0 \right\}$ *endowed with the norm:*

$$\Omega(f) = \sum_{i \sim j} \left(f(\mathbf{x}_i) - f(\mathbf{x}_j) \right)^2$$

is a RKHS whose reproducing kernel is $(-L)^*$*, the pseudo-inverse of the graph Laplacian.*

Third approach: The diffusion equation

Lemma

For any $\mathbf{x}_0 \in \mathbb{R}^d$*, the function:*

$$K_{\mathbf{x}_0}(\mathbf{x}, t) = K_t(\mathbf{x}_0, \mathbf{x}) = \frac{1}{(4\pi t)^{\frac{d}{2}}} \exp\left(-\frac{\| \mathbf{x} - \mathbf{x}_0 \|^2}{4t}\right).$$

is solution of the diffusion equation:

$$\frac{\partial}{\partial t} K_{\mathbf{x}_0}(\mathbf{x}, t) = \Delta K_{\mathbf{x}_0}(\mathbf{x}, t).$$

with initial condition $K_{\mathbf{x}_0}(\mathbf{x}, 0) = \delta_{\mathbf{x}_0}(\mathbf{x})$.

Discrete diffusion equation

- For finite-dimensional $f_t \in \mathbb{R}^m$, the diffusion equation becomes:

$$\frac{\partial}{\partial t} f_t = L f_t$$

which admits the following solution:

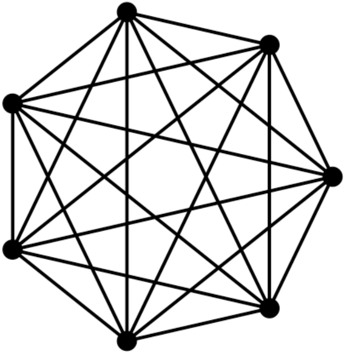
$$f_t = f_0 e^{tL}$$

- This suggest to consider:

$$K = e^{tL}$$

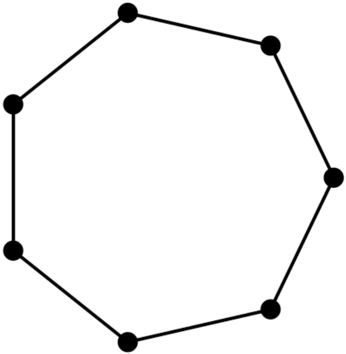
which is indeed symmetric positive semi-definite. We call it the diffusion kernel or heat kernel.

Example: complete graph



$$K_{i,j} = \begin{cases} \frac{1+(m-1)e^{-tm}}{m} & \text{for } i = j, \\ \frac{1-e^{-tm}}{m} & \text{for } i \neq j. \end{cases}$$

Example: closed chain



$$K_{i,j} = \frac{1}{m} \sum_{\nu=0}^{m-1} \exp\left[-2t\left(1 - \cos\frac{2\pi\nu}{m}\right)\right] \cos\frac{2\pi\nu(i-j)}{m}.$$

Spectrum of the diffusion kernel

- Let $0 = \lambda_1 > -\lambda_2 \geq \ldots \geq -\lambda_m$ be the eigenvalues of the Laplacian:

$$L = \sum_{i=1}^{m} (-\lambda_i) u_i u_i^\top \quad (\lambda_i \geq 0)$$

- The diffusion kernel K_t is an invertible matrix because its eigenvalues are strictly positive:

$$K_t = \sum_{i=1}^{m} e^{-t\lambda_i} u_i u_i^\top$$

Norm in the diffusion RKHS

- For any function $f \in \mathbb{R}^m$, let:

$$\hat{f}_i = u_i^\top f$$

be the Fourier coefficients of f (projection of f onto the eigenbasis of K).

- The RKHS norm of f is then:

$$\| f \|_{K_t}^2 = f^\top K^{-1} f = \sum_{i=1}^{m} e^{t\lambda_i} \hat{f}_i^2 .$$

Generalization

This observation suggests to define a whole family of kernels:

$$K_r = \sum_{i=1}^{m} r(\lambda_i) u_i u_i^\top$$

associated with the following RKHS norms:

$$\| f \|_{K_r}^2 = \sum_{i=1}^{m} \frac{\hat{f}_i^2}{r(\lambda_i)}$$

where $r : \mathbb{R}^+ \to \mathbb{R}_*^+$ is a non-increasing function.

Example : regularized Laplacian

$$r(\lambda) = \frac{1}{\lambda + \epsilon}, \qquad \epsilon > 0$$

$$K = \sum_{i=1}^{m} \frac{1}{\lambda_i + \epsilon} u_i u_i^\top = (-L + \epsilon I)^{-1}$$

$$\| f \|_K^2 = f^\top K^{-1} f = \sum_{i \sim j} \left(f(\mathbf{x}_i) - f(\mathbf{x}_j) \right)^2 + \epsilon \sum_{i=1}^{m} f(\mathbf{x}_i)^2 .$$

Applications 1: graph partitioning

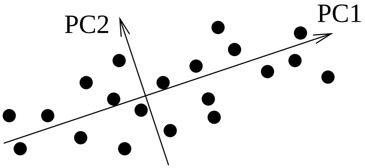
- A classical relaxation of graph partitioning is:

$$\min_{f \in \mathbb{R}^{\mathcal{X}}} \sum_{i \sim j} \left(f_i - f_j\right)^2 \quad \text{s.t.} \sum_i f_i^2 = 1$$

- This can be rewritten

$$\max_f \sum_i f_i^2 \text{ s.t.} \quad \| f \|_{\mathcal{H}} \leq 1$$

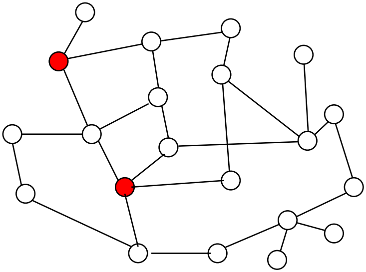
- This is principal component analysis in the RKHS ("kernel PCA")



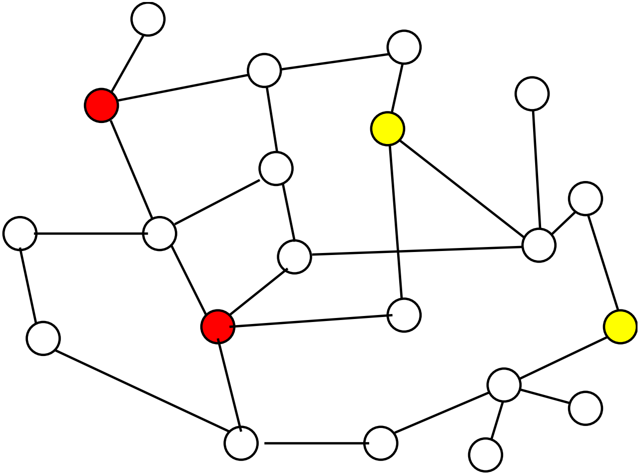
Applications 2: search on a graph

- Let $x_1, \ldots, x_q$ a set of q nodes (the query). How to find "similar" nodes (and rank them)?
- One solution:

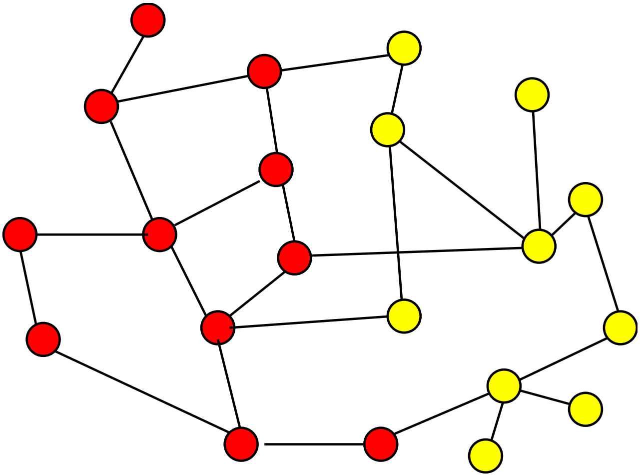
$$\min_f \| f \|_{\mathcal{H}} \quad \text{s.t.} \quad f(x_i) \geq 1 \text{ for } i = 1, \ldots, q.$$



Application 3: Semi-supervised learning



Application 3: Semi-supervised learning



Example 4: Tumor classification from microarray data

Data available

- Gene expression measures for more than 10k genes
- Measured on less than 100 samples of two (or more) different classes (e.g., different tumors)

Goal

- Design a classifier to automatically assign a class to future samples from their expression profile
- Interpret biologically the differences between the classes

Example 4: Tumor classification from microarray data

Data available

- Gene expression measures for more than 10k genes
- Measured on less than 100 samples of two (or more) different classes (e.g., different tumors)

Goal

- Design a classifier to automatically assign a class to future samples from their expression profile
- Interpret biologically the differences between the classes

Linear classifiers

The approach

- Each sample is represented by a vector $x = (x_1, \ldots, x_p)$ where $p > 10^5$ is the number of probes
- Classification: given the set of labeled sample, learn a linear decision function:

$$f(x) = \sum_{i=1}^{p} \beta_i x_i + \beta_0 \, ,$$

- Interpretation: the weight β_i quantifies the influence of gene i for the classification

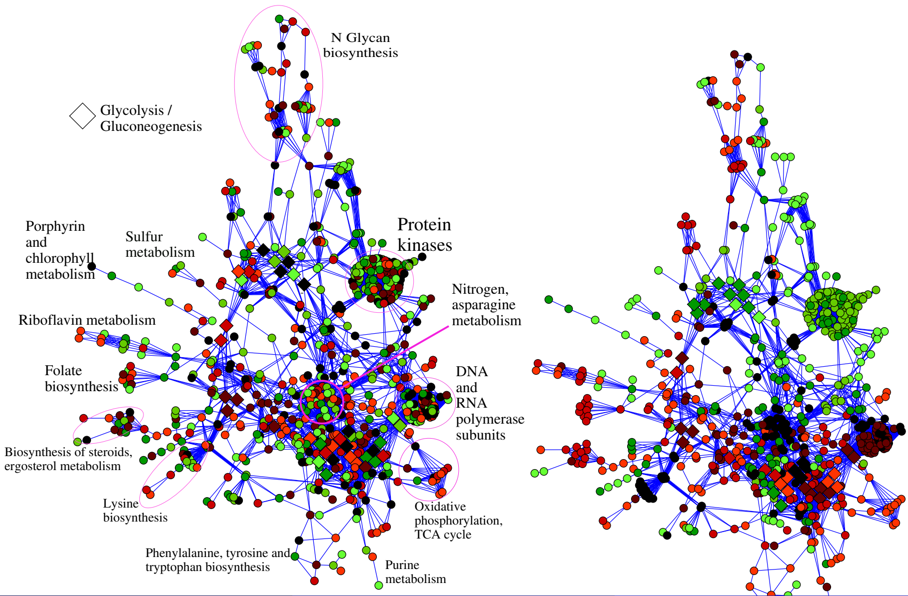
Pitfalls

- No robust estimation procedure exist for 100 samples in 10^5 dimensions!

Prior knowledge

- We know the functions of many genes, and how they interact together.
- This can be represented as a graph of genes, where connected genes perform some action together
- Prior knowledge: constraint the weights of genes that work together to be similar
- Mathematically: constrain the norm of the weight vector in the RKHS of the diffusion kernel.

Comparison



N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative phosphorylation, TCA cycle
Phenylalanine, tyrosine and tryptophan biosynthesis
Purine metabolism

Outline

Conclusion

- Implicit Hilbert space embedding through positive definite kernels
- State-of-the-art machine learning algorithms based on optimization in reproducing kernel Hilbert spaces
- P.d. kernels on groups and graphs allow the extension of these algorithms to non-vectorial data
- Making p.d. kernel for particular objects is a hot topic in machine learning!
- Many potential applications!

Further reading

Kernels and RKHS: general

N. Aronszajn.
Theory of reproducing kernels.
Trans. Am. Math. Soc., 68:337 – 404, 1950.

C. Berg, J. P. R. Christensen, and P. Ressel.
Harmonic analysis on semigroups.
Springer-Verlag, New-York, 1984.

G. Wahba.
Spline Models for Observational Data, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*.
SIAM, Philadelphia, 1990.

Further reading

Learning with kernels

V. N. Vapnik.
Statistical Learning Theory.
Wiley, New-York, 1998.

B. Schölkopf and A. J. Smola.
Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.
MIT Press, Cambridge, MA, 2002.

J. Shawe-Taylor and N. Cristianini.
Kernel Methods for Pattern Analysis.
Cambridge University Press, 2004.

B. Schölkopf, K. Tsuda, and J.-P. Vert.
Kernel Methods in Computational Biology.
MIT Press, 2004.

Further reading

Kernels on graphs

R. I. Kondor and J. Lafferty.
Diffusion Kernels on Graphs and Other Discrete Input.
In *ICML 2002*, 2002.

Semigroup kernels

C. Berg, J. P. R. Christensen, and P. Ressel.
Harmonic analysis on semigroups.
Springer-Verlag, New-York, 1984.

M. Cuturi, K. Fukumizu, and J.P. Vert.
Semigroup Kernels on Measures.
J. Mach. Learn. Res., 6:1169–1198, 2005.