Reproducing Kernel Hilbert Spaces

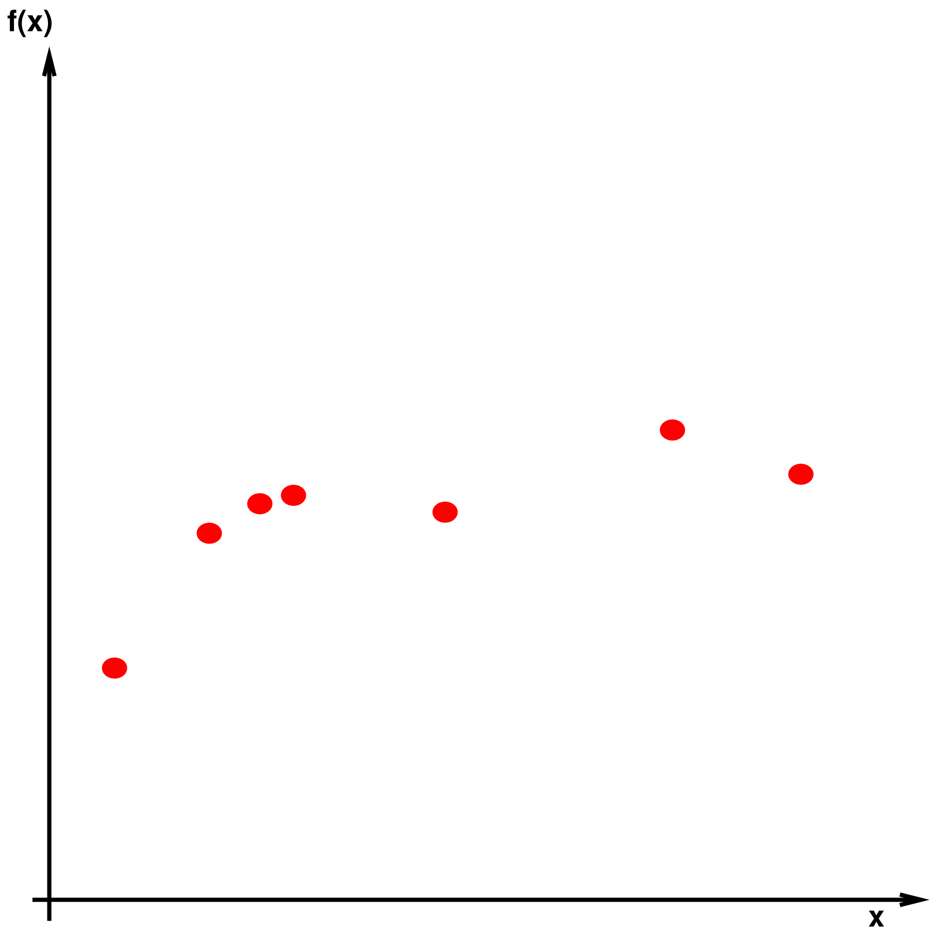
9.520 Class 03, 15 February 2006

Andrea Caponnetto

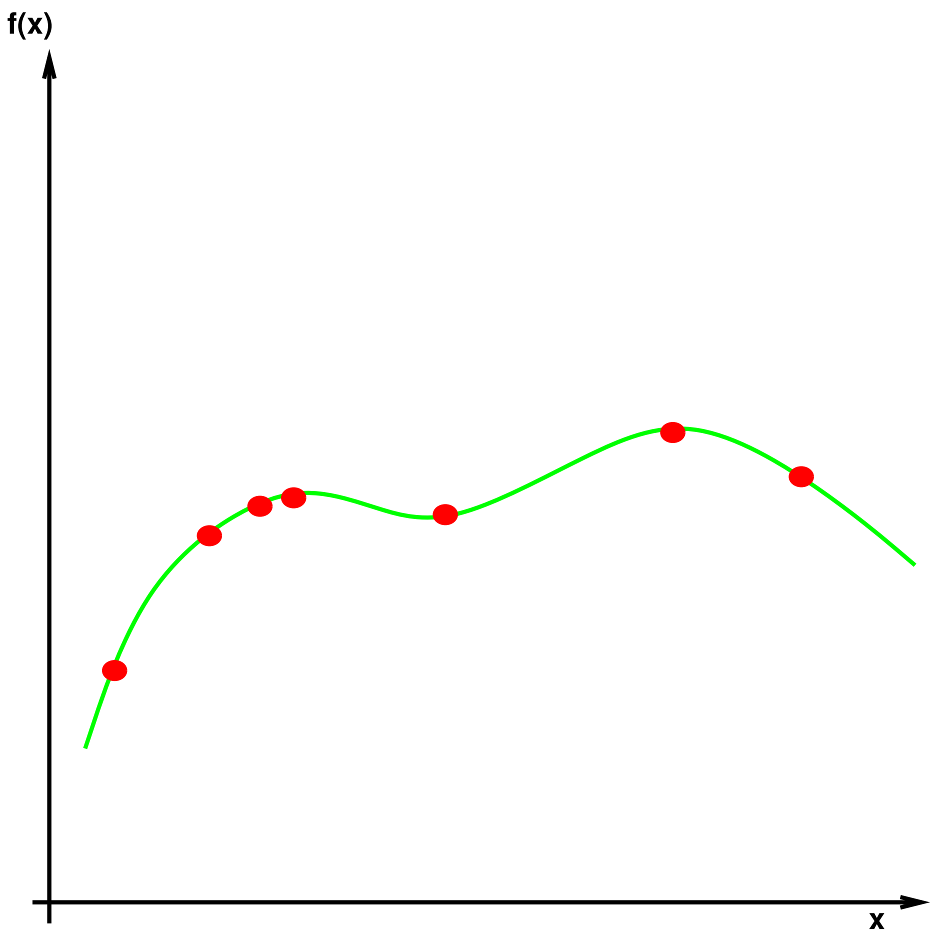
About this class

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.

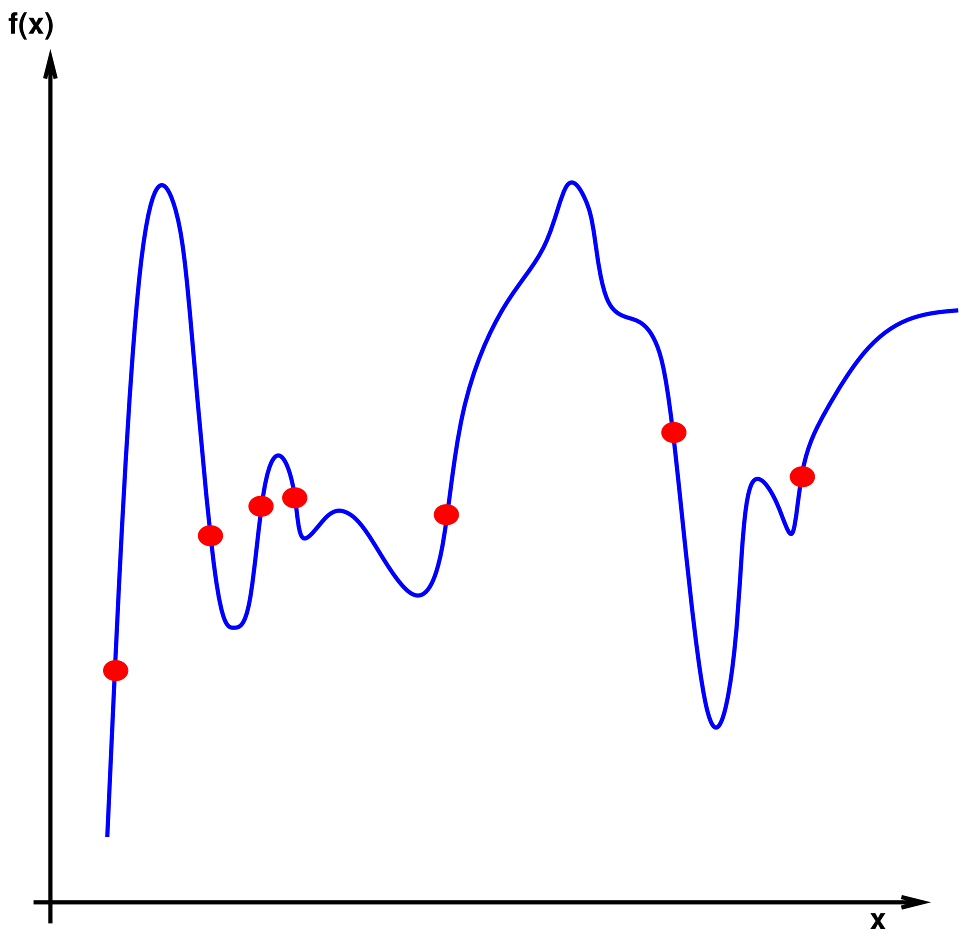
Here is a graphical example for generalization: given a certain number of samples...



suppose this is the "true" solution...



... 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 $\mathcal{H}$. The direct way – minimize the empirical error subject to f in a ball in an appropriate normed functional space $\mathcal{H}$ – is called Ivanov regularization. The indirect way is Tikhonov regularization (which is not ERM).

Ivanov regularization over normed spaces

ERM finds the function in $\mathcal{H}$ which minimizes

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$$

which in general – for arbitrary hypothesis space $\mathcal{H}$ – is *ill-posed*. Ivanov regularizes by finding the function that minimizes

$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i)$$

while satisfying

$$\|f\|_{\mathcal{H}}^2 \leq A,$$

with $\|\cdot\|$, the norm in the *normed function space* $\mathcal{H}$

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 $[a, b]$

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.

Examples

1. A norm in $C[a,b]$ can be established by defining

$$\|f\| = \max_{a \le t \le b} |f(t)|.$$

2. A norm in $L_1[a,b]$ can be established by defining

$$\|f\| = \int_a^b |f(t)| dt.$$

3. A norm in $L_2[a,b]$ can be established by defining

$$\|f\| = \left(\int_a^b f^2(t) dt \right)^{1/2}.$$

From Ivanov to Tikhonov regularization

Alternatively, by the *Lagrange multipler's technique*, Tikhonov regularization minimizes over the whole normed function space $\mathcal{H}$, for a fixed positive parameter λ, the regularized functional

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

In practice, the normed function space $\mathcal{H}$ that we will consider, is a *Reproducing Kernel Hilbert Space* (RKHS).

Lagrange multiplier's technique

Lagrange multiplier's technique allows the reduction of the constrained minimization problem

$$\begin{array}{ll} \textbf{Minimize} & I(x) \\ \textbf{subject to} & \Phi(x) \leq A \qquad\qquad \text{(for some } A\text{)} \end{array}$$

to the unconstrained minimization problem

$$\textbf{Minimize} \quad J(x) = I(x) + \lambda\Phi(x) \qquad\qquad \text{(for some } \lambda \geq 0\text{)}$$

Hilbert space

A **Hilbert space** is a normed space whose norm is induced by a *dot product* $\langle f, g \rangle$ by the relation

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

A Hilbert space must also be *complete* and *separable*.

• Hilbert spaces generalize the finite Euclidean spaces $\mathbb{R}^d$, and are generally *infinite dimensional*.

• Separability implies that Hilbert spaces have *countable orthonormal bases*.

Examples

• Euclidean d-space. The set of d-tuples $x = (x_1, ..., x_d)$ endowed with the dot product

$$\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i.$$

The corresponding norm is

$$\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}.$$

The vectors

$$e_1 = (1, 0, \ldots, 0),\ e_2 = (0, 1, \ldots, 0),\ \ldots,\ e_d = (0, 0, \ldots, 1)$$

form an orthonormal basis, that is $\langle e_i, e_j \rangle = \delta_{ij}$.

Examples (cont.)

• The function space $L_2[a,b]$ consisting of square integrable functions. The norm is induced by the dot product

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt.$$

It can be proved that this space is complete and separable.

An important example of orthogonal basis in this space is the following set of functions

$$1,\ \cos\frac{2\pi nt}{b-a},\ \sin\frac{2\pi nt}{b-a} \quad (n = 1, 2, ...).$$

• $C[a,b]$ and $L_1[a,b]$ are **not** Hilbert spaces.

Evaluation functionals

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

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

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

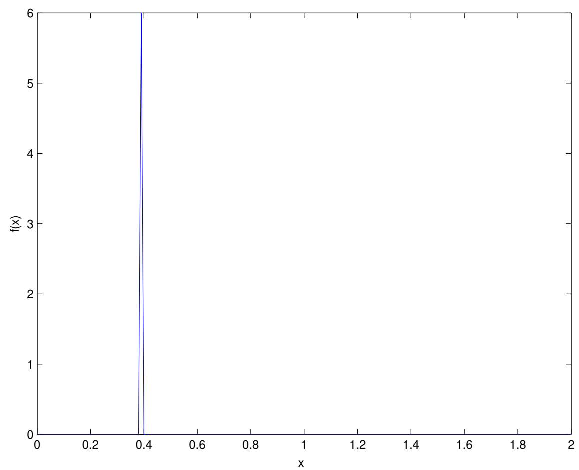
$$|\mathcal{F}_t[f]| = |f(t)| \leq M\|f\|_{\mathcal{H}} \ \forall f \in \mathcal{H}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm in the Hilbert space of functions.

• we don't like the space $L_2[a,b]$ because the its evaluation functionals are *unbounded*.

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.



RKHS

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

Reproducing kernel (rk)

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

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

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

$$K_t(x) = \langle K_t, K_x \rangle_K$$

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 $K(\mathbf{t}, \mathbf{s})$ 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 $c_1, ..., c_n \in \mathbb{R}$.

RKHS and kernels

The following theorem relates pd kernels and RKHS.

Theorem

a) For every RKHS the rk is a positive definite kernel on.

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

Sketch of proof

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

- Symmetry follows from the symmetry property of dot products

$$\langle K_t, K_x \rangle_K = \langle K_x, K_t \rangle_K$$

- K 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.)

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 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*

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

Examples of pd kernels

Very common examples of symmetric pd kernels are

- **Linear kernel**

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$$

- **Gaussian kernel**

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

- **Polynomial kernel**

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

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

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...).

Back to Tikhonov Regularization

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

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

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*.

Norms in RKHS, Complexity, and Smoothness

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

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

Regularity of functions in RKHS

Functions over X, in the RKHS $\mathcal{H}$ induced by a pd kernel K, fulfill a Lipschitz-like condition, with Lipschitz constant given by the norm $\|f\|_K$.

In fact, by the Cauchy-Schwartz inequality, we get $\forall x, x' \in X$

$$|f(x) - f(x')| = |\langle f, K_x - K_{x'} \rangle_K| \leq \|f\|_K \; d(x, x'),$$

with the distance d over X, defined by

$$d^2(x, x') = K(x, x) - 2K(x, x') + K(x', x').$$

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

$$d^2(x, x') = K(x, x) - 2K(x, x') + K(x', x') = |x - x'|^2,$$

and using the RKHS norm

$$\|f\|_K^2 = \langle f, f \rangle_K = \langle K_w, K_w \rangle_K = K(w, w) = w^2$$

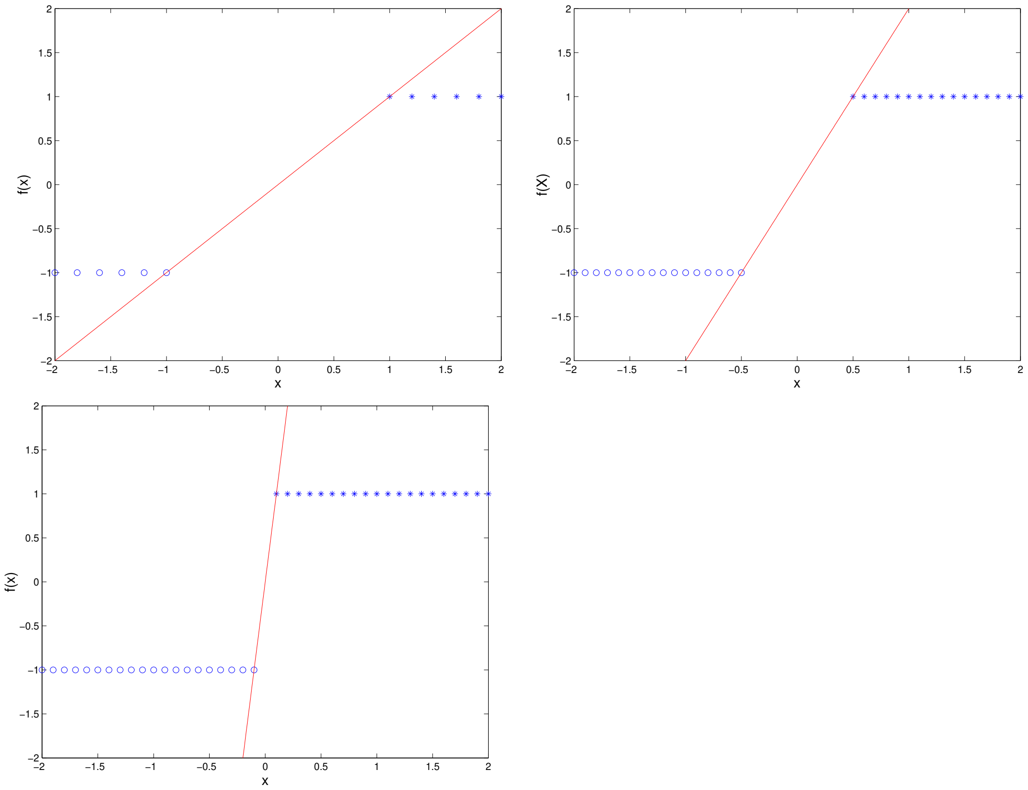
so our measure of complexity is the slope of the line.

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.



Again Tikhonov Regularization

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

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

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**.

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)

$$V = (y - f(\mathbf{x}))^2$$

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

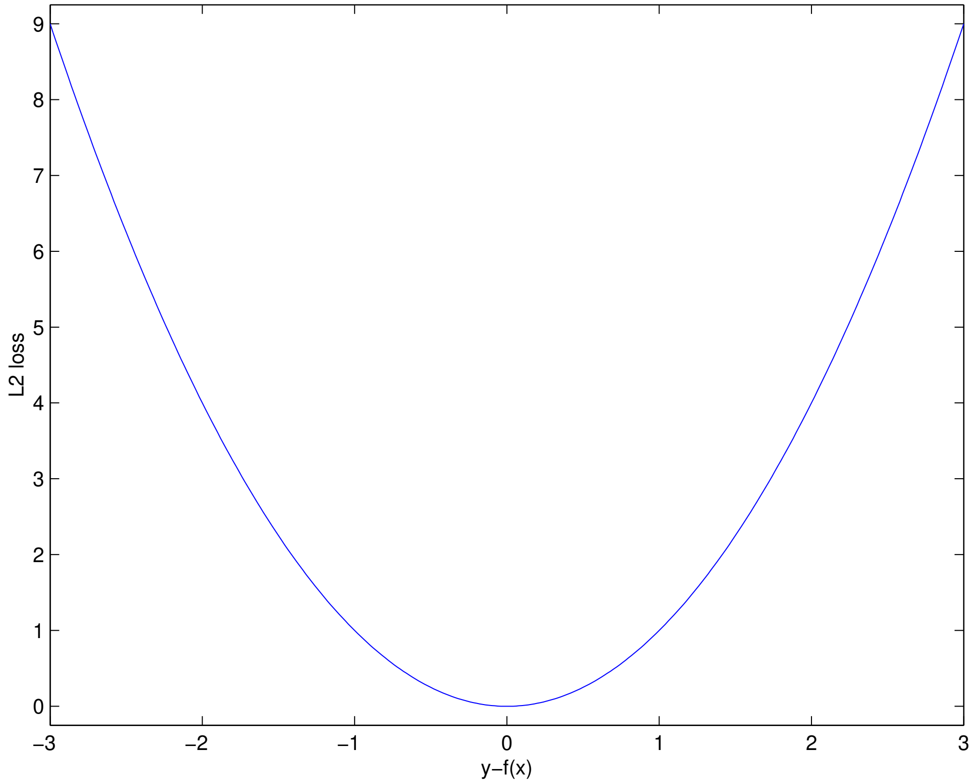
$$V = |1 - yf(\mathbf{x})|_+$$

where

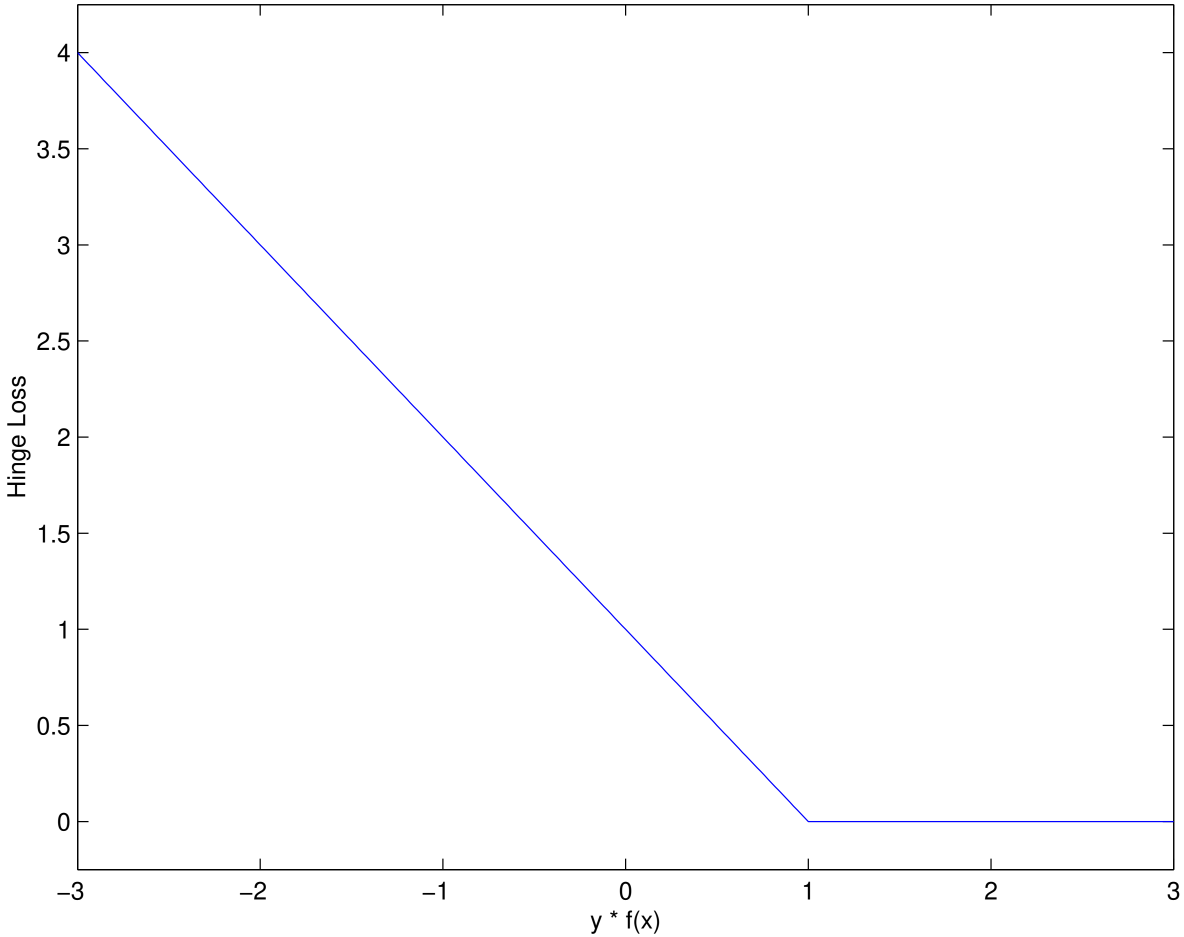
$$(k)_+ \equiv \max(k, 0).$$

The Square Loss

For regression, a natural choice of loss function is the square loss $V(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$.



The Hinge Loss



Existence and uniqueness of minimum

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

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

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

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

$$I_S[f] + \lambda \|f\|_K^2,$$

can be represented by the expression

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

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 $\mathcal{H}$,

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

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

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

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

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

$\mathcal{H}_0^\perp$ 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

$$I_S[f_0] + \lambda\|f_0\|_K^2 \leq I_S[f_0 + f_0^\perp] + \lambda\|f_0 + f_0^\perp\|_K^2.$$

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

Applying the Representer Theorem

Using the representer theorem the minimization problem over $\mathcal{H}$

$$\min_{f \in \mathcal{H}} I_S[f] + \lambda \|f\|_K^2$$

can be easily reduced to a minimization problem over $\mathbb{R}^n$

$$\min_{\mathbf{c} \in \mathbb{R}^n} g(\mathbf{c})$$

for a suitable function $g(\cdot)$.

If the loss function is convex, then g is convex, and finding the minimum reduces to computing the *subgradient* of g.

In particular, if the loss function is differentiable (eg. squared loss), we just have to compute (and set to zero) the *gradient* of g.

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.