KERNELS, DICTIONARY AND REGULARIZATION

Francesca Odone and **Lorenzo Rosasco**

RegML 2013

About this class

Goal To introduce a useful family of hypothesis spaces called Reproducing Kernel Hilbert Spaces (RKHS);
To discuss how to design regularization via RKHS;
To show how to solve computationally nonparametric learning models.

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

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

the regularization parameter λ trade-offs the two terms.

Tikhonov regularization amounts to minimize

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

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

The regularizer should encode some notion of smoothness of f.

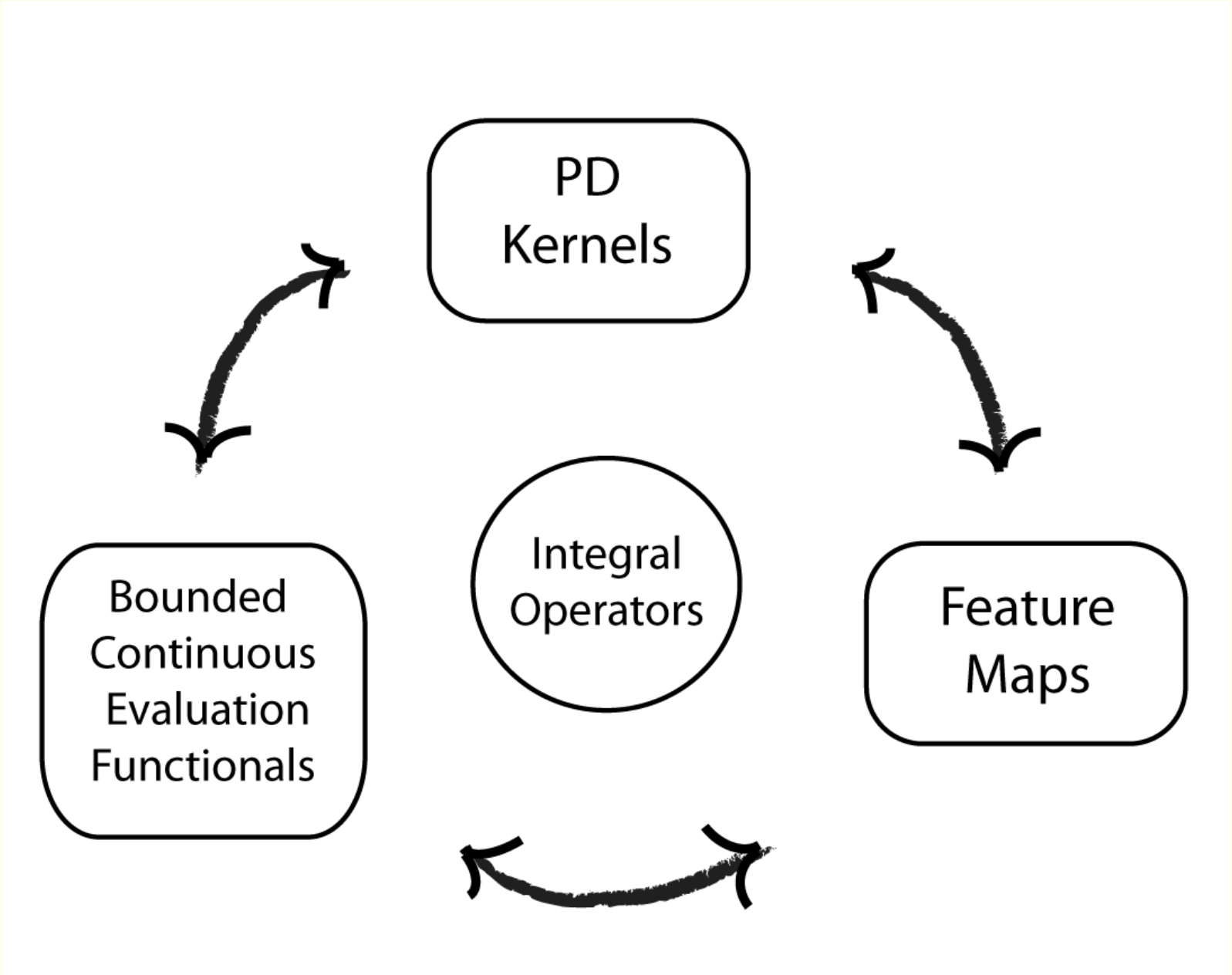
The "Ingredients" of Tikhonov Regularization

- 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

Plan

- Part I: Reproducing Kernels
- Part II: Feature Maps
- Part II: Representer Theorem

Different Views on RKHS



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** $\|f\| = \sqrt{\langle f, f \rangle}$.

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.

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!)

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

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

Hypothesis Space: Desiderata

- Hilbert Space.
- Point-wise defined functions.

Plan

- Part I: Reproducing Kernels
- Part II: Feature Maps
- Part III: Representer Theorem

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 $c_1, ..., c_n \in \mathbb{R}$.

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)

$$d(x, x')^2 \sim 2(1 - K(x, x')).$$

Examples of PD kernels

- Anova Kernels
- Diffusion Kernels
- String Kernels
- p-spectrum kernels
- All-subsequences kernels
- P Kernel
- Fisher Kernel
- Marginal Kernel
- Histogram Intersection Kernel
- ...

Building a Hypotheses Space from Kernels

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 suitable inner product (here $K_x(\cdot) = K(x, \cdot)$).

Building a Hypotheses Space from Kernels

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 suitable inner product (here $K_x(\cdot) = K(x, \cdot)$).
The dot product of two functions f and g in $\text{span}\{K_x | x \in X\}$

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

is *by definition*

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

Reproducing Property

By construction we have that

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

where we use the notation $K_x(\cdot) = K(x, \cdot)$.

As a consequence

$$\sup_{x \in X} |f(x)| \leq \sup_{x \in X} \sqrt{K(x,x)} \, \|f\|_{\mathcal{H}} .$$

Norms in RKHS and Smoothness

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

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

- Paley-Wiener Space: 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 norm

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

The kernel is $K(x, x') = sin(a(x - x'))/a(x - x')$.

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

- Sobolev Space: consider $f : \mathbb{R} \to \mathbb{R}$. The norm

$$\|f\|_{\mathcal{H}}^2 = \int (f(x))^2 dx + \int (f'(x))^2 dx = \int (\omega^2 + 1)|F(\omega)|^2 d\omega$$

The kernel is $K(x, x') = e^{-|x-x'|}$.

- Sobolev Space: consider $f : \mathbb{R} \to \mathbb{R}$. The norm

$$\|f\|_{\mathcal{H}}^2 = \int (f(x))^2 dx + \int (f'(x))^2 dx = \int (\omega^2 + 1)|F(\omega)|^2 d\omega$$

  The kernel is $K(x, x') = e^{-|x-x'|}$.
- Gaussian Kernel: the norm can be written as

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi d} \int |F(\omega)|^2 exp^{\frac{\sigma^2\omega^2}{2}} d\omega.$$

  The kernel is $K(x, x') = e^{-|x-x'|^2}$.

- Sobolev Space: consider $f : \mathbb{R} \to \mathbb{R}$. The norm

$$\|f\|_{\mathcal{H}}^2 = \int (f(x))^2 dx + \int (f'(x))^2 dx = \int (\omega^2 + 1)|F(\omega)|^2 d\omega$$

  The kernel is $K(x, x') = e^{-|x-x'|}$.
- Gaussian Kernel: the norm can be written as

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi d} \int |F(\omega)|^2 exp^{\frac{\sigma^2\omega^2}{2}} d\omega.$$

  The kernel is $K(x, x') = e^{-|x-x'|^2}$.

Linear RKHS

Our function space is 1-dimensional lines

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

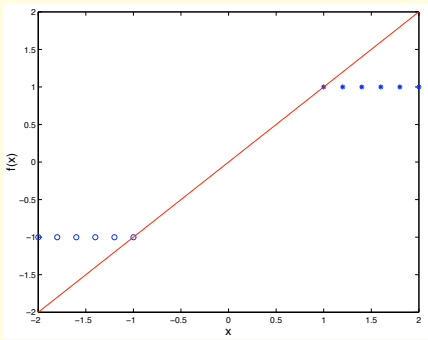
where the RKHS norm is simply

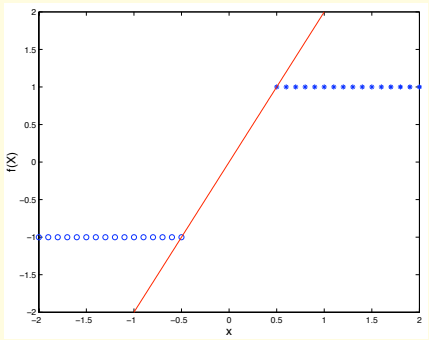
$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = w^2$$

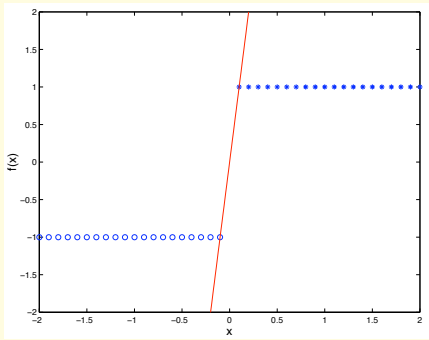
so that our measure of complexity is the slope of the line.
We want to separate two classes using lines and see how the magnitude of the slope corresponds to a measure of complexity.
We will look at three examples and see that each example requires more "complicated functions, functions with greater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







Plan

- Part I: Reproducing Kernels
- Part II: Feature Maps
- Part III: Representer Theorem

Feature Map and Feature Space

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

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

From Kernels to Feature Maps

We can associate one (in fact many!) feature map to every kernel.

- 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 $\Phi : X \to \ell^2$.
  Why?

From Feature Maps to Kernels

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 ϕ's as (possibly non linear) *measurements* on the inputs.

From Feature Maps to Kernels

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 ϕ's as (possibly non linear) *measurements* on the inputs.

$$K(x, x') = \sum_{j=1}^{p} \phi_j(x)\phi_j(x')$$

From Feature Maps to Kernels

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 ϕ's as (possibly non linear) *measurements* on the inputs.

$$K(x, x') = \sum_{j=1}^{p} \phi_j(x)\phi_j(x')$$

- If $p < \infty$ we can always define a feature map.
- If $p = \infty$ we need extra assumptions.
  Which ones?

From Feature Map to RKHS

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

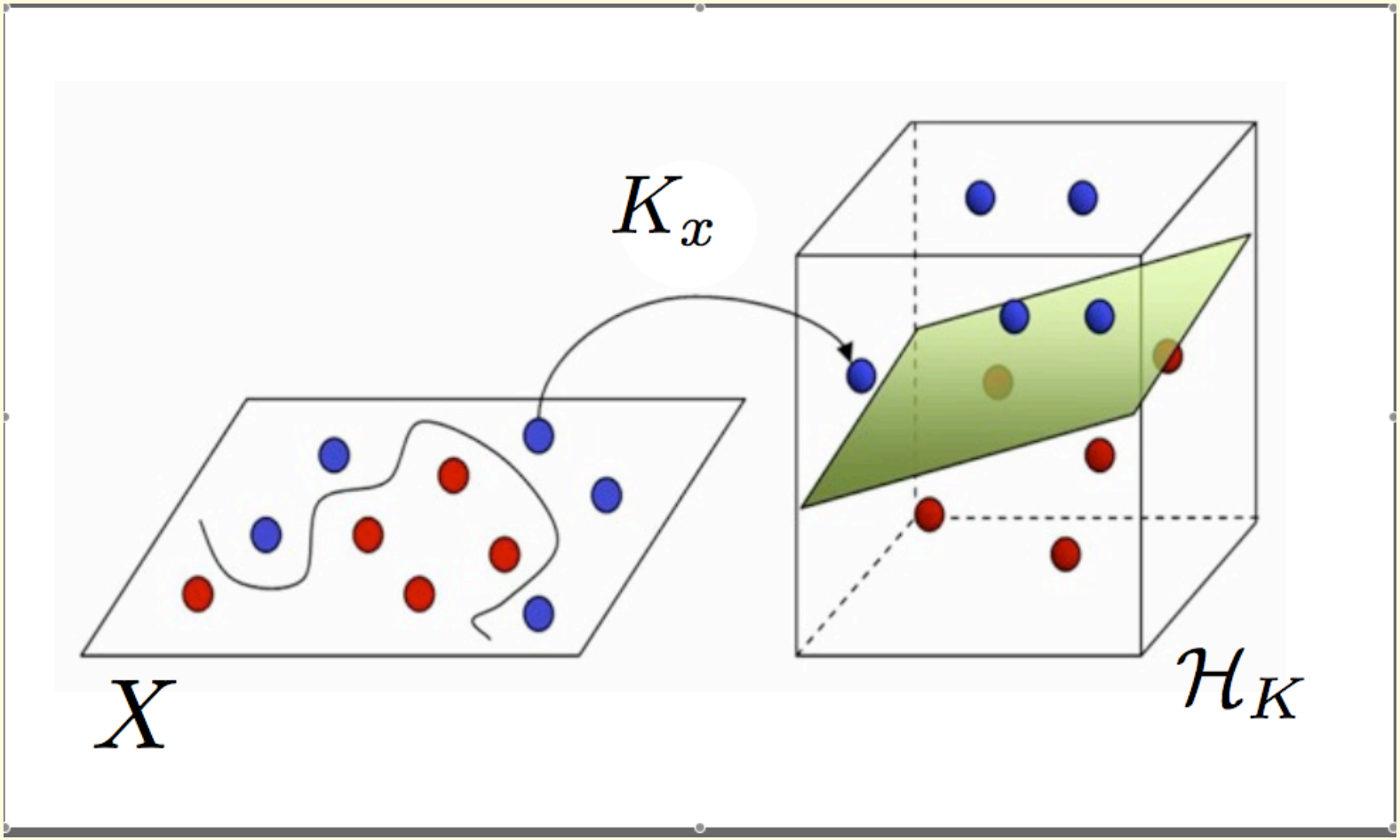
Functions can be seen as hyperplanes,

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

This can be seen for any of the previous examples.

- 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*

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

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

Part III: 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* λ 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!

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.

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* $\mathcal{H}_0$.

Common Loss Functions

The following two important learning techniques are implemented by different choices for the loss function $V(\cdot, \cdot)$

- **Regularized least squares** (RLS)

$$V = (y - f(x))^2$$

- **Support vector machines for classification** (SVMC)

$$V = |1 - yf(x)|_+$$

where

$$(k)_+ \equiv \max(k, 0).$$

Tikhonov Regularization for RLS and SVMs

In the next two classes we will study Tikhonov regularization with different loss functions for both regression and classification. We will start with the square loss and then consider SVM loss functions.