The Learning Problem and Regularization

9.520 Class 03, 12 February 2003

Tomaso Poggio and Sayan Mukherjee

About this class

Goal To introduce a particularly useful family of hypothesis spaces called Reproducing Kernel Hilbert Spaces (RKHS) and to derive the general solution of Tikhonov regularization in RKHS.

Function space

A **function space** is a space made of functions. Each function in the space can be thought of as a point. Examples:

1. $C[a, b]$, the set of all real-valued *continuous* functions in the interval $[a, b]$;

2. $L_1[a, b]$, the set of all real-valued functions whose absolute value is integrable in the interval $[a, b]$;

3. $L_2[a, b]$, the set of all real-valued functions square integrable in the interval $[a, b]$

Note that the functions in 2 and 3 are not necessarily continuous.

Normed space

A **normed** space is a linear (vector) space N in which a norm is defined. A nonnegative function $\|\cdot\|$ is a norm *iff* $\forall f, g \in N$ and $\alpha \in \mathbb{R}$

1. $\|f\| \geq 0$ and $\|f\| = 0$ *iff* $f = 0$;

2. $\|f + g\| \leq \|f\| + \|g\|$;

3. $\|\alpha f\| = |\alpha| \; \|f\|$.

Note, if all conditions are satisfied except $\|f\| = 0$ *iff* $f = 0$ then the space has a seminorm instead of a norm.

Euclidean space

A **Euclidean** space is a linear (vector) space E in which a dot product is defined. A real valued function $\langle \cdot, \cdot \rangle$ is a dot product *iff* $\forall f, g, h \in E$ and $\alpha \in \mathbb{R}$

1. $\langle f, g \rangle = \langle g, f \rangle$;
2. $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$ and $\langle \alpha f, g \rangle = \alpha \langle f, g \rangle$;
3. $\langle f, f \rangle \geq 0$ and $\langle f, f \rangle = 0$ *iff* $f = 0$.

In a Euclidean space we can speak of the *angle* between vectors and the norm of a vector is $\sqrt{\langle f, f \rangle}$.

Dense

Let A and B be subspaces of some normed (metric) space R. A is said to be **dense** in B *iff* $A \subset B$ and $B \subset \bar{A}$.

Example: the set of all rational points is dense in the real line.

Note: a hypothesis space that is dense in L_2 is a desirable property of many any approximation schemes.

Hilbert space

A **Hilbert space** is a Euclidean space that is *complete*, *separable*, and generally *infinite dimensional*.

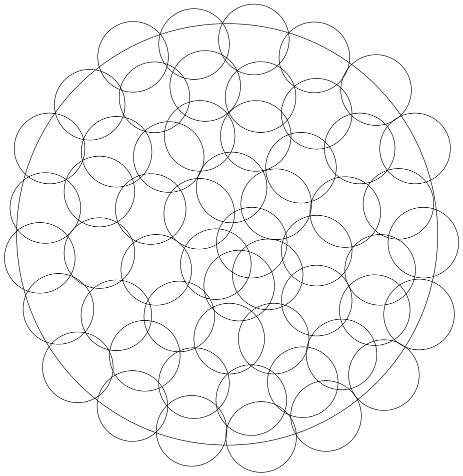
- A *Hilbert space* has a countable basis.

L_2 and $\mathbb{R}^n$ are examples of Hilbert spaces.

Compact spaces

A normed space is **compact** *iff* it is *totally bounded* and *complete*.

- A *compact space*, since it is totally bounded, has a **finite** ε-net for any $\varepsilon > 0$.



The large ball contains all functions in the space, for each function, f_i, f_j in a small ball, $\|f_i - f_j\|_\infty < \varepsilon$.

Evaluation functionals

A linear evaluation functional is a linear functional $\mathcal{F}_t$ that *evaluates* each function in the space at the point t, or

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

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

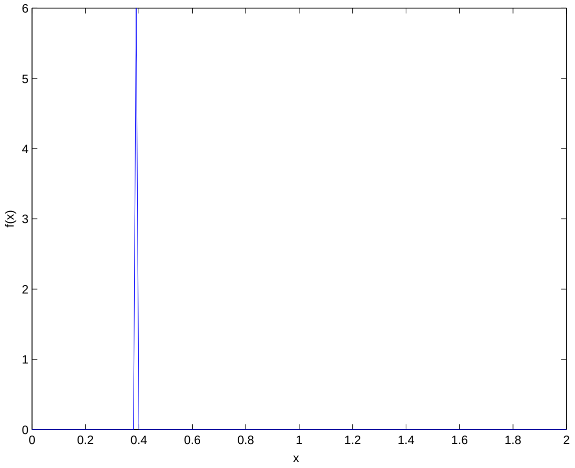
The functional is bounded if there exists a M s.t.

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

for all f where $\| \cdot \|_{Hil}$ is the norm in the Hilbert space.

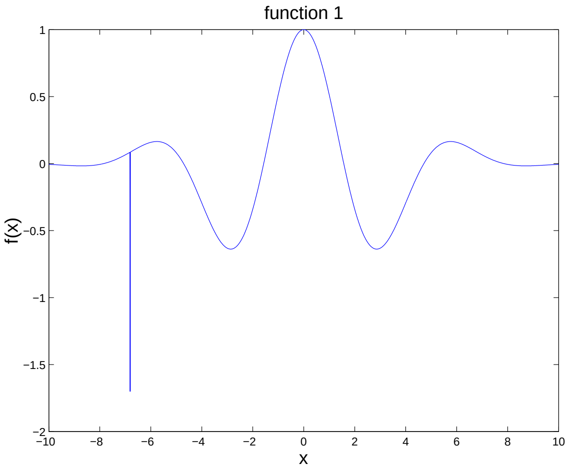
Evaluation functionals in Hilbert space

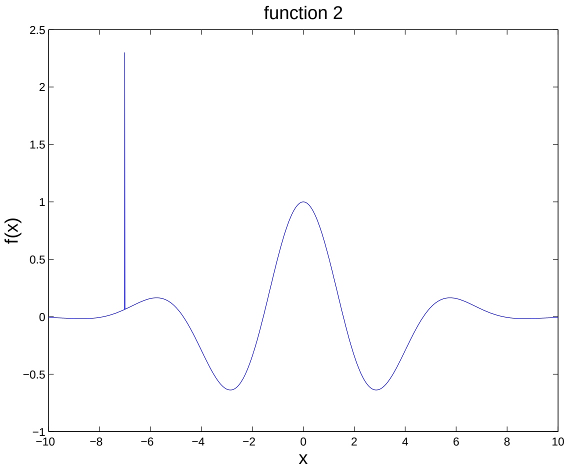
The evaluation functional is not bounded in the familiar Hilbert space $L_2([0,1])$, no such M exists and in fact elements of $L_2([0,1])$ are not even defined pointwise.



Evaluation functionals in Hilbert space

In the following pictures the two functions have the same norm but they are very different on sets of zero measure





RKHS

Definition A (real) RKHS is a Hilbert space of real-valued functions on a compact domain X with the property that for each t the evaluation functional $\mathcal{F}_t$ is a bounded linear functional.

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 $K(\mathbf{t}, \mathbf{s})$ is *positive definite (pd)* if

$$\sum_{i,j=1}^{n} c_i c_j K(\mathbf{t}_i, \mathbf{t}_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $\mathbf{t}_1, ..., \mathbf{t}_n \in X$ and $c_1, ..., c_n \in \mathbb{R}$.

Positive definite kernels (cont.)

An equivalent definition could be given in terms of positive semidefiniteness of the matrix

$$K_{ij} = K(\mathbf{t}_i, \mathbf{t}_j).$$

A *pd* kernel is *strictly positive definite* if for any distinct vectors $\mathbf{t}_1, ..., \mathbf{t}_n \in X$ the above inequality holds strictly when the c_i are not all zero (in that case the matrix K_{ij} is positive definite and not just positive semidefinite). Definitions in the literature are often inconsistent and confusing (especially the different use for matrices vs. functions).

RKHS and kernels

The following theorem relates kernels and RKHS.

Theorem

a) For every RKHS there exists a unique, positive definite function called the reproducing kernel (rk)

b) Conversely for every positive definite function K on $X \times X$ there is a unique RKHS on X with K as its rk

Sketch of proof $+$ some important concepts

If $\mathcal{H}$ is a RKHS, then for each $x \in X$ there exists by the Riesz representation theorem an element K_x of $\mathcal{H}$ (called *representer of evaluation*) with the property – called by Aronszajn – the *reproducing property*

- $\mathcal{F}_x[f] = \langle K_x, f \rangle_K = f(x).$

The rk is $K_x(t)$.

Sketch of proof (cont.)

The rk $K_x(t)$ can be thought of as effectively a delta function for that space but the rk belongs to the Hilbert space.

$$\begin{aligned} K_x(t) &\Leftrightarrow \delta_x(t) \\ f(x) = \langle K_x(t), f(t) \rangle_K &\Leftrightarrow f(x) = \langle \delta_x(t), f(t) \rangle \\ K_x(t) \in \mathcal{H} &\not\Leftrightarrow \delta_x(t) \notin L_2. \end{aligned}$$

Sketch of proof (cont.)

It is called a *reproducing kernel* also because

- $K(t,x) = \langle K(t,\cdot), K(x,\cdot)\rangle_K.$

It is pd because

$$\sum_{i,j=1}^{n} c_i c_j K(\mathbf{t}_i, \mathbf{t}_j) = \sum_{i,j=1}^{n} c_i c_j \langle K_{\mathbf{t}_i}, K_{\mathbf{t}_j}\rangle_K = ||\sum c_j K_{\mathbf{t}_j}||_K^2 \geq 0.$$

Sketch of proof (cont.)

Conversely, given K one can construct the RKHS $\mathcal{H}$ as the completion of the space of functions spanned by the set K_x with $x \in X$ with a inner product defined as follows.

The dot product of two functions f and g in $\mathcal{H}$

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

where $s \in \mathbb{N}$, (s is finite or infinite depending on the kernel), can be written as

$$\begin{aligned}\langle f, g\rangle_K &= \left\langle \sum_{i=1}^{s} \alpha_i K(x_i, \cdot), \sum_{i=1}^{s} \beta_i K(x_i, \cdot) \right\rangle_K \\ &= \sum_{i,j=1}^{s} \alpha_i \beta_j K(x_i, x_j).\end{aligned}$$

Norms in RKHS, Complexity, and Smoothness

We will measure the complexity of a hypothesis space using the the RKHS norm, $\|f\|_K$.

The next example illustrate how bounding the RKHS norm corresponds to enforcing some kind of "simplicity" or smoothness of the functions.

A linear example

Our function space is 1-dimensional lines

$$f(x) = w\,x \text{ and } K(x, x_i) \equiv x\,x_i.$$

For this kernel

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) = \sum_{i=1}^{n} \alpha_i x_i\, x = x \sum_{i=1}^{n} \alpha_i x_i = xw$$

so $w = \sum_{i=1}^{n} \alpha_i x_i$.

Using the RKHS norm

$$\|f(x)\|_K^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j x_i x_j = \left(\sum_{i=1}^{n} \alpha_i x_i \right) \left(\sum_{j=1}^{n} \alpha_j x_j \right) = w^2$$

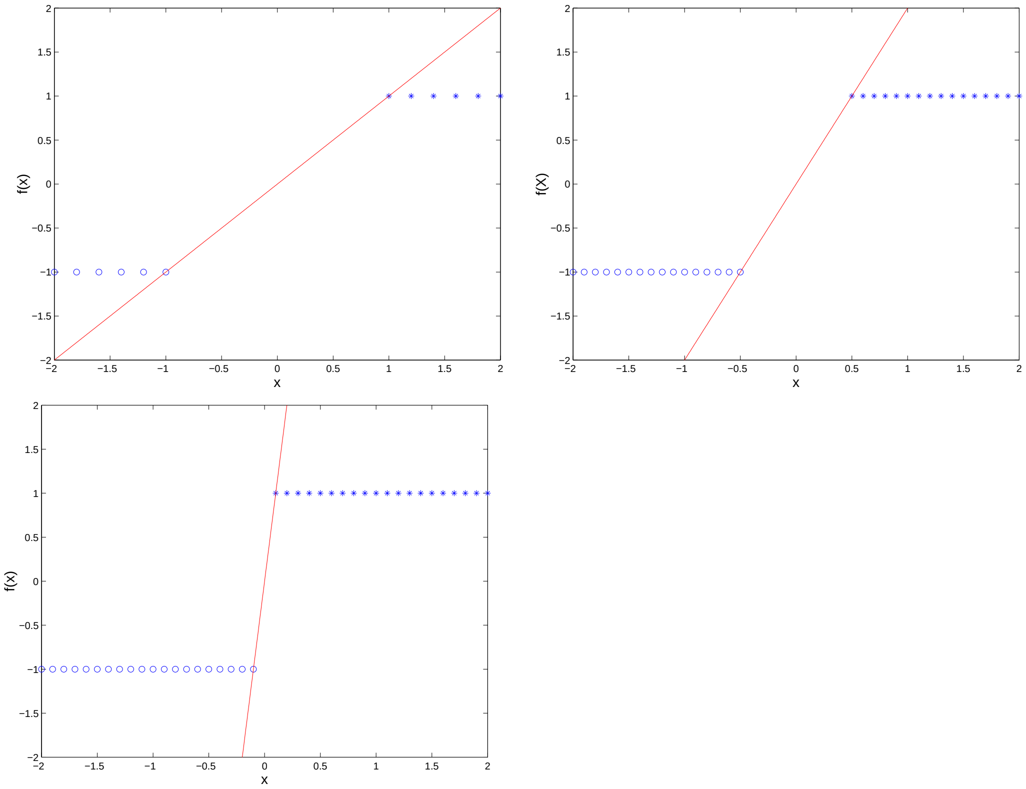
so our measure of complexity is the slope of the line.

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.

A linear example (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.



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. Of course, RKHS were used much earlier in approximation theory (eg Wahba, 1990...) and computer vision (eg Bertero, Torre, Poggio, 1988...).

Tikhonov Regularization

As we saw in the last class we replace ERM with minimization of the following functional, trading off the training error and the complexity of the hypothesis (measured by the radius of the ball in the RKHS):

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

where $\mathcal{H}$ is the RKHS as defined by the kernel $K(\cdot, \cdot)$.

The general solution to Tikhonov regularization (in RKHS): the Representer Theorem

Theorem. The solution to the Tikhonov regularization problem

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

can be written in the form

$$f(x) = \sum_{i=1}^{\ell} c_i K(x, x_i).$$

This theorem is exceedingly useful — it says that to solve the Tikhonov regularization problem, we need only find the best function of the form $f(x) = \sum_{i=1}^{\ell} c_i K(x, x_i)$. Put differently, all we have to do is find the c_i.

): The Representer Theorem: short proof

Very short proof. We consider the square loss case. Apply the operator $\int dt\overline{f(t)}\frac{\partial}{\partial f}$, that is the integral of the functional derivative, to

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(f(x_i)-y_i)^2+\lambda\|f\|_K^2$$

and set it equal zero.

Thus

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(f(x_i)-y_i)\overline{f}(x_i)+\lambda\langle f,\overline{f}\rangle=0.$$

The equation must be valid for any $\overline{f}$. In particular, setting $\overline{f} = K_x$ gives

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(f(x_i) - y_i)K_x(x_i) + \lambda\langle f, K_x\rangle = 0$$

): The Representer Theorem: short proof

$$\begin{aligned}\frac{1}{\lambda\ell}\sum_{i=1}^{\ell}(y_i - f(x_i))K_x(x_i) &= \langle f, K_x\rangle \\ \frac{1}{\lambda\ell}\sum_{i=1}^{\ell}(y_i - f(x_i))K_x(x_i) &= f(x)\end{aligned}$$

so we can write

$$f(x) = \sum_{i=1}^{\ell} c_i K_{x_i}(x)$$

where

$$c_i = \frac{y_i - f(x_i)}{\ell\lambda}$$

since $\langle f, K_x\rangle = f(x)$.

Tikhonov Regularization 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.