# Different views on Reproducing Kernel Hilbert Spaces

## Regularization Methods for High Dimensional Learning

**Francesca Odone** and **Lorenzo Rosasco**
odone@disi.unige.it - lrosasco@mit.edu

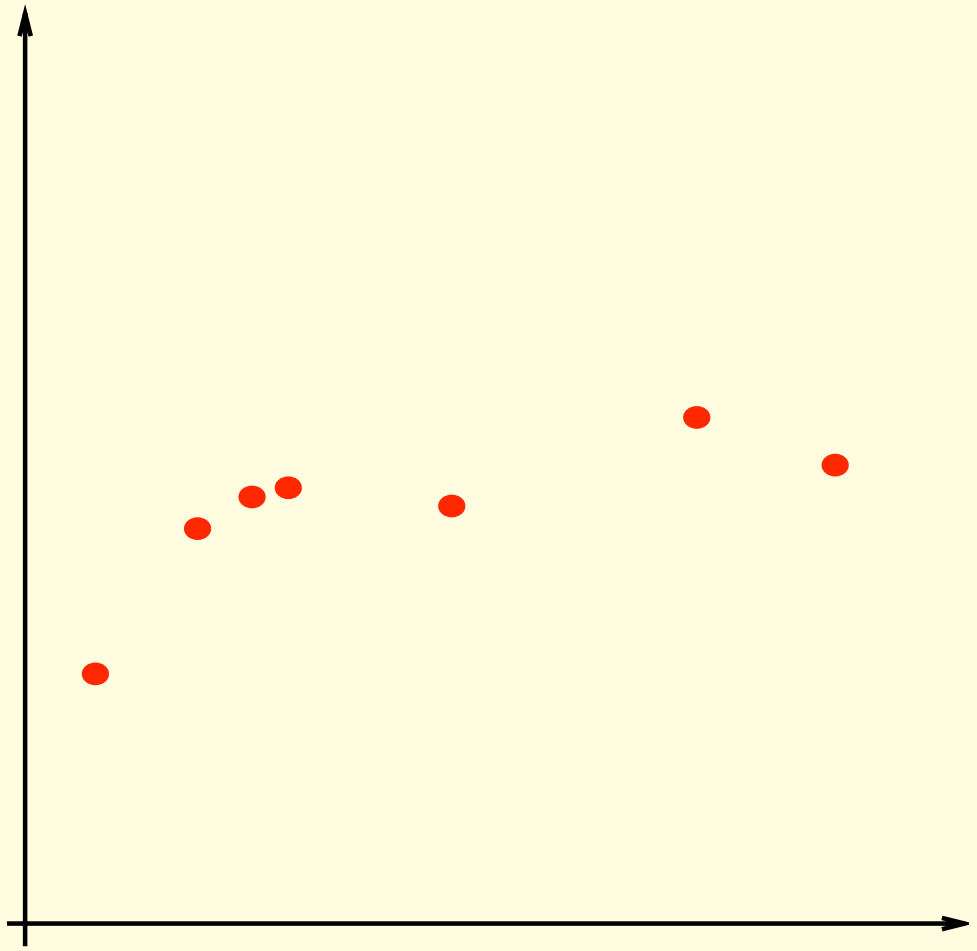
March 13, 2012

# About this class

**Goal** To introduce Reproducing Kernel Hilbert Spaces (RKHS) from different perspectives and to derive the general solution of Tikhonov regularization in RKHS.

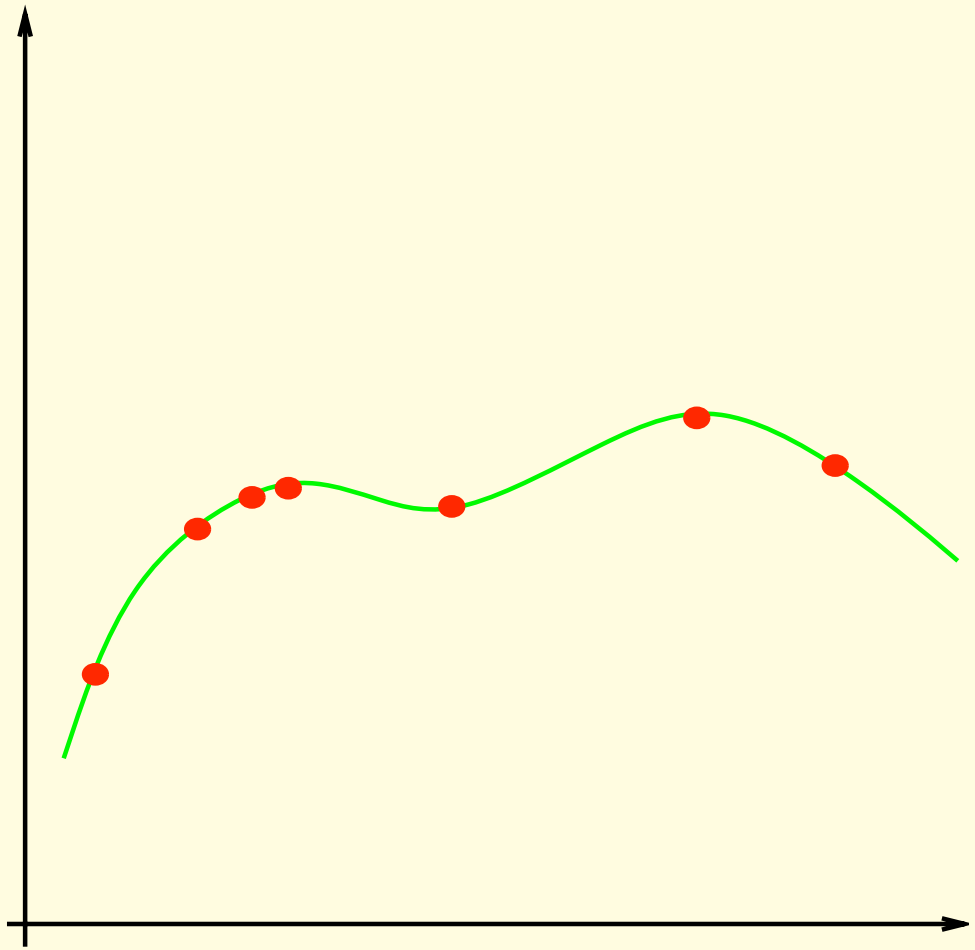
# Function Approximation from Samples

Here is a graphical example for generalization: given a certain number of samples...



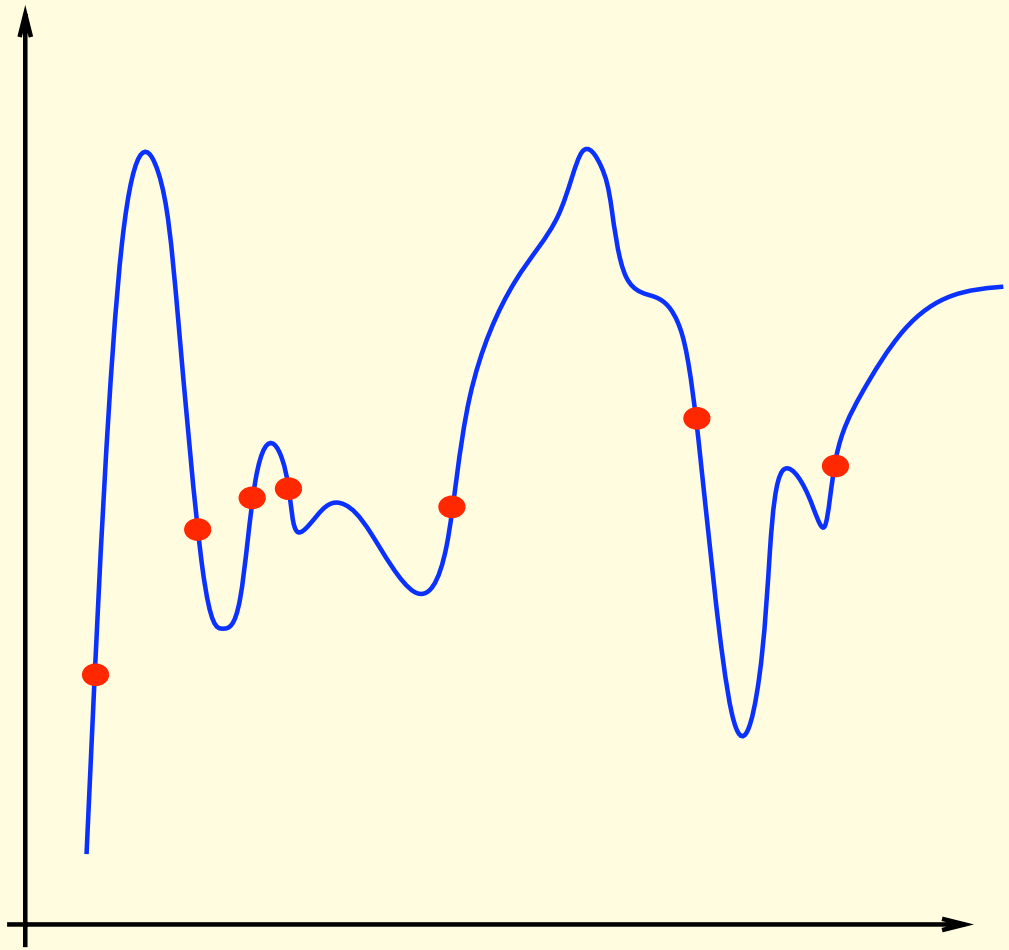
# Function Approximation from Samples (cont.)

Suppose this is the “true” solution...



# The Problem is Ill-Posed

... but suppose ERM gives this solution!



# Regularization

The basic idea of regularization (originally introduced independently of the learning problem) is to restore well-posedness of ERM by constraining the hypothesis space $\mathcal{H}$.

## Regularization

A possible way to do this is considering *regularized* empirical risk minimization, that is we look for solutions minimizing a two term functional

$$\underbrace{ERR(f)}_{\text{empirical error}} + \lambda \underbrace{R(f)}_{\text{regularizer}}$$

the regularization parameter $\lambda$ trade-offs the two terms.

# Tikhonov Regularization

Tikhonov regularization amounts to minimize

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \mathcal{R}(f) \qquad \lambda > 0 \tag{1}$$

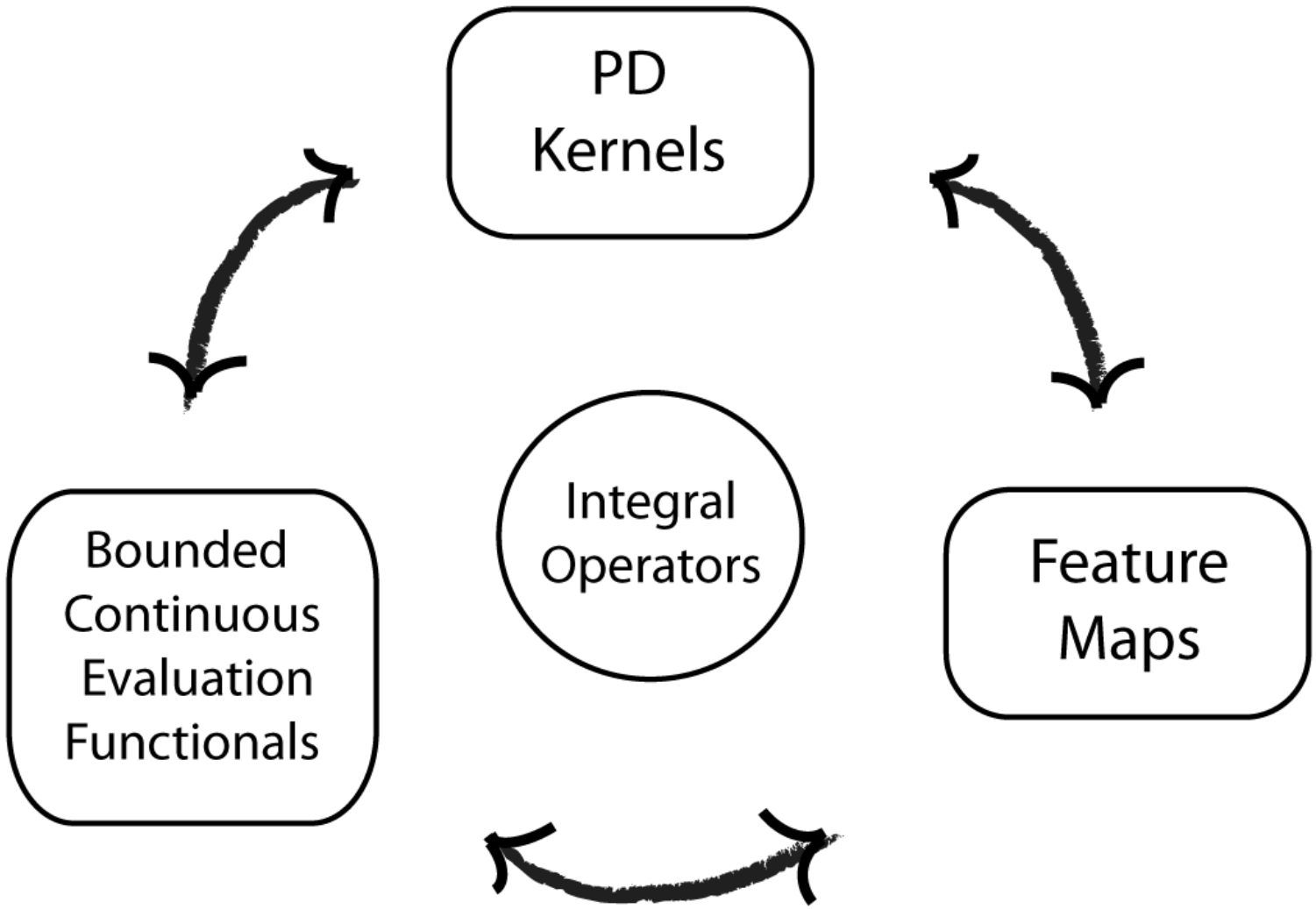
- $V(f(x), y)$ is the loss function, that is the price we pay when we predict $f(x)$ in place of $y$
- $\mathcal{R}(f)$ is a regularizer– often $\mathcal{R}(f) = \|\cdot\|_{\mathcal{H}}$, the norm in the *function space* $\mathcal{H}$

The regularizer should encode some notion of smoothness of $f$, choosing different loss functions $V(f(x), y)$ we can recover different algorithms.

# The "Ingredients" of Tikhonov Regularization

- The scheme we just described is very general and by choosing different loss functions $V(f(x), y)$ we can recover different algorithms
- The main point we want to discuss is how to choose a norm encoding some notion of smoothness/complexity of the solution
- Reproducing Kernel Hilbert Spaces allow us to do this in a very powerful way

# Different Views on RKHS



# Part I: Evaluation Functionals

A **function space** $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \to \mathbb{R}$.

A **function space** $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \to \mathbb{R}$.
A **norm** is a nonnegative function $\|\cdot\|$ such that $\forall f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ *iff* $f = 0$;
2. $\|f + g\| \leq \|f\| + \|g\|$;
3. $\|\alpha f\| = |\alpha| \, \|f\|$.

A **function space** $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \to \mathbb{R}$.
A **norm** is a nonnegative function $\|\cdot\|$ such that $\forall f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ *iff* $f = 0$;
2. $\|f + g\| \leq \|f\| + \|g\|$;
3. $\|\alpha f\| = |\alpha| \, \|f\|$.

A norm can be defined via a **inner product** $\|f\| = \sqrt{\langle f, f \rangle}$.

# Some Functional Analysis

A **function space** $\mathcal{F}$ is a space whose elements are functions $f$, for example $f : \mathbb{R}^d \to \mathbb{R}$.
A **norm** is a nonnegative function $\|\cdot\|$ such that $\forall f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ *iff* $f = 0$;
2. $\|f + g\| \leq \|f\| + \|g\|$;
3. $\|\alpha f\| = |\alpha| \, \|f\|$.

A norm can be defined via a **inner product** $\|f\| = \sqrt{\langle f, f \rangle}$.
A **Hilbert space** is a complete inner product space.

# Spaces of Functions: Examples

- Continuous functions $C[a, b]$ :
  a norm can be established by defining

$$\|f\| = \max_{a \leq x \leq b} |f(x)|$$

  (not a Hilbert space!)

# Spaces of Functions: Examples

- Continuous functions $C[a, b]$ :
  a norm can be established by defining

$$\|f\| = \max_{a \leq x \leq b} |f(x)|$$

  (not a Hilbert space!)
- Square integrable functions $L_2[a, b]$:
  it is a Hilbert space where the norm is induced by the dot product

$$\langle f, g \rangle = \int_a^b f(x)g(x)dx$$

An evaluation functional over the *Hilbert space of functions* $\mathcal{H}$ is a linear functional $\mathcal{F}_t : \mathcal{H} \to \mathbb{R}$ that *evaluates* each function in the space at the point $t$, or

$$\mathcal{F}_t[f] = f(t).$$

# RKHS

An evaluation functional over the *Hilbert space of functions* $\mathcal{H}$ is a linear functional $\mathcal{F}_t : \mathcal{H} \to \mathbb{R}$ that *evaluates* each function in the space at the point $t$, or

$$\mathcal{F}_t[f] = f(t).$$

## Definition

A Hilbert space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if the evaluation functionals are bounded and continuous, i.e. if there exists a $M$ s.t.

$$|\mathcal{F}_t[f]| = |f(t)| \leq M\|f\|_{\mathcal{H}} \ \ \forall f \in \mathcal{H}$$

# Evaluation functionals

Evaluation functionals are not always bounded.
Consider $L_2[a, b]$:

- Each element of the space is an equivalence class of functions with the same integral $\int |f(x)|^2 dx$.
- An integral remains the same if we change the function in a countable set of points.

# Linear RKHS

Our function space is 1-dimensional lines

$$f(x) = w\,x$$

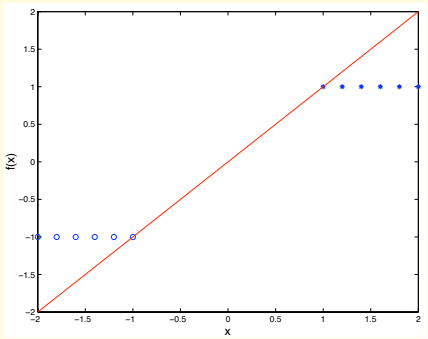
where the RKHS norm is simply

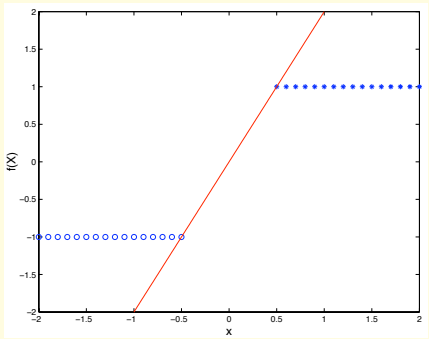
$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = w^2$$

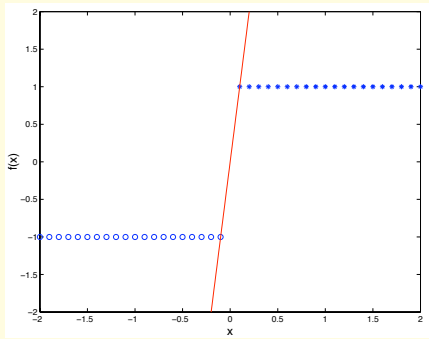
so that our measure of complexity is the slope of the line.
We want to separate two classes using lines and see how the magnitude of the slope corresponds to a measure of complexity.
We will look at three examples and see that each example requires more ”complicated functions, functions with greater slopes, to separate the positive examples from negative examples.

# Linear case (cont.)

here are three datasets: a linear function should be used to separate the classes. Notice that as the class distinction becomes finer, a larger slope is required to separate the classes.



# Part II: Kernels

# Reproducing kernel (RK)

- If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists a function $K_t$ in $\mathcal{H}$ (called *representer*) with the **reproducing** property

$$\mathcal{F}_t[f] = \langle K_t, f \rangle_{\mathcal{H}} = f(t).$$

# Reproducing kernel (RK)

- If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists a function $K_t$ in $\mathcal{H}$ (called *representer*) with the **reproducing** property

$$\mathcal{F}_t[f] = \langle K_t, f \rangle_{\mathcal{H}} = f(t).$$

- Since $K_t$ is a function in $\mathcal{H}$, by the reproducing property, for each $x \in X$

$$K_t(x) = \langle K_t, K_x \rangle_{\mathcal{H}}$$

The *reproducing kernel* (rk) of $\mathcal{H}$ is

$$K(t, x) := K_t(x)$$

# Positive definite kernels

Let $X$ be some set, for example a subset of $\mathbb{R}^d$ or $\mathbb{R}^d$ itself. A *kernel* is a symmetric function $K : X \times X \to \mathbb{R}$.

## Definition

A kernel $K(t, s)$ is *positive definite (pd)* if

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, ..., t_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

The following theorem relates pd kernels and RKHS

**THEOREM**

a) For every RKHS there exist an associated reproducing kernel which is symmetric and positive definite

b) Conversely every symmetric, positive definite kernel $K$ on $X \times X$ defines a unique RKHS on $X$ with $K$ as its reproducing kernel

# SKETCH OF PROOF

**a)** We must prove that the rk $K(t,x) = \langle K_t, K_x \rangle_{\mathcal{H}}$ is *symmetric* and *pd*.

- Symmetry follows from the symmetry property of dot products

$$\langle K_t, K_x \rangle_{\mathcal{H}} = \langle K_x, K_t \rangle_{\mathcal{H}}$$

- $K$ is pd because

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) = \sum_{i,j=1}^{n} c_i c_j \langle K_{t_i}, K_{t_j} \rangle_{\mathcal{H}} = || \sum c_j K_{t_j} ||^2_{\mathcal{H}} \geq 0.$$

# Sketch of proof (cont.)

**b)** Conversely, given $K$ one can construct the RKHS $\mathcal{H}$ as the *completion* of the space of functions spanned by the set $\{K_x | x \in X\}$ with a inner product defined as follows.
The dot product of two functions $f$ and $g$ in $\mathrm{span}\{K_x | x \in X\}$

$$\begin{aligned} f(x) &= \sum_{i=1}^{s} \alpha_i K_{x_i}(x) \\ g(x) &= \sum_{i=1}^{s'} \beta_i K_{x'_i}(x) \end{aligned}$$

is *by definition*

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{s} \sum_{j=1}^{s'} \alpha_i \beta_j K(x_i, x'_j).$$

# EXAMPLES OF PD KERNELS

Very common examples of symmetric pd kernels are

- **Linear kernel**

$$K(x, x') = x \cdot x'$$

- **Gaussian kernel**

$$K(x, x') = \mathrm{e}^{-\frac{\|x-x'\|^2}{\sigma^2}}, \qquad \sigma > 0$$

- **Polynomial kernel**

$$K(x, x') = (x \cdot x' + 1)^d, \qquad d \in \mathbb{N}$$

For specific applications, designing an effective kernel is a challenging problem.

# Examples of PD kernels

- Kernel are a very general concept. We can have kernel on vectors, string, matrices, graphs, probabilities...
- Combinations of Kernels allow to do integrate different kinds of data.
- Often times Kernel are views and designed to be similarity measure (in this case it make sense to have normalized kernels)

$$d(x, x')^2 = \|K_x - K'_x\|^2 = 2(1 - K(x, x')).$$

# Part III: Feature Map

# Feature Maps and Kernels

Kernels can be seen as inner products.

- Let $\Phi(x) = K_x$. Then $\Phi : X \to \mathcal{H}$.

# Feature Maps and Kernels

Kernels can be seen as inner products.

- Let $\Phi(x) = K_x$. Then $\Phi : X \to \mathcal{H}$.
- Let $\Phi(x) = (\psi_j(x))_j$, where $(\psi_j(x))_j$ is an orthonormal basis of $\mathcal{H}$. Then $\Phi : X \to \ell^2$.

# Mercer Theorem and Feature Map

A well know example comes from a result due to Mercer.

## Mercer Theorem

The operator

$$L_K f(x) = \int_X K(x,s) f(s) p(s) dx$$

is symmetric positive compact with eigenvalues/functions $(\sigma_I, \phi)_i$.
It can be shown that

$$K(x,s) = \sum_{i \geq 1} \sigma_i \phi_i(x) \phi_i(s).$$

Then, let $\Phi(x) = (\sqrt{\sigma_i}\phi_i(x))_i$, $\Phi : X \to \ell^2$ so that by the above result

$$K(x,s) = \langle \Phi(x), \Phi(x) \rangle .$$

# Feature Map and Feature Space

In general a feature map is a map $\Phi : X \to \mathcal{F}$, where $\mathcal{F}$ is a Hilbert space and is called Feature Space.
Every feature map defines a kernel via

$$K(x, s) = \langle \Phi(x), \Phi(x) \rangle .$$

# Kernel from Feature Maps

Often times, feature map, and hence kernels, are defined through a dictionary of features

$$\mathcal{D} = \{\phi_j,\ i = 1, \ldots, p \mid \phi_j : X \to \mathbb{R},\ \ \forall j\}$$

where $p \leq \infty$.

# Kernel from Feature Maps

Often times, feature map, and hence kernels, are defined through a dictionary of features

$$\mathcal{D} = \{\phi_j,\ i = 1, \ldots, p \mid \phi_j : X \to \mathbb{R},\ \ \forall j\}$$

where $p \leq \infty$.
We can interpret the above functions as (possibly non linear) *measurements* on the inputs.

# Function as Hyperplanes in the Feature Space

The concept of feature map allows to give a new interpretation of RKHS.

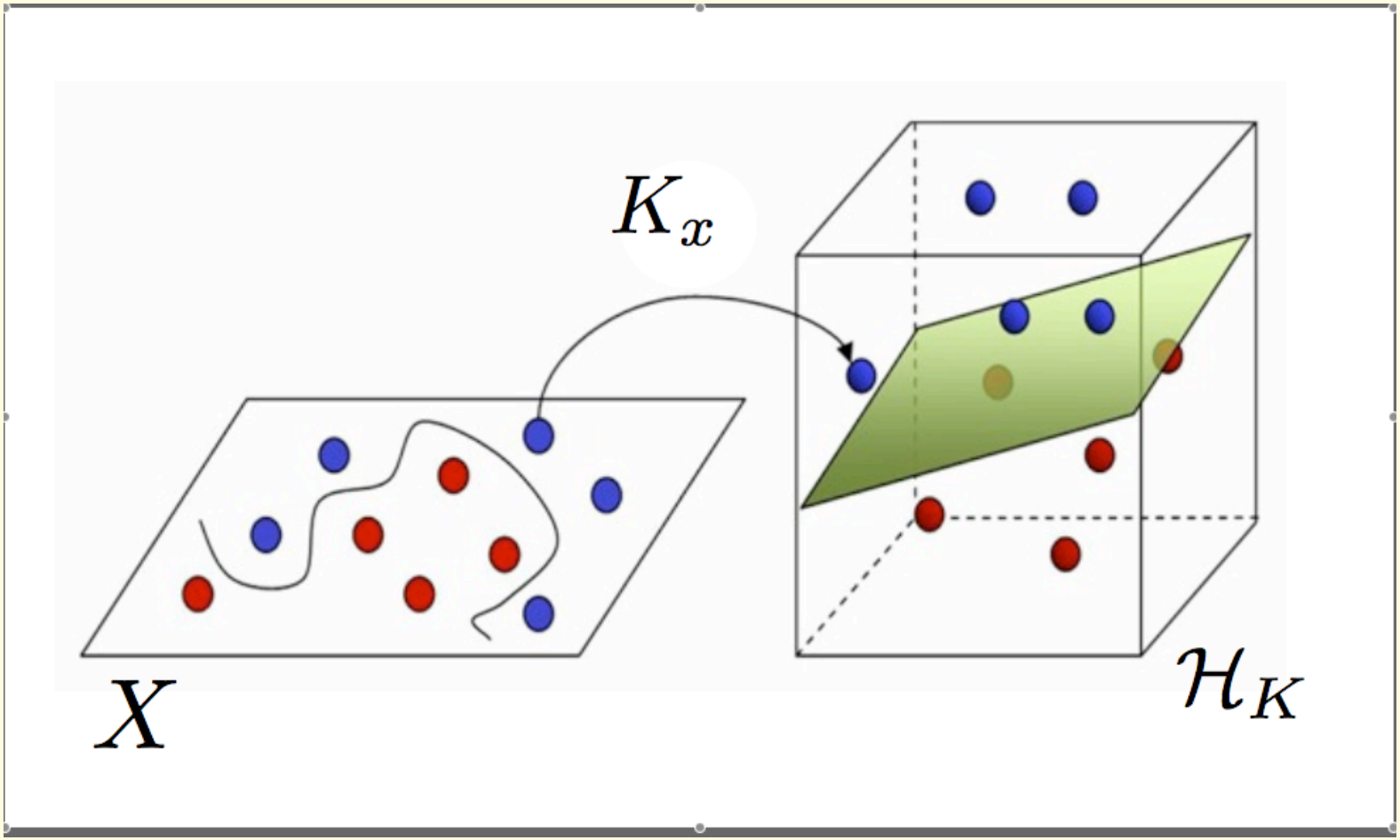
Functions can be seen as hyperplanes,

$$f(x) = \langle w, \Phi(x) \rangle .$$

This can be seen for any of the previous examples.

- Let $\Phi(x) = (\sqrt{\sigma_j}\phi_j(x))_j$.
- Let $\Phi(x) = K_x$.
- Let $\Phi(x) = (\psi_j(x))_j$.

# Feature Maps Illustrated



# Kernel "Trick" and Kernelization

Any algorithm which works in a euclidean space, hence requiring only inner products in the computations, can be *kernelized*

$$K(x, s) = \langle \Phi(x), \Phi(x) \rangle .$$

- Kernel PCA.
- Kernel ICA.
- Kernel CCA.
- Kernel LDA.
- Kernel...

# Part IV: Regularization Networks and Representer Theorem

# Again Tikhonov Regularization

The algorithms (*Regularization Networks*) that we want to study are defined by an optimization problem over RKHS,

$$f_S^\lambda = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

where the *regularization parameter* $\lambda$ is a positive number, $\mathcal{H}$ is the RKHS as defined by the *pd kernel* $K(\cdot, \cdot)$, and $V(\cdot, \cdot)$ is a **loss function**.
Note that $\mathcal{H}$ is possibly infinite dimensional!

# Existence and uniqueness of minimum

If the positive loss function $V(\cdot,\cdot)$ is convex with respect to its first entry, the functional

$$\Phi[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

is *strictly convex* and *coercive*, hence it has exactly one local (global) minimum.
Both the squared loss and the hinge loss are convex.
On the contrary the 0-1 loss

$$V = \Theta(-f(x)y),$$

where $\Theta(\cdot)$ is the Heaviside step function, is **not** convex.

# The Representer Theorem

## An important result

The minimizer over the RKHS $\mathcal{H}$, $f_S$, of the regularized empirical functional

$$I_S[f] + \lambda \|f\|_{\mathcal{H}}^2,$$

can be represented by the expression

$$f_S^{\lambda}(x) = \sum_{i=1}^{n} c_i K(x_i, x),$$

for some $n$-tuple $(c_1, \ldots, c_n) \in \mathbb{R}^n$.
Hence, minimizing over the (possibly infinite dimensional) Hilbert space, *boils down to minimizing over* $\mathbb{R}^n$.

# Sketch of proof

Define the linear subspace of $\mathcal{H}$,

$$\mathcal{H}_0 = \text{span}(\{K_{x_i}\}_{i=1,\ldots,n})$$

Let $\mathcal{H}_0^\perp$ be the linear subspace of $\mathcal{H}$,

$$\mathcal{H}_0^\perp = \{f \in \mathcal{H} | f(x_i) = 0,\ i = 1, \ldots, n\}.$$

From the reproducing property of $\mathcal{H}$, $\forall f \in \mathcal{H}_0^\perp$

$$\langle f, \sum_i c_i K_{x_i} \rangle_{\mathcal{H}} = \sum_i c_i \langle f, K_{x_i} \rangle_{\mathcal{H}} = \sum_i c_i f(x_i) = 0.$$

$\mathcal{H}_0^\perp$ is the orthogonal complement of $\mathcal{H}_0$.

# Sketch of proof (cont.)

Every $f \in \mathcal{H}$ can be uniquely decomposed in components along and perpendicular to $\mathcal{H}_0$: $f = f_0 + f_0^{\perp}$.
Since by orthogonality

$$\|f_0 + f_0^{\perp}\|^2 = \|f_0\|^2 + \|f_0^{\perp}\|^2,$$

and by the reproducing property

$$I_S[f_0 + f_0^{\perp}] = I_S[f_0],$$

then

$$I_S[f_0] + \lambda \|f_0\|_{\mathcal{H}}^2 \leq I_S[f_0 + f_0^{\perp}] + \lambda \|f_0 + f_0^{\perp}\|_{\mathcal{H}}^2.$$

Hence the minimum $f_S^{\lambda} = f_0$ *must belong to the linear space* $\mathcal{H}_0$.

# Norms in RKHS and Smoothness

Choosing different kernels one can show that the norm in the corresponding RKHS encodes different notions of smoothness.

# Norms in RKHS and Smoothness

Choosing different kernels one can show that the norm in the corresponding RKHS encodes different notions of smoothness.

- Band limited functions. Consider the set of functions

$$\mathcal{H} := \{f \in L^2(\mathbb{R}) \mid F(\omega) \in [-a, a], a < \infty\}$$

with the usual $L^2$ inner product. the function at every point is given by the convolution with a sinc function $sin(ax)/ax$. The norm

$$\|f\|_{\mathcal{H}}^2 = \int f(x)^2 dx = \int_a^a |F(\omega)|^2 d\omega$$

Where $F(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\, dt$ is the Fourier tranform of $f$.

# Norms in RKHS and Smoothness

- Sobolev Space: consider $f : [0, 1] \to \mathbb{R}$ with $f(0) = f(1) = 0$. The norm

$$\|f\|_{\mathcal{H}}^2 = \int (f'(x))^2 dx = \int \omega^2 |F(\omega)|^2 d\omega$$

# Norms in RKHS and Smoothness

- Sobolev Space: consider $f : [0, 1] \to \mathbb{R}$ with $f(0) = f(1) = 0$. The norm
$$\|f\|_{\mathcal{H}}^2 = \int (f'(x))^2 dx = \int \omega^2 |F(\omega)|^2 d\omega$$
- Gaussian Space: the norm can be written as
$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi^d} \int |F(\omega)|^2 exp^{\frac{\sigma^2\omega^2}{2}} d\omega$$

# Historical Remarks

- RKHS were explicitly introduced in learning theory by Girosi (1997).
- Poggio and Girosi (1989) introduced Tikhonov regularization in learning theory and worked with RKHS only implicitly, because they dealt mainly with hypothesis spaces on unbounded domains, which we will not discuss here.
- RKHS were used much earlier in approximation theory (eg Wahba, 1990...) and computer vision (eg Bertero, Torre, Poggio, 1988...).

# References


- Aronszajn. *Theory of reproducing kernels.* Transactions of the American Mathematical Society, 686, 337-404, 1950.
- Cucker and Smale. *On the mathematical foundations of learning.* Bulletin of the American Mathematical Society, 2002.
- Evgeniou, Pontil and Poggio. *Regularization Networks and Support Vector Machines* Advances in Computational Mathematics, 2000.
- Wahba, G. *Spline Models for Observational Data* Series in Applied Mathematics, Vol. 59, SIAM, 1990. (Chapter 1)