# EPAT 2010

# Kernel Methods

## Definitions & Kernel Design

Marco Cuturi

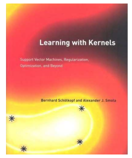
# Kernel Methods

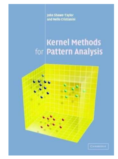
A reasonably large academic subfield

- Widespread popularity in machine learning now

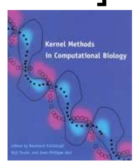

- Gained momentum in the late 90's with the support vector machine,
- Rooted in much older maths.
- Kernel methods are a pluridisciplinary field, publications appearing in
  - computer science (*nips, journ. of machine learning, ICML..*),
  - statistics and functional analysis (*annals of statistics..*),
  - optimization (*Mathematical Programming..*),
  - Different application subfields (*Neural Computation..*)

# Kernel Methods

- Standard text-books:
  - Introduction [SS02] 

  - More about kernels [STC04] 

  - More learning theory [SC08] 

  - First chapters [STV04] 

  - “Mathematical” perspective [BTA03]. The real deal: [BCR84].
- Some short surveys,
  - journal papers [HHS08], [MMR+01]
  - a survey on my webpage (local copy, not arxiv): key to all citations!
- On the web:
  - Courses by J.-P. Vert, Francis Bach, Kenji Fukumizu, Stéphane Canu.

# Some terminology

**Etymology** : from old english *cyrnel*, diminutive of corn (seed)

the word **kernel** appears in different different contexts...

- The *linux* kernel...
- Kernel of a linear operator of $\mathcal{X}$: $\ker(L) = \{x \in \mathcal{X} | L(x) = 0\}$.
- Kernel of a matrix in $\mathbb{R}^{d \times d}$, $i.e.$ its nullspace $\{\mathbf{x} \in \mathbb{R}^d | A\mathbf{x} = \mathbf{0}\}$.
- In set theory, for a function $f : \mathcal{X} \mapsto \mathcal{Y}$, $\ker(f) = \{(x, x') | \, f(x) = f(x')\}$.
- Kernel of an integral transform $T$, $Tf(u) = \int_{t_1}^{t_2} k(t, u) f(t) dt$
- Smoothing kernel, a function $k \geq 0, k(u) = k(-u), \int_{-\infty}^{\infty} k(u) du = 1$.
- $K(t, x, y) = \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|x-y\|^2}{4t}}$ solves heat equation $K(t, x, y) = \Delta_x K(t, x, y)$

sets, subspaces, **one-**variable, **two-**variables, **three-**variables function...

# Moral of the story

$\boxed{\textbf{No need} \text{ to look for a common or primitive meaning}}$

- Kernel is just a word mathematicians fancy (unfortunately!)
- People enjoy it because of its vague "core" meaning.

- Don't feel you have missed something if you do not see the connection between different *kernel* objects in mathematics. There might be none...

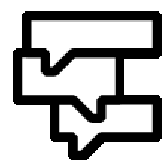
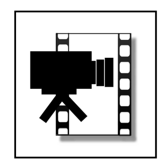
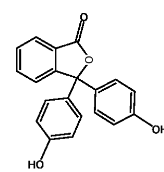
- Will mention some links during the lecture between different definitions.

# What is a kernel

In the context of these lectures...

- A kernel $k$ is a function

$$\begin{array}{rccc} k: & \mathcal{X} \times \mathcal{X} & \longmapsto & \mathbb{R} \\ & (\mathbf{x}, \mathbf{y}) & \longrightarrow & k(\mathbf{x}, \mathbf{y}) \end{array}$$

- which compares two objects of a space $\mathcal{X}$, *e.g.*...
  - strings, texts and sequences, 
  - images, audio and video feeds,   
  - graphs, interaction networks and 3D structures 

- whatever actually... time-series of graphs of images? graphs of texts?...

# Fundamental properties of a kernel

**symmetric**

$$k(\mathbf{x},\mathbf{y}) = k(\mathbf{y},\mathbf{x}).$$

**positive-(semi)definite**

for any *finite* family of points $\mathbf{x}_1,\cdots,\mathbf{x}_n$ of $\mathcal{X}$, the matrix

$$K = \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}_1) & k(\mathbf{x}_1,\mathbf{x}_2) & \cdots & k(\mathbf{x}_1,\mathbf{x}_i) & \cdots & k(\mathbf{x}_1,\mathbf{x}_n) \\ k(\mathbf{x}_2,\mathbf{x}_1) & k(\mathbf{x}_2,\mathbf{x}_2) & \cdots & k(\mathbf{x}_2,\mathbf{x}_i) & \cdots & k(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ k(\mathbf{x}_i,\mathbf{x}_1) & k(\mathbf{x}_i,\mathbf{x}_2) & \cdots & k(\mathbf{x}_i,\mathbf{x}_i) & \cdots & k(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n,\mathbf{x}_1) & k(\mathbf{x}_n,\mathbf{x}_2) & \cdots & k(\mathbf{x}_n,\mathbf{x}_i) & \cdots & k(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix} \succeq 0$$

is positive semidefinite (has a nonnegative spectrum).

$K$ is often called the **Gram matrix** of $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ using $k$

# What can we do with a kernel?

# The setting

- Pretty simple setting: a set of objects $\mathbf{x}_1, \cdots, \mathbf{x}_n$ of $\mathcal{X}$
- **Sometimes** additional information on these objects
  - labels $\mathbf{y}_i \in \{-1, 1\}$ or $\{1, \cdots, \#(\text{classes})\}$,
  - scalar values $\mathbf{y}_i \in \mathbb{R}$,
  - associated object $\mathbf{y}_i \in \mathcal{Y}$
- A kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$.

# A few intuitions on the possibilities of kernel methods

Important concepts and perspectives

- The functional perspective: represent **points as functions**.
- The new or **alternative dot-product** perspective.
- **Nonlinearity** : linear combination of kernel evaluations.
- Summary of a sample through its **kernel matrix**.

# Represent any point in $\mathcal{X}$ as a function

For every $\mathbf{x}$, the map
$$\mathbf{x} \longrightarrow k(\mathbf{x}, \cdot)$$
associates to $\mathbf{x}$ a function $k(\mathbf{x}, \cdot)$ from $\mathcal{X}$ to $\mathbb{R}$.

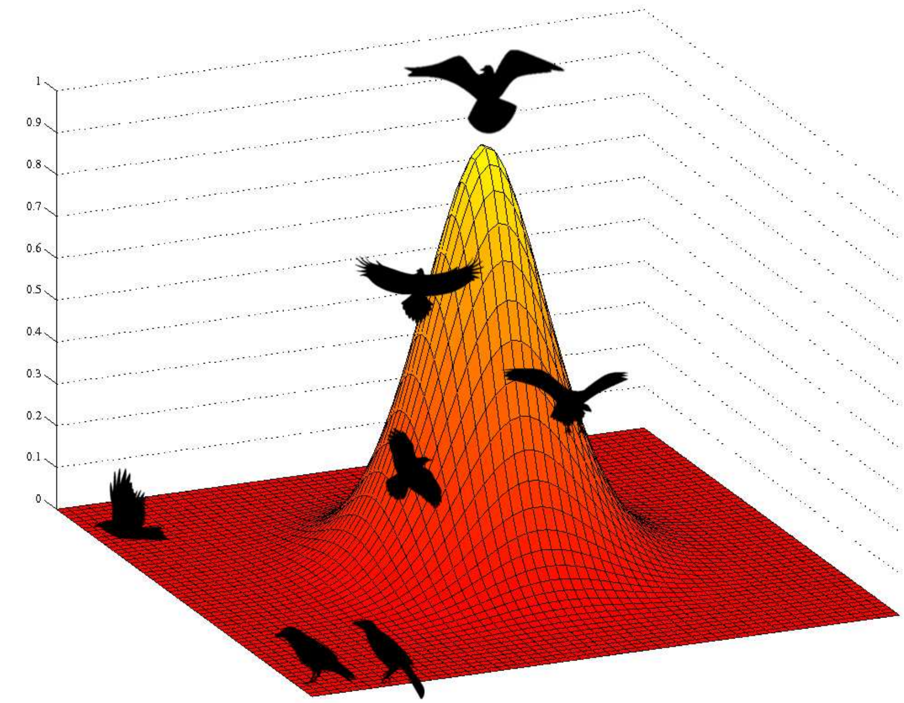
- Suppose we have a kernel $k$ on bird images

- Suppose for instance

$k\,($  $,$  $) = .32$

# Represent any point in $\mathcal{X}$ as a function

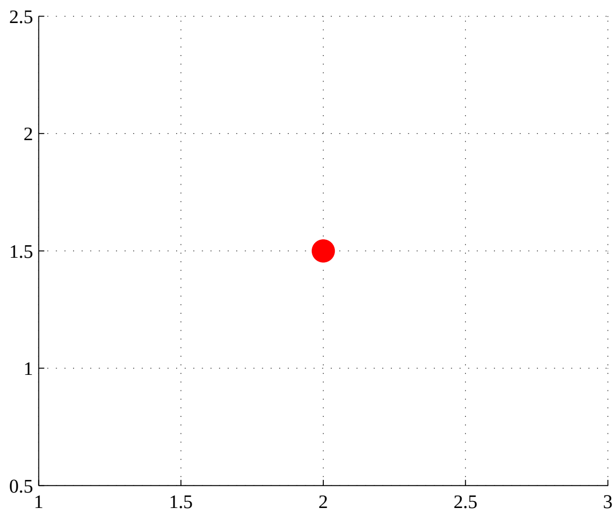
- We examine one image in particular: 
- With kernels, we get a **representation** of that bird as a real-valued function, defined on the space of birds, represented here as $\mathbb{R}^2$ for simplicity.



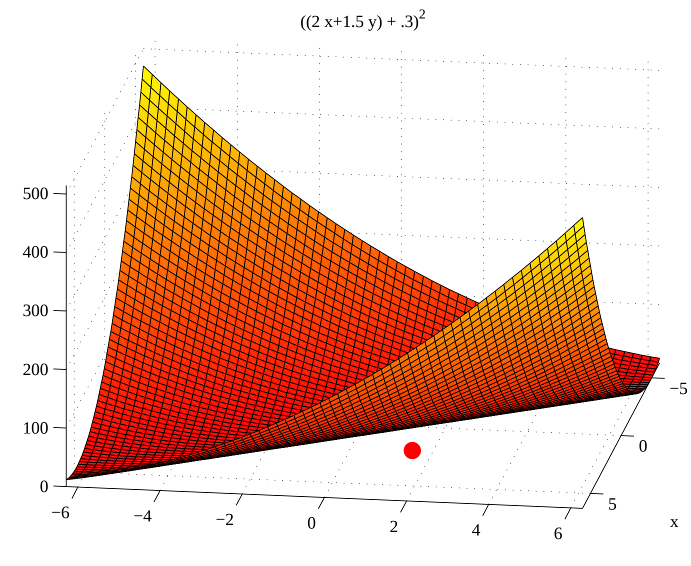

schematic plot of $k($  $, \cdot\ )$.

# Represent any point in $\mathcal{X}$ as a function

- If the bird example was confusing...
- $k\left(\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right],\begin{bmatrix}x'\\y'\end{bmatrix}\right) = \left(\left[\begin{smallmatrix}x & y\end{smallmatrix}\right]\begin{bmatrix}x'\\y'\end{bmatrix} + .3\right)^2$
- From a point in $\mathbb{R}^2$ to a function defined over $\mathbb{R}^2$.




$\longrightarrow$




- We assume implicitly that the **functional representation** will be more useful than the **original representation**.

# Dot-product perspective

- Suppose $\mathcal{X} = \mathbb{R}^d$.
- The simplest kernel: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$.
- For a data sample $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$.
- In matrix form, $X = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} \in \mathbb{R}^{d \times n}$.
- In standard linear algebra, the Gram matrix of $X$ is

$$K = \left[\mathbf{x}_i^T \mathbf{x}_j\right]_{1 \le i,j \le n} = X^T X.$$

# Dot-product perspective

- Consider a different kernel $k_G(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$,

$$K_G = \left[k_G(\mathbf{x}_i,\mathbf{x}_j)\right]_{1\leq i,j\leq n}.$$

- obviously $\mathbf{x}_i^T\mathbf{x}_j \neq k_G(\mathbf{x}_i,\mathbf{y}_j)$.
- is there a representation $\xi_i \in \mathbb{R}^{??}$ for each point such that $\xi_i^T\xi_j = k_G(\mathbf{x}_i,\mathbf{x}_j)$?
- Linear algebra to the rescue: $K = PDP^T$, $U = P\sqrt{D}P^T$, hence $K = U^TU$, providing $U = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \xi_1 & \xi_2 & \cdots & \xi_n \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} \in \mathbb{R}^{\boldsymbol{n}\times n}$.

# Dot-product perspective

- In summary, we have defined $n$ vectors such that

$$\left[k_G(\mathbf{x}_i, \mathbf{x}_j)\right] = \left[\xi_i^T \xi_j\right]$$

- Great: for each $\mathbf{x}_i$ we have a vector representation $\xi_i$.
- Problem:
  - this representation depends explicitly on the sample $X$.
  - For a new $\mathbf{x}_{n+1}$, difficult to find $\xi_{n+1}$ such that $\xi_{n+1}^T \xi_j = k_G(\mathbf{x}_{n+1}, \mathbf{x}_j)$.
- **We will see that there exists a mapping $\phi$**, such that
  - $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a dot-product space,
  - which gives a dot product representation for $k$,

$$k_G(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle.$$

for **all points** $(\mathbf{x}, \mathbf{y})$...

# Decision functions as linear combination of kernel evaluations

- Linear decisions functions are a major tool in statistics, that is functions

$$f(\mathbf{x}) = \beta^T \mathbf{x} + \beta_0.$$

- Implicitly, a point $\mathbf{x}$ is processed depending on its characteristics $x_i$,

$$f(\mathbf{x}) = \sum_{i=1}^{d} \boldsymbol{\beta_i} x_i + \boldsymbol{\beta_0}.$$

  the free parameters are scalars $\boldsymbol{\beta_0, \beta_1, \cdots, \beta_d}$.

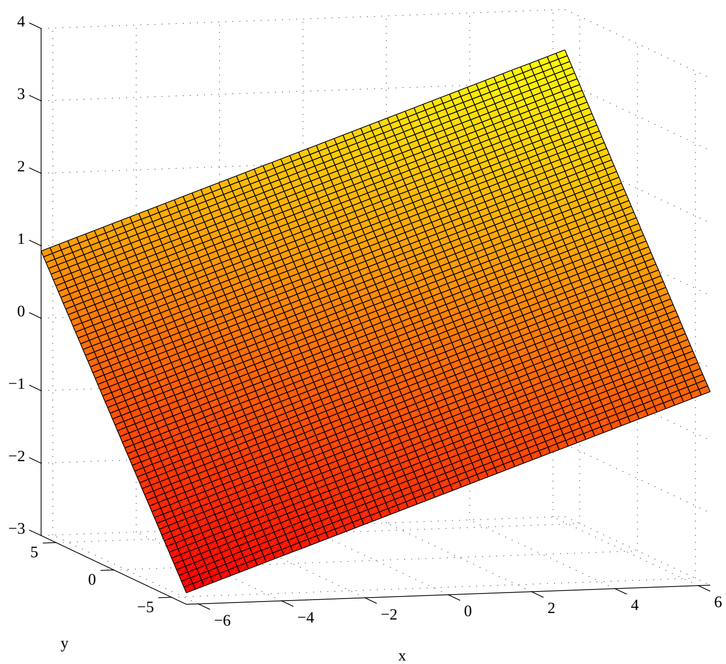
- Kernel methods yield candidate decision functions

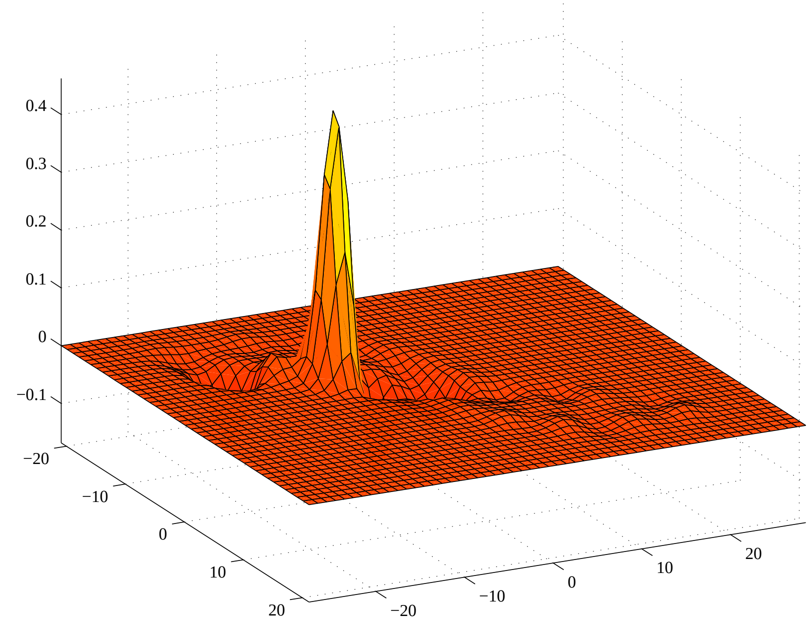
$$f(\mathbf{x}) = \sum_{j=1}^{\boldsymbol{n}} \boldsymbol{\alpha_j} k(\mathbf{x}_j, \mathbf{x}) + \boldsymbol{\alpha_0}.$$

  the free parameters are scalars $\boldsymbol{\alpha_0, \alpha_1, \cdots, \alpha_n}$.

# Decision functions as linear combination of kernel evaluations

- linear decision surface / linear expansion of **kernel surfaces** (here $k_G(\mathbf{x}_i, \cdot)$)




- Kernel methods are considered **non-linear** tools.
- Yet not completely "nonlinear" $\rightarrow$ only one-layer of nonlinearity.

$$\boxed{\text{kernel methods use the data as a functional base to define decision functions}}$$

# Decision functions as linear combination of kernel evaluations

database $\{\mathbf{x}_i, i = 1, \dots, N\}$

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \; k\,(\mathbf{x}_i, \mathbf{x})$$

**kernel definition**

weights $\alpha$ estimated
with a kernel machine

- $f$ is any predictive function of interest of a new point $\mathbf{x}$.
- Weights $\alpha$ are **optimized** with a kernel machine ($e.g.$ support vector machine)

**intuitively, kernel methods provide decisions based on how *similar* a point $\mathbf{x}$ is to each instance of the training set**

# The Gram matrix perspective

- Imagine a little task: you have read 100 novels so far.

- You would like to know whether you will enjoy reading a **new** novel.
- A few options:
  - read the book...
  - have friends read it for you, read reviews.
  - try to guess, based on the novels you read, if you will like it

# The Gram matrix perspective

$\boxed{\text{Two distinct approaches}}$

- Define what **features** can characterize a book.
  - Map each book in the library onto vectors

$$\longrightarrow \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

typically the $x_i$'s can describe...

▷ # pages, language, year 1st published, country,

▷ coordinates of the main action, keyword counts,

▷ author's prizes, popularity, booksellers ranking

- Challenge: find a decision function using 100 ratings and features.

# The Gram matrix perspective

- Define what makes **two novels similar**,
    - Define a kernel $k$ which quantifies novel similarities.
    - Map the library onto a Gram matrix

$$\longrightarrow \quad K = \begin{bmatrix} k(b_1,b_1) & k(b_1,b_2) & \cdots & k(b_1,b_{100}) \\ k(b_2,b_1) & k(b_2,b_2) & \cdots & k(b_2,b_{100}) \\ \vdots & \vdots & \ddots & \vdots \\ k(b_n,b_1) & k(b_n,b_2) & \cdots & k(b_{100},b_{100}) \end{bmatrix}$$

- Challenge: find a decision function that takes this $100 \times 100$ matrix as an input.

# The Gram matrix perspective

Given a new novel,

- with the **features approach**, the prediction can be rephrased as **what are the features of this new book**? what **features** have I found in the past that were good indicators of my taste?

- with the **kernel approach**, the prediction is rephrased as **which novels this book is similar or dissimilar to?** what **pool of books** did I find the most influentials to define my tastes accurately?

kernel methods **only use kernel similarities**, do not consider features.

Features can help define similarities, but **never considered elsewhere**.

# The Gram matrix perspective

$\boxed{\text{In summary}}$

- A feature based analysis of a data-driven problem:

$$\text{objects } o_1, \cdots, o_n \quad \longrightarrow \quad \text{feature vectors } X = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} \in \mathbb{R}^{\boldsymbol{d} \times n}$$

- A similarity based analysis of a data driven problem:

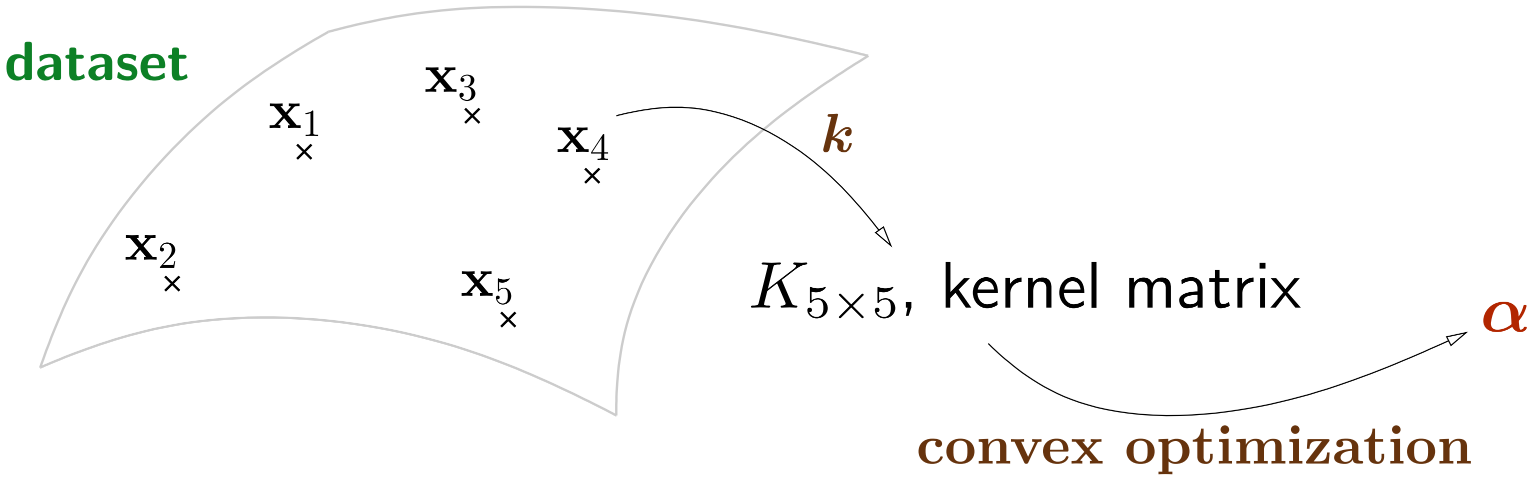
$$\text{objects } o_1, \cdots, o_n \quad \rightarrow \quad \text{Gram } K = \begin{bmatrix} k(o_1, o_1) & k(o_1, o_2) & \cdots & k(o_1, o_n) \\ k(o_2, o_1) & k(o_2, o_2) & \cdots & k(o_2, o_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(o_n, o_1) & k(o_n, o_2) & \cdots & k(o_n, o_n) \end{bmatrix} \in \mathbb{R}^{\boldsymbol{n} \times n}$$

- Some parallels (can define $K = X^T X$ or $X = \sqrt{K}$ or Cholesky) but...

$\boxed{\text{Algorithms use either features } \textit{or} \text{ (kernel) similarities.}}$

# The Gram matrix perspective

in kernel methods, clear separation between the kernel...




and **Convex optimization** (thanks to psdness of $K$, more later) to output the $\alpha$'s.

# Outline of the lectures

# Outline

- Mathematical considerations $(\leq 80's)$
  - Reproducing Kernel Hilbert Spaces
  - positive-definiteness, negative definiteness $etc.$.
  - kernels, similarities and distances
- Defining kernels
  - Standard kernels $(\leq 80's)$
  - Statistical modeling & kernels $(> 1998)$
  - Algebraic structures and kernels
- Kernel algorithms
  - representer theorem
  - unsupervised techniques, eigenfunctions of samples $(\geq 1998)$
  - supervised learning, SVM $(\geq 1995)$
  - density estimation and novelty detection $(\geq 1999)$
- Selecting kernels
  - parameter tuning $(\geq 00's)$
  - multiple kernel learning $(\geq 2004)$

# Mathematical Considerations

different definitions and properties of the same mathematical object

# space of functions

- In the next slides we focus on

$$\boxed{\textbf{reproducing kernel Hilbert spaces} \text{ (RKHS)}}$$

- This term is ubiquitous in the kernel methods literature.
- "Old" mathematics [Mer09], [Aro50]. Survey in [BTA03].

- Reminder: a **Hilbert space** is a
  - vector space, possibly infinite dimensional,
  - equipped with a dot-product, $i.e.$
    - ▷ a bilinear symmetric application
    - ▷ which satisfies $\langle x, x\rangle \geq 0$, equal to $0$ only with $x = 0$.
  - complete (all Cauchy sequences **converge** inside the space).
- **reproducing kernel**... a new term.

# reproducing kernels

- Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$.

**Definition 1** (RKHS). *$\mathcal{H}$ is said to be a reproducing kernel Hilbert space if every linear map of the form $L_{\mathbf{x}} : f \mapsto f(\mathbf{x})$ from $\mathcal{H}$ to $\mathbb{R}$ is continuous for any $\mathbf{x}$ in $\mathcal{X}$.*

Where is the **reproducing kernel** in this definition?

# reproducing kernels

- By the **Riesz representation theorem**
  - *Any continuous linear functional* $L(\cdot)$ *on* $\mathcal{H}$ *can be written uniquely* $\langle \mathbf{u}, \cdot\rangle_{\mathcal{H}}$

  we hence have that:

$$\forall \mathbf{x} \in \mathcal{X}, \exists!\, k_{\mathbf{x}} \in \mathcal{H} \quad | \quad f(\mathbf{x}) = \langle f,\ k_{\mathbf{x}}\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}$$

  $k_{\mathbf{x}}$ is called the point-evaluation functional at the point $\mathbf{x}$.

- Since $\mathcal{H}$ is a space of functions, $k_{\mathbf{x}}$ is itself a function. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by

$$k(\mathbf{x}, \mathbf{y}) \overset{\text{def}}{=} k_{\mathbf{x}}(\mathbf{y}).$$

- $k$ is the **reproducing kernel** of $\mathcal{H}$ and it is determined entirely by $\mathcal{H}$ through the Riesz representation theorem which guarantees the **unicity** of $k_{\mathbf{x}}$ for each $\mathbf{x}$.

# positive definite kernels

**Definition 2** (Real-valued Positive Definite Kernels). *A symmetric function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is a positive definite (p.d.) kernel on* $\mathcal{X}$ *if*

$$\sum_{i,j=1}^{n} c_i c_j \, k\left(x_i, x_j\right) \geq 0,$$

*holds for any* $n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}$ *and* $c_1 \ldots, c_n \in \mathbb{R}$.

With this definition, the set of p.d. kernels $\mathcal{P}(\mathcal{X})$ is a closed, convex pointed cone:

- $\forall \lambda \geq 0, k$ p.d.kernel $\Rightarrow \lambda k$ is p.d.
- $\forall \lambda \geq 0, k_1, k_2$ p.d.kernel, $\lambda k_1 + (1-\lambda) k_2$ p.d. kernel.
- $k$ p.d. kernel, $-k$ p.d. kernel $\Rightarrow k = 0$.
- if $k_n \in \mathcal{P}(\mathcal{X})$ and $\lim_{n\infty} k_n = k$ then $k \in \mathcal{P}(\mathcal{X})$.

# kernels: two definitions

- Have mathematicians screwed up again and used the term kernel separately?

**reproducing kernels** (functional analysis, topology)

$$\overset{?}{\neq}$$

**positive definite kernels** (positivity and linear algebra)

- luckily, no screw up: the two notions are equivalent.

# Moore-Aronszajn (1950) theorem

**Theorem 1.** *Let $\mathcal{X}$ be any set. An application $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a reproducing kernel iff it is a positive definite kernel*

- A first proof was given by Mercer (1909) when $\mathcal{X}$ is compact.
- Hence the *Mercer* kernel term sometimes used.
- In many applications compacity is never really mentioned...
- ... hence *positive definite* or *reproducing* are more accurate terms.
- In the general case the result was proved by Moore & Aronszajn in 1950 (separately).

# Moore-Aronszajn (1950) theorem, proof outline

- If $k$ is a r.k., $k(\mathbf{x},\mathbf{y}) = \langle k(\mathbf{x},\cdot), k(\mathbf{y},\cdot)\,\rangle = \langle k(\mathbf{y},\cdot), k(\mathbf{x},\cdot)\,\rangle = k(\mathbf{y},\mathbf{x})$,

$$\sum_{i,j=1}^{n} c_i c_j k(\mathbf{x}_i,\mathbf{x}_j) = \left\| \sum_{i=1}^{n} k(\mathbf{x}_i,\cdot) \right\|_{\mathcal{H}}^2 \geq 0.$$

- if $k$ is a p.d. kernel,
  - Define the vector space $\tilde{\mathcal{H}} = \mathrm{span}\{k(\mathbf{x},\cdot)\}$.
  - Define $\langle\cdot,\cdot\rangle_{\tilde{\mathcal{H}}}$ for $f = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i,\cdot)$ and $g = \sum_{j=1}^{n} \beta_j k(\mathbf{y}_j,\cdot)$ as

$$\langle f, g\,\rangle = \sum_{i,j=1}^{m,n} \alpha_i \beta_j k(\mathbf{x}_i,\mathbf{y}_j).$$

  - even if $\{k(\mathbf{x},\cdot)\}_{\mathbf{x}\in\mathcal{X}}$ is not a l.i. family ($i.e.$ no unicity of $\alpha$ or $\beta$) we have

$$\langle f, g\,\rangle = \sum_{i=1}^{m} \alpha_i g(\mathbf{x}_i) = \sum_{j=1}^{n} \beta_i f(\mathbf{y}_i).$$

- $\langle \cdot, \cdot \rangle_{\tilde{\mathcal{H}}}$ is **bilinear symmetric** and **p.d.** through the p.d. of $k$.
- Cauchy-Schwartz is verified thanks to p.d. of the Gram matrix on all $\mathbf{x}_i, \mathbf{y}_j$.

$$\begin{bmatrix} \alpha^T & \mathbf{0}_n^T \\ \mathbf{0}_m^T & \beta^T \end{bmatrix} \begin{bmatrix} K_{\mathbf{x}} & K_{\mathbf{x},\mathbf{y}} \\ K_{\mathbf{x},\mathbf{y}}^T & K_{\mathbf{y}} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{0}_m \\ \mathbf{0}_n & \beta \end{bmatrix} = \begin{bmatrix} \alpha^T K_{\mathbf{x}} \alpha & \alpha^T K_{\mathbf{x},\mathbf{y}} \beta \\ \beta^T K_{\mathbf{x},\mathbf{y}}^T \alpha & \beta^T K_{\mathbf{y}} \beta \end{bmatrix} \succeq 0$$

hence

$$\|f\|^2 \|g\|^2 = (\alpha^T K_{\mathbf{x}} \alpha)(\beta^T K_{\mathbf{y}}) \geq (\alpha^T K_{\mathbf{x},\mathbf{y}} \beta)^2 = \langle f, g \rangle^2.$$

- Hence $\|f\| = 0 \Rightarrow f = 0$ since

$$\forall \mathbf{x} \in \mathcal{X}, |f(\mathbf{x})| = \langle f, k(\mathbf{x}, \cdot) \rangle \leq \|f\| \sqrt{k(\mathbf{x}, \mathbf{x})} = 0.$$

- $\tilde{\mathcal{H}}$ is a pre-Hilbertian. For any Cauchy sequence $f_n$ in $\tilde{\mathcal{H}}$, and $\mathbf{x} \in \mathcal{X}$

$$|f_m(\mathbf{x}) - f_n(\mathbf{x})| = \langle f_n - f_m, k(\mathbf{x}, \cdot) \rangle \leq \|f_n - f_m\| \sqrt{k(\mathbf{x}, \mathbf{x})} \rightarrow 0,$$

  $f_n(\mathbf{x})$ is thus Cauchy in $\mathbb{R}$ and has thus a limit. $f_n$ has thus a limit.
- We add all such limits to **complete** $\tilde{\mathcal{H}}$ into $\mathcal{H}$.
- still a few steps more (show the r.k. of $\mathcal{H}$ is still $k$).

# The kernel paradigm

- A simple function $k$ that is p.d. defines a Hilbert space of functions:
  - its elements,
$$f(\mathbf{x}) = \sum_{i=1}^{\infty} \alpha_i k(\mathbf{x}_i, \mathbf{x}),$$
and Cauchy limits of such functions,
  - their dot-product,
$$\langle f, g \rangle_{\mathcal{H}} = \langle \sum_{i=1}^{\infty} \alpha_i k(\mathbf{x}_i, \cdot), \sum_{i=1}^{\infty} \beta_i k(\mathbf{y}_i, \cdot) \rangle_{\mathcal{H}} = \sum_{i,j=1}^{\infty} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{y}_j).$$
  - their norm,
$$\|f\|^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{i,j=1}^{\infty} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j).$$

We usually focus on *positive definite* kernels but don't forget the **reproducing** story

# Another alternative definition

**Definition 3** (Reproducing Kernel). *A real-valued function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is a reproducing kernel of a Hilbert space* $\mathcal{H}$ *of real-valued functions on* $\mathcal{X}$ *if and only if*

*(i)* $\forall t \in \mathcal{X}, \;\; k(\cdot, t) \in \mathcal{H}$;

*(ii)* $\forall t \in \mathcal{X}, \forall f \in \mathcal{H}, \;\; \langle f, k(\cdot, t) \rangle = f(t)$.

- straightforward to prove equivalence with the first characterization.

# A word on continuity

**Proposition 2.** *Let $k$ be a positive denite kernel on a* ***topological*** *space $\mathcal{X}$ , and $\mathcal{H}$ the associated RKHS. If $k(\mathbf{x},\mathbf{y})$ is continuous for all $\mathbf{x},\mathbf{y}\in\mathcal{X}$, then all the functions in $\mathcal{H}$ are* ***continuous*** *functions of $\mathcal{X}\mapsto\mathbb{R}$.*

***Proof.*** Let $f$ be an arbitrary function in $\mathcal{H}$,

$$|f(x)-f(y)| = |\langle f, k(\mathbf{x},\cdot)-k(\mathbf{y},\cdot)\rangle| \underset{CS}{\leq} \|f\|\,\|k(\mathbf{x},\cdot)-k(\mathbf{y},\cdot)\|,$$

Remember that $\|k(\mathbf{x},\cdot)-k(\mathbf{y},\cdot)\| = k(\mathbf{x},\mathbf{x})+k(\mathbf{y},\mathbf{y})-2k(\mathbf{x},\mathbf{y})$.

# A more intuitive perspective: Feature maps

**Theorem 3.** *A function $k$ on $\mathcal{X} \times \mathcal{X}$ is a positive definite kernel if and only if there exists a set $T$ and a mapping $\phi$ from $\mathcal{X}$ to $l^2(T)$, the set of real sequences $\{u_t, t \in T\}$ such that $\sum_{t\in T} |u_t|^2 < \infty$, where*

$$\forall(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{X}, \; k(\mathbf{x},\mathbf{y}) = \sum_{t\in T} \phi(\mathbf{x})_t \, \phi(\mathbf{y})_t = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{l^2(X)}$$

- A very popular perspective in the machine learning world.
- Equivalent to previous definitions, less stressed in the RHKS literature.

$$\mathbf{x} \longrightarrow \phi(\mathbf{x}) = \begin{bmatrix} \vdots \\ \vdots \\ \phi(\mathbf{x})_t \\ \vdots \\ \vdots \end{bmatrix}_{t\in T}$$

where the $\phi_t$ are a set of – possibly infinite but countable – features.

# kernels $\rightarrow$ Gram matrices

- If $X = \{\mathbf{x}_i\}_{i \in I}$ in $\mathcal{X}$,
$$K_X = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j \in I} \succeq 0.$$

- If one applies *any* transformation of $K_X$ which keeps eigenvalues nonnegative,
$$r : \begin{array}{ccc} \mathbf{S}_n & \longmapsto & \mathbf{S}_n \\ K & \longrightarrow & r(K), \end{array}$$

  $r(K)$ is a valid positive definite matrix and hence a kernel on $X$.

- examples: $K + t (t > 0), K^2, e^K, etc.$
- in fact, if $K = P \Delta P^T$, any transformation that preserves the spectrum's non-negativity would be ok.
- Yet... this kernel is only valid on $X$, the sample, not the whole space $\mathcal{X}$.

$$\boxed{\text{Meaning somehow... Gram matrices} \nrightarrow \text{kernels}}$$

# positive definite kernels and distances

- Kernels are often called similarities.
- the **higher** $k(\mathbf{x},\mathbf{y})$, the more similar $\mathbf{x}$ and $\mathbf{y}$.
- With distances, the **lower** $d(\mathbf{x},\mathbf{y})$, the closer $\mathbf{x}$ and $\mathbf{y}$.

- Many distances exist in the literature. Can they be used to define kernels?

| |
|---|
| what is the link between kernels and distances? |
| **high similarity** $\stackrel{?}{=}$ **small distance** |

- At least true for the Gaussian kernel $k(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}$...
- Important theorems taken from [BCR84].

# Distances

**Definition 4** (Distances, or metrics). *A* ***nonnegative-valued*** *function $d$ on $\mathcal{X} \times \mathcal{X}$ is a distance if it satisfies, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$:*

*(i)* $d(\mathbf{x}, \mathbf{y}) \geq 0$, *and* $d(\mathbf{x}, \mathbf{y}) = 0$ *if and only if* $\mathbf{x} = \mathbf{y}$ *(non-degeneracy)*

*(ii)* $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ *(symmetry),*

*(iii)* $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ *(triangle inequality)*

- Very simple example: if $\mathcal{X}$ is a Hilbert space, $\|\mathbf{x} - \mathbf{y}\|$ is a distance. It is usually called a... Hilbertian distance.
- By extension, any distance $d(\mathbf{x}, \mathbf{y})$ which can be written as $\|\phi(\mathbf{x}) - \phi(\mathbf{y})\|$ where $\phi$ maps $\mathcal{X}$ to any Hilbert space is called a **Hilbertian metric**.
- Useful. To build Gaussian kernel, Laplace kernels $k(\mathbf{x}, \mathbf{y}) = e^{-t\|\mathbf{x}-\mathbf{y}\|}$...
- Yet does not suffice:

# the missing link: negative definite kernels

**Definition 5** (Negative Definite Kernels)**.** *A symmetric function* $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is a negative definite (n.d.) kernel on* $\mathcal{X}$ *if*

$$\sum_{i,j=1}^{n} c_i c_j \, \psi \left( x_i, x_j \right) \leq 0 \tag{1}$$

*holds for any* $n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}$ *and* $c_1 \ldots, c_n \in \mathbb{R}$ *such that* $\sum_{i=1}^{n} c_i = 0$.

- Example $\psi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$.
  - prove by decomposing into $\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle$
- $\mathcal{N}(\mathcal{X})$ is also a closed convex cone.

important example: $k$ is p.d. $\Rightarrow -k$ is n.d.
Converse completely false.

## negative definite kernels & positive definite kernels

A first link between these two kernels:

**Proposition 4.** *Let* $x_0 \in \mathcal{X}$ *and let* $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a symmetric kernel. Let*

$$\varphi(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \psi(\mathbf{x}, x_0) + \psi(\mathbf{y}, x_0) - \psi(\mathbf{x}, \mathbf{y}) - \psi(x_0, x_0).$$

*Then* $k$ *is positive definite* $\Leftrightarrow$ $\psi$ *is negative definite.*

- Example: $\|\mathbf{x} - x_0\|^2 + \|\mathbf{y} - x_0\|^2 - \|\mathbf{x} - \mathbf{y}\|^2$ is a p.d. kernel.

***Proof.***

- $\Rightarrow$ For $\mathbf{x}_1, \cdots, \mathbf{x}_n$, and $c_1, \cdots, c_n$ s.t. $\sum_{i=1}^n c_i = \mathbf{0}$,

$$\sum_{i,j=1}^n c_i c_j \varphi(\mathbf{x}_i, \mathbf{x}_j) = -\sum_{i,j=1}^n c_i c_j \psi(\mathbf{x}_i, \mathbf{x}_j) \geq 0.$$

- $\Leftarrow$ For $\mathbf{x}_1, \cdots, \mathbf{x}_n$ and $c_1, \cdots, c_n$, let $c_0 = -\sum_{i=1}^n$. Set $\mathbf{x}_0 = x_0$. Then

$$\begin{aligned}
0 &\geq \sum_{i,j=0}^n c_i c_j \psi(\mathbf{x}_i, \mathbf{x}_j) \\
&= \sum_{i,j=1}^n c_i c_j \psi(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^n c_i c_0 \psi(\mathbf{x}_i, x_0) + \sum_{j=1}^n c_0 c_j \psi(x_0, \mathbf{x}_j) + c_0^2 \psi(x_0, x_0). \\
&= \sum_{i,j=1}^n [\psi(\mathbf{x}_i, x_0) + \psi(\mathbf{x}_j, x_0) - \psi(\mathbf{x}_i, \mathbf{y}_j) - \psi(x_0, x_0)] = \sum_{i,j=1}^n c_i c_j \varphi(\mathbf{x}_i, \mathbf{x}_j).
\end{aligned}$$

## negative definite kernels & positive definite kernels

**Proposition 5.** *For a p.d. kernel $k \geq 0$ on $\mathcal{X} \times \mathcal{X}$, the following conditions are equivalent*

*(i)* $-\log k \in \mathcal{N}(\mathcal{X})$,

*(ii)* $k^t$ *is positive definite for all* $t > 0$.

*If $k$ satisfies either, $k$ is said to be* ***infinitely divisible****,*

***Proof.***

- $-\log k = \lim_{n\to\infty} n(1 - k^{\frac{1}{n}})$ which is the limit of a series of n.d. kernels if $(ii)$ is true, hence $(ii) \Rightarrow (i)$.
- conversely, if $-\log k \in \mathcal{N}(\mathcal{X})$ we use Proposition 4. Writing $\psi = -\log k$ and choosing $x_0 \in \mathcal{X}$ we have

$$k^t = e^{-t\psi(\mathbf{x},\mathbf{y})} = e^{t\psi(x_0,x_0)} \boldsymbol{e^{t\varphi(\mathbf{x},\mathbf{y})}} \boldsymbol{e^{-t\psi(\mathbf{x},x_0)}} \boldsymbol{e^{-t\psi(\mathbf{y},x_0)}} \in \mathcal{P}(\mathcal{X})$$

# negative definite kernels: $(\text{Hilbertian distance})^2 + \ldots$

**Proposition 6.** *Let $\psi : \mathcal{X} \times \mathcal{X}$ be a n.d. kernel. Then there is a Hilbert space $H$ and a mapping $\phi$ from $X$ to $H$ such that*

$$\psi(\mathbf{x}, \mathbf{y}) = \|\phi(\mathbf{x}) - \phi(\mathbf{y})\|^2 + f(\mathbf{x}) + f(\mathbf{y}), \tag{2}$$

*where $f : \mathcal{X} \to \mathbb{R}$. If $\psi(x, x) = 0$ for all $\mathbf{x} \in \mathcal{X}$ then $f$ can be chosen as zero. If the set $\{(\mathbf{x}, \mathbf{y}) | \, \psi(\mathbf{x}, \mathbf{y}) = 0\}$ is exactly $\{(\mathbf{x}, \mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ then $\sqrt{\psi}$ is a Hilbertian distance.*

***Proof.*** Fix $x_0$ and define

$$\varphi(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\mathbf{1}}{\mathbf{2}} \left[\psi(\mathbf{x}, x_0) + \psi(\mathbf{y}, x_0) - \psi(\mathbf{x}, \mathbf{y}) - \psi(x_0, x_0)\right].$$

By Proposition 4 $\varphi$ is p.d. hence there is a RKHS and mapping $\phi$ such that $\varphi(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. Hence

$$\begin{aligned} \|\phi(\mathbf{x}) - \phi(\mathbf{y})\|^2 &= \varphi(\mathbf{x}, \mathbf{x}) + \varphi(\mathbf{y}, \mathbf{y}) - 2\varphi(\mathbf{x}, \mathbf{y}) \\ &= \psi(\mathbf{x}, \mathbf{y}) - \frac{\psi(\mathbf{x}, \mathbf{x}) + \psi(\mathbf{y}, \mathbf{y})}{2}. \end{aligned}$$

# distances & negative definite kernels

- whenever a n.d. kernel $\psi$
  - vanishes on the *diagonal*, $i.e.$ on $\{(x, x), x \in \mathcal{X}\}$,
  - is 0 only on the diagonal, to ensure non-degeneracy,

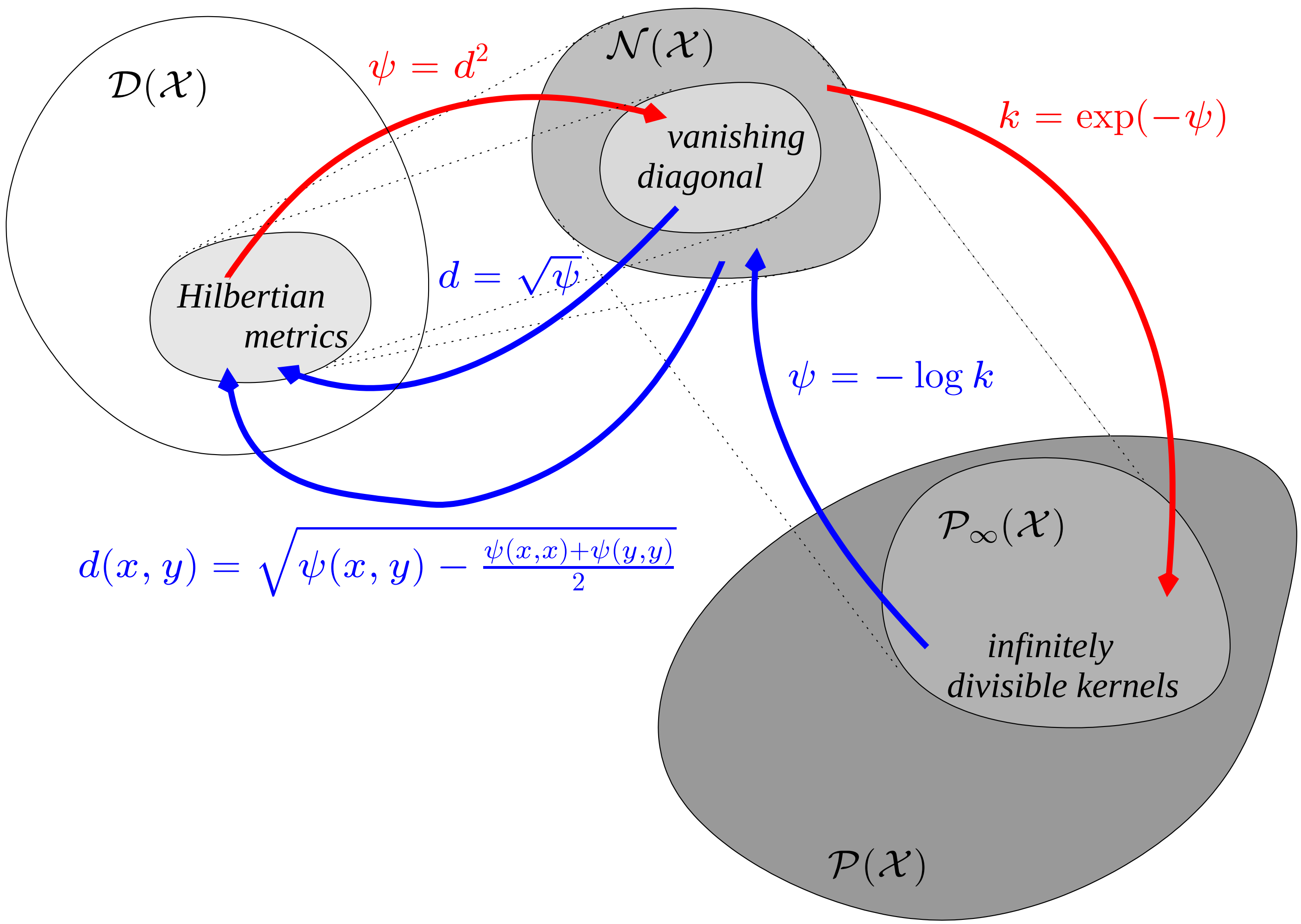
$$\rightarrow \sqrt{\psi} \text{ is a Hilbertian distance for } \mathcal{X}.$$

- **More generally**, for a n.d. kernel $\psi$,

$$\sqrt{\psi(\mathbf{x},\mathbf{y}) - \frac{\psi(\mathbf{x},\mathbf{x})}{2} - \frac{\psi(\mathbf{y},\mathbf{y})}{2}} \text{ is a (pseudo)}\textbf{metric} \text{ for } \mathcal{X}\ .$$

- On the contrary, to each distance does not always correspond a n.d. kernel (Monge-Kantorovich distance, edit-distance $etc.$.)

# In summary...




- Set of distances on $\mathcal{X}$ is $\mathcal{D}(\mathcal{X})$, Negative definite kernels $\mathcal{N}(\mathcal{X})$, positive and infinitely divisible positive kernels $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}_\infty(\mathcal{X})$ respectively.

# Some final remarks on $\mathcal{N}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$

- $\mathcal{N}(\mathcal{X})$ is a cone. Additionally,
  - if $\psi \in \mathcal{N}(\mathcal{X}), \forall c \in \mathbb{R}, \psi + c \in \mathcal{N}(\mathcal{X})$.
  - if $\psi(x,x) \geq 0$ for all $x \in \mathcal{X}$, $\psi^\alpha \in \mathcal{N}(\mathcal{X})$ for $0 < \alpha < 1$ since

$$\psi^\alpha = \frac{\alpha}{\Gamma(1-\alpha)} \int_0^\infty t^{-\alpha-1}(1 - e^{-t\psi})dt$$

    and $\log(1+\psi) \in \mathcal{N}(\mathcal{X})$ since

$$\log(1+\psi) = \int_0^\infty (1 - e^{-t\psi})\frac{e^{-t}}{t}dt.$$

  - if $\psi > 0$, then $\log(\psi) \in \mathcal{N}$ since

$$\log(\psi) = \lim_{c\to\infty} \log\left(\psi + \frac{1}{c}\right) = \lim_{c\to\infty} \log\left(1 + c\psi\right) - \log c$$

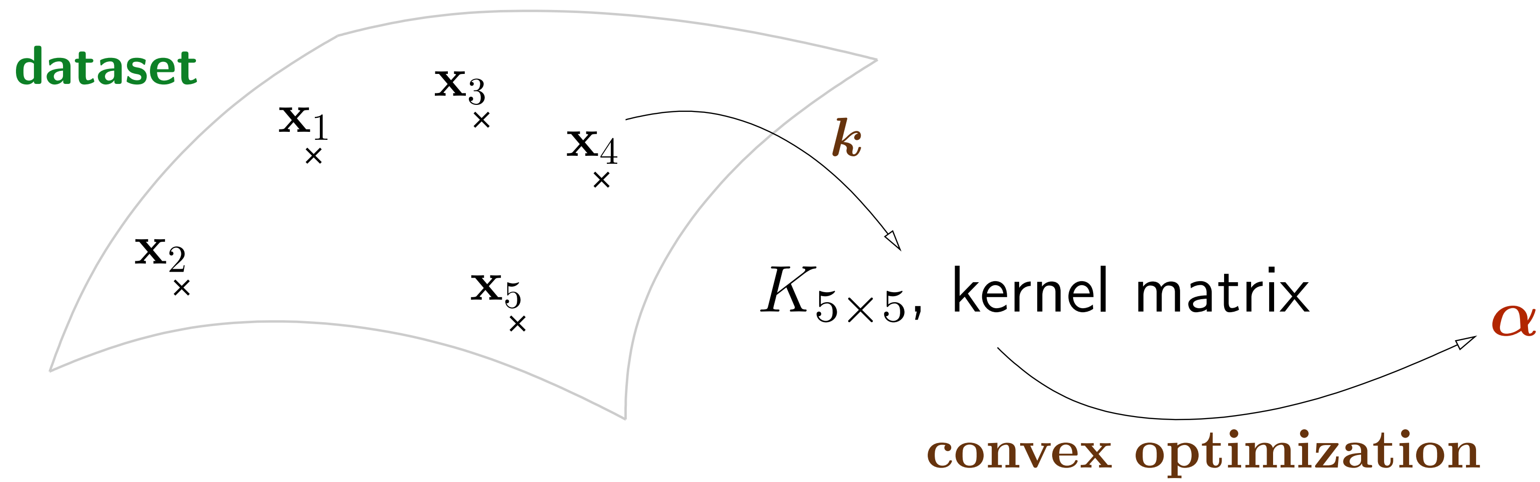
# Some final remarks on $\mathcal{D}(\mathcal{X}), \mathcal{N}(\mathcal{X}), \mathcal{P}(\mathcal{X})$

- $\mathcal{P}(\mathcal{X})$ is a cone. Additionally,
  - The pointwise product $k_1 k_2$ of two p.d. kernels if a p.d. kernel
  - $k^n \in \mathcal{P}(\mathcal{X})$ for $n \in \mathbb{N}$. $(k+c)^n$ too...as well as $\exp(k) \in \mathcal{P}(\mathcal{X})$:
    - ▷ $\exp(k) = \sum_{i=0}^{\infty} \frac{k^i}{i!}$, a limit of p.d. kernels.
    - ▷ $\exp(k) = \exp(-(-k))$ where $-k \in \mathcal{N}(\mathcal{X})$.
- The sum of two infinitely divisible kernels is not necessarily infinitely divisible.
  - $-\log k_1$ and $-\log k_2$ might be in $\mathcal{N}(\mathcal{X})$, but $-\log(k_1 + k_2)$?...

# Defining kernels

# Intuitively an important issue...

Remember that kernel methods drop all previous information




to proceed exclusively with $K$.

if the kernel $K$ is poorly informative, the optimization cannot be very useful...
it is therefore **crucial** that the kernel quantifies **noteworthy similarities**.

# Kernels on vectors

(relatively) easy case: **we are only given feature vectors**, with **no access** to the original data.

- Reminder (copy paste of previous slide!): for a family of kernels $k_1, \cdots, k_n, \cdots$
  - The sum $\sum_{i=1}^{n} \lambda_i k_i$ is p.d., given $\lambda_1, \ldots, \lambda_n \geq 0$
  - The product $k_1^{a_1} \cdots k_n^{a_n}$ is p.d., given $a_1, \ldots, a_n \in \mathbb{N}$
  - $\lim_{n\to\infty} k_n$ is p.d. (if the limit exists!).
- Using these properties we can prove the p.d. of
  - the polynomial kernel $k_p(x, y) = (\langle \mathbf{x}, \mathbf{y} \rangle + b)^d, \quad b > 0, d \in \mathbb{N}$,
  - the Gaussian kernel $k_\sigma(x, y) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$ which can be rewritten as

$$k_\sigma(x, y) = \left[ e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}} e^{-\frac{\|\mathbf{y}\|^2}{2\sigma^2}} \right] \cdot \left[ \sum_{i=0}^{\infty} \frac{\langle \mathbf{x}, \mathbf{y} \rangle^i}{i!} \right]$$

# Kernels on vectors

- the Laplace kernels, using some n.d. kernel weaponry,

$$k_\lambda(x,y) = e^{-\lambda\|\mathbf{x}-\mathbf{y}\|^{\boldsymbol{a}}}, \quad 0<\lambda,\, 0<\boldsymbol{a}\leq 2$$

- the all-subset Gaussian kernel in $\mathbb{R}^d$,

$$k(x,y) = \prod_{i=1}^{d}\left(1+ae^{-b(x_i-y_i)^2}\right) = \sum_{I\subset\{1,\cdots,d\}} a^{\#(I)}e^{-b\|\mathbf{x}_I-\mathbf{y}_I\|^2}.$$

- A variation on the Gaussian kernel: Mahalanobis kernel,

$$k_\Sigma(x,y) = e^{-(\mathbf{x}-\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{y})},$$

  idea: correct for discrepancies between the magnitudes and correlations of different variables.
- Usually $\Sigma$ is the empirical covariance matrix of a sample of points.

# Kernels on vectors

- These kernels can be seen as *meta*-kernels which can use any feature representation.
- Example: Gaussian kernel of Gaussian kernel feature maps,

$$k_{G^2}(\mathbf{x},\mathbf{y}) = k_G\left(e^{-\frac{\|\mathbf{x}-\cdot\|^2}{2\sigma^2}}, e^{-\frac{\|\mathbf{y}-\cdot\|^2}{2\sigma^2}}\right) = e^{-\frac{2-e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}}{2\lambda^2}}.$$

- Not sure this is very useful though!

- Indeed, the real challenge is not to define funky kernels,

$$\boxed{\text{the challenge is to tune the parameters } b, d, \sigma, \Sigma.}$$

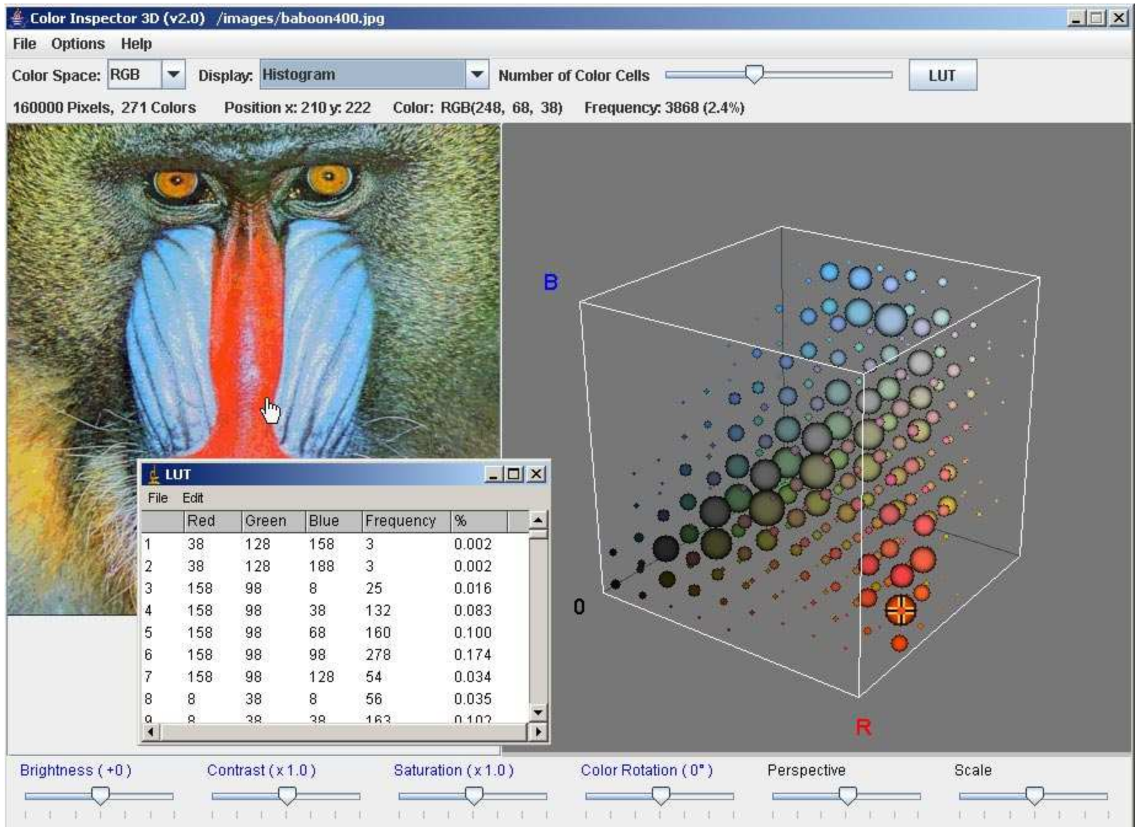
# Kernels on structured objects

- Structured objects?
  - texts, webpages, documents
  - sounds, speech, music,
  - images, video segments, movies,
  - 3d structures, sequences, trees, graphs
- Structured objects means
  - objects with **a tricky structure**,
  - which cannot be simply embedded in a vector space of small dimensionality,
  - without obvious algebraic properties,

$$\boxed{\text{structured object} = \text{that which cannot be represented in a (small) Euclidian space}}$$

# Vectors in $\mathbb{R}^n_+$ and Histograms

- A powerful and popular feature representation for structured objects: **histograms of smaller building-blocks of the object**:




- histograms are simple instances of **probability measures**,
  - nonnegative coordinates, sum up to 1.

# Standard metrics for Histograms

**Information geometry**, introduced yesterday, studies distances between densities.

- Reference : [AN01]
- An abridged bestiary of **negative definite distances** on the probability simplex:

$$\psi_{JD}(\theta,\theta') = h\left(\frac{\theta+\theta'}{2}\right) - \frac{h(\theta)+h(\theta')}{2},$$

$$\psi_{\chi^2}(\theta,\theta') = \sum_i \frac{(\theta_i-\theta'_i)^2}{\theta_i+\theta'_i}, \quad \psi_{TV}(\theta,\theta') = \sum_i |\theta_i-\theta'_i|,$$

$$\psi_{H_2}(\theta,\theta') = \sum_i |\sqrt{\theta_i}-\sqrt{\theta'_i}|^2, \quad \psi_{H_1}(\theta,\theta') = \sum_i |\sqrt{\theta_i}-\sqrt{\theta'_i}|.$$

- Recover kernels through

$$k(\theta,\theta') = e^{-t\psi}, \quad t>0$$

# Information Diffusion Kernel [LL05,ZLC05]

- Solve the heat equation on the multinomial manifold, using the Fisher metric
- Approximate the solution with

$$k_{\Sigma_d}(\theta,\theta') = e^{-\frac{1}{t}\arccos^2(\sqrt{\theta}\cdot\sqrt{\theta'})},$$

- $\arccos^2$ is the **squared** **geodesic distance** between $\theta$ and $\theta'$ as elements from the unit sphere ($\theta_i \rightarrow \sqrt{\theta_i}$).
- In [ZLC05]: the use of

$$k_{\Sigma_d}(\theta,\theta') = e^{-\frac{1}{t}\arccos(\sqrt{\theta}\cdot\sqrt{\theta'})},$$

  is advocated.
- the geodesic distance is a n.d. kernel on the *whole sphere* ($\arccos^2$ is not).

# Transportation Metrics for Histograms

Beyond information geometry, the family of **transportation distances**.

- Suppose $\mathbf{r} = (r_1, \cdots, r_d)$ and $\mathbf{c} = (c_1, \cdots, c_d)$ are two histograms in $\mathbb{R}_+^n$.
- Define the set of transportations

$$U(\mathbf{r}, \mathbf{c}) = \{F \in \mathbb{R}_+^{d\times d} \,|\, F\mathbf{1} = \mathbf{r}, F^T\mathbf{1} = \mathbf{c}\}.$$

- Transportation distances between $\mathbf{r}$ and $\mathbf{c}$:

$$d_{\text{cost}}(\mathbf{rc}) = \min_{F\in U(\mathbf{r},\mathbf{c})} \text{cost}(F).$$

**Monge-Kantorovich**: $\text{cost}(F) = \langle F, D\rangle$ where $D$ is a n.d. matrix.

- $d_{\text{cost}}$ is **not** n.d. in the general case.
- Alternatives:

$$k_{\text{cost}}(\mathbf{rc}) = \int_{F\in U(\mathbf{r},\mathbf{c})} e^{-\text{cost}(F)}.$$

- works when cost $= 0$: the volume of $U(\mathbf{r}, \mathbf{c})$ is a p.d. kernel of $\mathbf{r}$ and $\mathbf{c}$. [Cut07]

# Statistical Modeling and Kernels

| Histograms cannot always summarize efficiently the structures of $\mathcal{X}$ |
|---|

- Statistical models of complex objects provide richer explanations:
  - Hidden Markov Models for sequences and time-series,
  - VAR, VARMA, ARIMA *etc.* models for time-series,
  - Branching processes for trees and graphs
  - Random Markov Fields for images *etc.*
- $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ are interpreted as i.i.d realizations of one or many densities on $\mathcal{X}$.
- These densities belong to a model $\{\boldsymbol{p_\theta}, \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^{\boldsymbol{d}}\}$

| Can we use **generative** (statistical) **models**<br>in<br>**discriminative** (kernel and metric based) **methods**? |
|---|

# Fisher Kernel

- The Fisher kernel [JH99] between two elements $\mathbf{x}, \mathbf{y}$ of $\mathcal{X}$ is

$$k_{\hat\theta}(\mathbf{x},\mathbf{y}) = \left(\frac{\partial \ln \boldsymbol{p_\theta}(\mathbf{x})}{\partial\theta}\Big|_{\hat{\boldsymbol\theta}}\right)^T \boldsymbol{J}_{\hat{\boldsymbol\theta}}^{-1} \left(\frac{\partial \ln \boldsymbol{p_\theta}(\mathbf{y})}{\partial\theta}\Big|_{\hat{\boldsymbol\theta}}\right),$$

  - $\hat{\boldsymbol\theta}$ has been selected using sample data ($e.g.$ MLE),
  - $\boldsymbol{J}_{\hat{\boldsymbol\theta}}^{-1}$ is the Fisher information matrix computed in $\hat\theta$.

- The statistical model $\{\boldsymbol{p_\theta}, \boldsymbol\theta \in \boldsymbol\Theta\}$ provides:

  - finite dimensional *features* through the **score vectors**,
  - A ***Mahalanobis metric*** associated with these vectors through $J_{\hat\theta}$.

- Alternative formulation:

$$k_{\hat\theta}(x,y) = e^{-\frac{1}{\sigma^2}\left(\nabla_{\hat\theta}\ln p_\theta(\mathbf{x}) - \nabla_{\hat\theta}\ln p_\theta(\mathbf{y})\right)^T J_{\hat\theta}^{-1}\left(\nabla_{\hat\theta}\ln p_\theta(\mathbf{x}) - \nabla_{\hat\theta}\ln p_\theta(\mathbf{y})\right)}.$$

  with the meta-kernel idea.

# Fisher Kernel Extended [TKR+02,SG02]

- Minor extensions, useful for binary classification:
- Estimate $\hat{\theta}_1$ and $\hat{\theta}_2$ for each class respectively,
- consider the score vector of the likelihood ratio

$$\phi_{\hat{\theta}_1,\hat{\theta}_2} : \mathbf{x} \mapsto \left( \frac{\partial \ln \frac{p_{\theta_1}(\mathbf{x})}{p_{\theta_2}(\mathbf{x})}}{\partial \vartheta} \bigg|_{\hat{\vartheta}=(\hat{\theta}_1,\hat{\theta}_2)} \right),$$

  where $\vartheta = (\theta_1, \theta_2)$ is in $\Theta^2$.

- Use this logratio's score vector to propose instead the kernel

$$(x, y) \mapsto \phi_{\hat{\theta}_1,\hat{\theta}_2}(\mathbf{x})^T \phi_{\hat{\theta}_1,\hat{\theta}_2}(\mathbf{y}).$$

# Mutual Information Kernel: densities as feature extractors

- More **bayesian** flavor $\rightarrow$ drops maximum-likelihood estimation of $\theta$. [See02]
- Instead, use **prior knowledge** on $\boldsymbol{\{p_\theta, \theta \in \Theta\}}$ through a **density** $\omega$ on $\Theta$
- Mutual information kernel $k_\omega$:

$$k_\omega(\mathbf{x}, \mathbf{y}) = \int_\Theta p_\theta(\mathbf{x}) p_\theta(\mathbf{y})\, \omega(d\theta).$$

- The feature maps $0 \leq p_\theta(\mathbf{x}) \leq 1$ and $0 \leq p_\theta(\mathbf{y}) \leq 1$.

> $k_\omega$ is big whenever many **common** densities $p_\theta$ score high probabilities for **both** $\mathbf{x}$ and $\mathbf{y}$

- Explicit computations sometimes possible, **namely conjugate priors**.
- Example: context-tree kernel for strings.

# Mutual Information Kernel & Fisher Kernels

The Fisher kernel is a maximum *a posteriori* approximation of the MI kernel.

- What? How? by setting the prior $\omega$ to the multivariate Gaussian density

$$\mathcal{N}(\hat{\theta}, J_{\hat{\theta}}^{-1}),$$

  an approximation known as Laplace's method,

- Writing

$$\Phi(x) = \nabla_{\hat{\theta}} \ln p_\theta(x) = \frac{\partial \ln p_\theta(x)}{\partial \theta}\Big|_{\hat{\theta}}$$

  we get

$$\log p_\theta(x) \approx \log p_{\hat{\theta}}(x) + \Phi(x)(\theta - \hat{\theta}).$$

# Mutual Information Kernel & Fisher Kernels

- Using $\mathcal{N}(\hat{\theta}, J_{\hat{\theta}}^{-1})$ for $\omega$ yields

$$
\begin{aligned}
k(x,y) &= \int_\Theta p_\theta(\mathbf{x}) p_\theta(\mathbf{y})\, \omega(d\theta),\\
&\approx C \int_\Theta e^{\log p_{\hat{\theta}}(x) + \Phi(x)^T(\theta-\hat{\theta})} e^{\log p_{\hat{\theta}}(y) + \Phi(y)^T(\theta-\hat{\theta})} \;\; e^{-(\theta-\hat{\theta})^T J_{\hat{\theta}}(\theta-\hat{\theta})} d\theta\\
&= C p_{\hat{\theta}}(x) p_{\hat{\theta}}(y) \int_\Theta e^{(\Phi(x)+\Phi(y))^T(\theta-\hat{\theta}) + (\theta-\hat{\theta})^T J_{\hat{\theta}}(\theta-\hat{\theta})} d\theta\\
&= C' p_{\hat{\theta}}(x) p_{\hat{\theta}}(y) e^{\frac{1}{2}(\Phi(x)+\Phi(y))^T J_{\hat{\theta}}^{-1}(\Phi(x)+\Phi(y))}
\end{aligned}
\tag{1}
$$

- the kernel

$$\tilde{k}(x,y) = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

is equal to the Fisher kernel in exponential form.

# Marginalized kernels - Graphs and Sequences

- Similar ideas: leverage **latent variable models.** [TKA02,KTI03]
- For **location** or **time-based** data,
  - the probability of emission of a token $x_i$ is conditioned by
  - an **unobserved** latent variable $s_i \in \mathcal{S}$, where $\mathcal{S}$ is a finite space of possible states.
- for observed sequences $\mathbf{x} = (x_1, \cdots, x_n), \mathbf{y} = (y_1, \cdots, y_n)$, sum over all possible state sequences the **weighted** product of **these probabilities**:

$$k(x, y) = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \boldsymbol{p(s|x)\, p(s'|y)\, \kappa\left((x, s), (y, s')\right)}$$

- closed form computations exist for graphs & sequences.

# Kernels on MLE parameters

- Use model directly to extract a single representation from observed points:

$$x \mapsto \hat{\theta}_x, \quad y \mapsto \hat{\theta}_y,$$

  through MLE for instance.

- compare $\mathbf{x}$ and $\mathbf{y}$ through a kernel $k_\Theta$ on $\Theta$,

$$k(x, y) = k_\Theta(\hat{\theta}_\mathbf{x}, \hat{\theta}_\mathbf{y}).$$

- Bhattacharrya affinities:

$$k_\beta(\mathbf{x}, \mathbf{y}) = \int_\mathcal{X} p_{\hat{\theta}_\mathbf{x}}(z)^\beta p_{\hat{\theta}_\mathbf{y}}(z)^\beta dz$$

  for $\beta > 0$.

# Semigroup Kernels : Building blocks

Loose algebraic structure: **Semigroups** [BCR84]

- Importance: unifying theory for many kernels, constructive perspective.
- a **semigroup** $(\mathcal{S},+)$ is a set $\mathcal{S}\neq\emptyset$ endowed with an ***associative composition*** $+$ with neutral element $0$.
- An **involutive semigroup** $(\mathcal{S},+,*)$ is endowed with an involution $*:\mathcal{S}\rightarrow\mathcal{S}$ such that $\forall x$ in $\mathcal{S},(x^*)^*=x$.
- Examples:
  - $\mathcal{S}$ is the set of strings, $+$ is the concatenation, $0$ is the empty string. $*$ is either the identity or the operation $ABCD\rightarrow DCBA$.
  - $\mathcal{S}$ is a group, and $*$ is the inverse. $e.g.$ $(\mathbb{R}^d,+,-)$
  - $\mathcal{S}$ is $\mathbb{R}^d_+$ with the $+$ operation and $*$ is the identity.
- We only consider ***abelian*** ($+$ is commutative) semigroups.

# Semigroup Kernels

- a **semigroup kernel** is a kernel $k$ defined as

$$k(x,y) \stackrel{\text{def}}{=} \varphi(x + y^*),$$

  where $\varphi : \mathcal{S} \mapsto \mathbb{R}$.

- $\rightarrow$ quantify similarity by looking only at $x + y^*$.
- Examples in $\mathbb{R}^d$,

$$k(x,y) = \varphi(x - y), \quad *(x) = -x,$$

  or

$$k(x,y) = \phi(x + y), \quad *(x) = x$$

- Example in $M_1(\mathbb{R}^d)$, the space of probability measures on $\mathbb{R}^d$,

$$k(\mu,\mu') \stackrel{\text{def}}{=} \frac{1}{\sqrt{\det \Sigma\left(\frac{\mu+\mu'}{2}\right)}},$$

# Semigroup Kernels and Semicharacters

- **Semicharacters**: real-valued function $\rho$ on an Abelian semigroup $(S,+)$ s.t.

(i) $\rho(0)=1$,
(ii) $\forall s,t\in\mathcal{S},\ \rho(s+t)=\rho(s)\overline{\rho(t)}$,
(iii) $\forall s\in\mathcal{S},\ \rho(s)=\overline{\rho(s^*)}$.

- For $(\mathbb{R}^+,+,\text{Id})$, semicharacters are exactly functions $s\rightarrow e^{\lambda s}$. indeed,
  - $e^{\lambda(s+t)}=e^{\lambda s}e^{\lambda t}$
- For $(\mathbb{R},+,-)$, semicharacters are exactly functions $s\rightarrow e^{i\lambda s}$. indeed,
  - $e^{i\lambda(s-t)}=e^{i\lambda s}e^{-i\lambda t}$, $e^{i\lambda s}=\overline{e^{-i\lambda s}}$.
- $\hat{S}$ is the **set of bounded semicharacters**.

| The building blocks of (bounded) semigroup kernels are semicharacters. |
|---|

# Semigroup Kernels and Semicharacters

- Proved in a fundamental theorem of Bochner [Boc33], generalized by [BCR84]:

  **Theorem 7** (Integral representation of p.d. functions). *A bounded function* $\varphi : S \to \mathbb{R}$ *is p.d. if and only if it there exists a non-negative measure* $\omega$ *on* $\hat{S}$ *such that:*
  $$\varphi(s) = \int_{\hat{S}} \rho(s)\, d\omega(\rho).$$
  *In that case the measure* $\omega$ *is unique.*

- ***Proof idea***
  - Semicharacters are **extreme rays** of the cone of positive definite kernels.
  - Choquet's theory helps us prove that any point in that cone is a convex combination of extreme rays (a barycentre)

# Bochner Theorems in $(\mathbb{R}^d, +, -)$ and $(\mathbb{R}_+^d, +, \mathbf{Id})$

- $* = -$: $\exists!$ non-negative measure $\omega$ on $\mathbb{R}^d$ s.t.

$$\varphi(x) = \int_{\mathbb{R}^d} e^{ix^T r} d\omega(r);$$

$\varphi$ is the Fourier transform of a non-negative measure $\omega$ on $\mathbb{R}^d$.

  - Kernels of the type $k(x, y) = \varphi(x - y)$ also known as ***Radial Basis Functions*** have such a decomposition.

- $* = \mathbf{Id}$: Suppose $k$ is bounded & s.t. $k(x, y) = \psi(x + y)$. $\exists!$ non-negative measure $\omega$ on $\mathbb{R}^d$ s.t.

$$\psi(x) = \int_{\mathbb{R}^d} e^{-x^T r} d\omega(r);$$

$\psi$ is the Laplace transform of a non-negative measure $\omega$ on $\mathbb{R}^d$.