9.520: Statistical Learning Theory and Applications February 10th, 2010

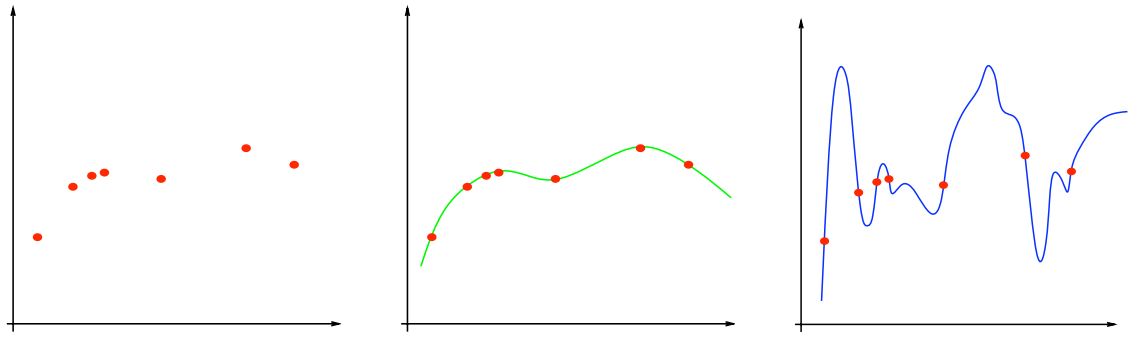
Reproducing Kernel Hilbert Spaces

Lecturer: Lorenzo Rosasco *Scribe: Greg Durrett*

1 Introduction

In the previous two lectures, we've discussed the connections of the learning problem to statistical inference. However, unlike in traditional statistics, our primary goal with learning is to predict the future rather than describe the data at hand. We also typically have a much smaller sample of data in a much higher-dimensional space, so we cannot blindly choose a model and assume it will be accurate. If the model is too highly-parameterized, it will react too strongly to the data, we will overfit the data, and we will fail to learn the underlying phenomenon (see Figure 1 for an example of this behavior). However, models with too few parameters may not even describe the training data adequately, and will provide similarly bad performance.

Regularization provides us with one way to strike the appropriate balance in creating our model. It requires a (possibly large) class of models and a method for evaluating the complexity of each model in the class. The concept of "kernels" will provide us with a flexible, computationally feasible method for implementing this scheme.



(a) Data set for which we wish to learn a function

(b) Smooth function that will likely be a good predictor of future points

(c) Function that probably does not model the data well, but still minimizes empirical error

Figure 1: Two different functions learned from a small training set.

The goal of these notes will be to introduce a particularly useful family of hypothesis spaces called reproducing kernel Hilbert spaces (RKHS), each of which is associated with a particular kernel, and to derive the general solution of Tikhonov regularization in RKHS, known as the representer theorem.

2 Regularization

The goal of regularization is to restore the well-posedness (specifically, making the result depend smoothly on the data) of the empirical risk minimization (ERM) technique by effectively restricting the hypothesis space $\mathcal{H}$. One way of doing this is introduce a penalization term in our minimization as follows:

$$\underbrace{\text{ERR}(f)}_{\text{empirical error}} + \lambda \underbrace{\text{pen}(f)}_{\text{penalization term}}$$

where the regularization parameter λ controls the tradeoff between the two terms. This will then cause the minimization to seek out simpler functions, which incur less of a penalty.

Tikhonov regularization can be written in this way, as

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

where

- $\lambda > 0$ is a regularization parameter,
- $V(f(x), y)$ is the loss function, that is the price we pay when we predict $f(x)$ in place of y, and
- $\|\cdot\|_{\mathcal{H}}$ is the norm in the function space $\mathcal{H}$.

This formulation is powerful, as it does not present a specific algorithm, but rather a large class of algorithms. By choosing V and $\mathcal{H}$ differently, we can derive a wide variety of commonly-used techniques, including traditional linear regression and support vector machines (SVMs).

Given our intuition about what causes overfitting, the penalization should somehow force us to choose f to be as smooth as possible while still fitting the data. The norm from our function space $\mathcal{H}$ will allow us to encode this criterion, but in order to design this norm appropriately, we need to describe reproducing kernel Hilbert spaces.

3 Functional Analysis Background

In order to define RKHS, we will make use of several terms from functional analysis, which we define here. Additional review on functional analysis can be found in the notes from the math camp, available on the website.

Definition 1 *A* **function space** $\mathcal{F}$ *is a space whose elements are functions, e.g.* $f : \mathbb{R}^d \to \mathbb{R}$.

Definition 2 *An* **inner product** *is a function* $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ *that satisfies the following properties for every* $f, g \in \mathcal{F}$ *and* $\alpha \in \mathbb{R}$*:*

1. *Symmetric:* $\langle f, g \rangle = \langle g, f \rangle$
2. *Linear:* $\langle r_1 f_1 + r_2 f_2, g \rangle = r_1 \langle f_1, g \rangle + r_2 \langle f_2, g \rangle$
3. *Positive-definite:* $\langle f, f \rangle \geq 0$ *for all* $f \in \mathcal{F}$ *and* $\langle f, f \rangle = 0$ *iff* $f = 0$.

Definition 3 *A* **norm** *is a nonnegative function* $\|\cdot\| : \mathcal{F} \to \mathbb{R}$ *such that for all* $f, g \in \mathcal{F}$ *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 an inner product, as $\|f\| = \sqrt{\langle f, f \rangle}$

Note that while the dot product in $\mathbb{R}^d$ is an example of an inner product, an inner product is more general than this, and requires only those properties given above. Similarly, while the Euclidean norm is an example of a norm, we consider a wider class of norms on the function spaces we will use.

Definition 4 *A* **Hilbert space** *is a complete, (possibly) infinite-dimensional linear space endowed with an inner product.*

A norm in $\mathcal{H}$ can be naturally defined from the given inner product, as $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. Although it is possible to impose a different norm so long as it satisfies the criteria given above, we will not do this in general; our norm is assumed to be the norm derived from the inner product. Furthermore, we always assume that $\mathcal{H}$ is **separable** (contains a countable dense subset) so that $\mathcal{H}$ has a countable orthonormal basis.

While this tells us what a Hilbert space is, it is not intuitively clear why we need this mechanism, or what we gain by using it. Essentially, a Hilbert space lets us apply concepts from finite-dimensional linear algebra to infinite-dimensional spaces of functions. In particular, the fact that a Hilbert space is **complete** will guarantee the convergence of certain algorithms. More importantly, the presence of an inner product allows us to make use of orthogonality and projections, which will later become important.

3.1 Examples of function spaces

- One function space is the space of continuous functions on the interval $[a, b]$, denoted by $C[a, b]$. A norm can be established by defining

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

  However, there is no inner product for the space that induces this norm, so it is not a Hilbert space.

- Another example is square integrable functions on the interval $[a, b]$, denoted by $L_2[a, b]$. We define the inner product as

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

  This produces the correct norm:

  $$\|f\| = \int_a^b f^2(x)dx$$

  It can be checked that this space is complete, so it is a Hilbert space. However, there is one problem with the functions in this space. Consider trying to evaluate the function $f(x)$ at the point $x = k$. There exists a function g in the space defined as follows:

  $$g(x) = \begin{cases} c & \text{if } x = k \\ f(x) & \text{otherwise} \end{cases}$$

  Because it differs from f only at one point, g is clearly still square-integrable, and moreover, $\|f - g\| = 0$. However, we can set the constant c (or, more generally, the value of $g(x)$ at any finite number of points) to an arbitrary real value. What this means is that a condition on the integrability of the function is not strong enough to guarantee that we can use it predictively, since prediction requires evaluating the function at a particular data value. This characteristic is what will differentiate reproducing kernel Hilbert spaces from ordinary Hilbert spaces, as we discuss in the next section.

4 Reproducing Kernel Hilbert Spaces

Definition 5 *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*

$$\mathcal{F}_t[f] = f(t) \quad \text{for all } f \in \mathcal{H}.$$

Definition 6 *A Hilbert space $\mathcal{H}$ is a* **reproducing kernel Hilbert space (RKHS)** *if the evaluation functionals are bounded, i.e. if for all t there exists some $M > 0$ such that*

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

This condition is not trivial. For $L_2[a, b]$, we showed above that there exist functions that are square-integrable, but which have arbitrarily large values on finite point sets. In this case, no choice of M will give us the appropriate bound on these functions on these point sets.

While this condition might seem obscure or specific, it is actually quite general and is the weakest possible condition that ensures us both the existence of an inner product and the ability to evaluate each function in the space at every point in the domain. In practice, it is difficult to work with this definition directly. We would like to establish an equivalent notion that is more useful in practice. To do this, we will need the "reproducing kernel" from which the reproducing kernel Hilbert space takes its name.

First, from the definition of the reproducing kernel Hilbert space, we can use the Riesz representation theorem to prove the following property.

Theorem 7 *If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists a function $K_t \in \mathcal{H}$ (called the representer of t) with the* **reproducing property**

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

This allows us to represent our linear evaluation functional by taking the inner product with an element of $\mathcal{H}$. Since K_t is a function in $\mathcal{H}$, by the reproducing property, for each $x \in X$ we can write

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

We take this to be the definition of reproducing kernel in $\mathcal{H}$.

Definition 8 *The* **reproducing kernel** *(rk) of $\mathcal{H}$ is a function $K : X \times X \to \mathbb{R}$, defined by*

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

In general, we have the following definition of a reproducing kernel.

Definition 9 *Let X be some set, for example a subset of $\mathbb{R}^d$ or $\mathbb{R}^d$ itself. A function $K : X \times X \to \mathbb{R}$ is a* **reproducing kernel** *if it is symmetric, i.e. $K(x, y) = K(y, x)$, and positive definite:*

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

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

Having this general notion of a reproducing kernel is important because it allows us to define an RKHS in terms of its reproducing kernel, rather than attempting to derive the kernel from the definition of the function space directly. The following theorem formally establishes the relationship between the RKHS and a reproducing kernel.

Theorem 10 *A RKHS defines a corresponding reproducing kernel. Conversely, a reproducing kernel defines a unique RKHS.*

Proof: To prove the first statement, we must prove that the reproducing kernel $K(t, x) = \langle K_t, K_x\rangle_{\mathcal{H}}$ is symmetric and positive-definite.

Symmetry follows from the symmetry property of inner products:

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

K is positive-definite because

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

To prove the second statement, 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 an inner product defined as follows: given two functions f and g in $\text{span}\{K_x | x \in X\}$

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

we define their inner product to be

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

(This is only a sketch of the proof.) □

Now we have a more concrete concept of what an RKHS is and how we might create such spaces for ourselves. If we can succeed at writing down a reproducing kernel, we know that there exists an associated RKHS, and we need not concern ourselves with the particulars of the boundedness criterion.

4.1 Examples of reproducing kernels

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

4.2 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...).

In general, it is quite difficult to find useful function spaces that *aren't* RKHS.

5 Norms and Smoothness

We established earlier that if a space of functions can be represented as an RKHS, it has useful properties (namely the inner product and the ability for each function to be evaluated at each point) that allow us to use it to solve learning problems. Armed with the notion of kernels, we can now describe specific examples of RKHS and examine how their different norms provide different forms of regularization.

Sobolev kernel Consider functions $f : [0, 1] \to \mathbb{R}$ with $f(0) = f(1) = 0$. The kernel

$$K(x, y) = \Theta(y - x)(1 - y)x + \Theta(x - y)(1 - x)y$$

induces the norm

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

where $F(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\, dt$ is the Fourier tranform of f. Such a norm is very useful because it allows us to regularize on the basis of frequency content. In particular, the more prominent the high-frequency components of f, the higher $\|f\|_{\mathcal{H}}^2$ will be; in fact, the norm will be infinite for any function whose frequency magnitudes do not decay faster than $\frac{1}{\omega}$. This imposes a condition on the smoothness of the functions, since a high derivative gives rise to high frequency components.

The (somewhat mysterious) reproducing kernel written above was designed to yield this useful norm, and was not created arbitrarily.

Gaussian kernel It is possible to see that the Gaussian kernel yields as the norm:

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

which penalizes high-frequency components even more harshly.

5.1 Linear case

We illustrate how regularization controls complexity through a simple linear case. Our function space is 1-dimensional lines through the origin with a linear kernel:

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

giving an RKHS norm of

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

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.

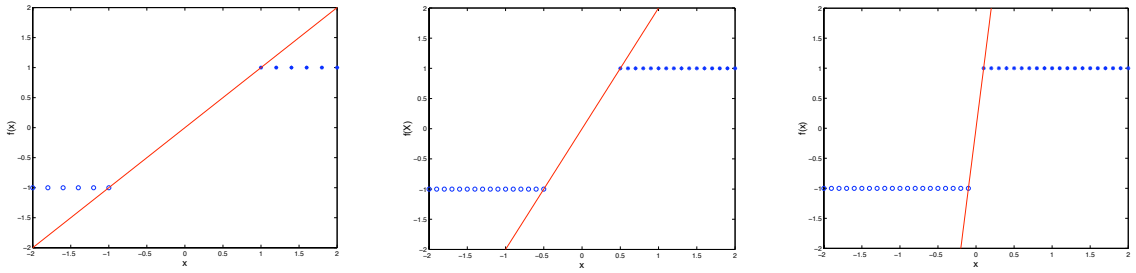
A classification problem can be thought of as "harder" when the distinctions between the two classes are less pronounced. In Figure 2, we see that the less separation there is between the x-values of the two classes, the steeper the slope of the line that is required to model the relationship. Having a norm that increases with slope is therefore a good choice in this case: by penalizing lines with high slope, we only use complex solutions to problems if doing so is necessary to reduce the training error.

6 Solving Tikhonov Regularization: The Representer Theorem

6.1 Well-posedness, existence, and uniqueness

Now that we have RKHS and sensible norms to use for regularization, we can revisit Tikhonov regularization in a more concrete setting. The algorithms (*regularization networks*) that we want to study are defined by an optimization problem over RKHS,

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



(a) Wide margin (b) Moderate margin (c) Small margin

Figure 2: Three different training sets to demonstrate that higher slopes are necessary to describe the data as the class distinctions become finer.

where the regularization parameter λ is a positive real number, $\mathcal{H}$ is the RKHS as defined by the reproducing kernel $K(\cdot,\cdot)$, and $V(\cdot,\cdot)$ is the loss function.

We have imposed stability on this problem through the use of regularization, but we still need to check the other two criteria of well-posedness. Does there always exist a solution to the minimization, and is that solution unique? As it turns out, this requires a condition on the loss function. If the positive loss function $V(\cdot,\cdot)$ is convex with respect to its first entry, the functional

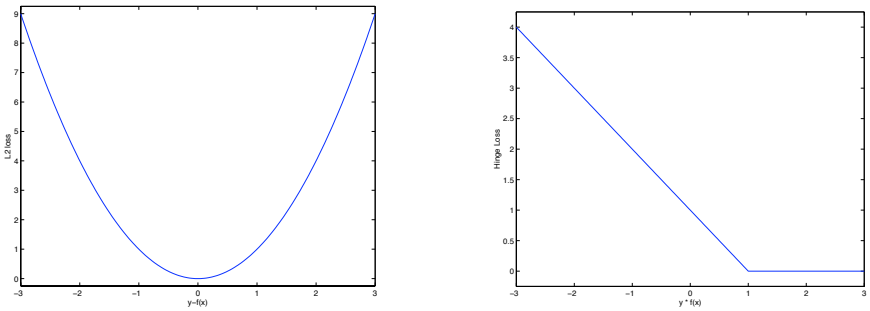
$$\Phi[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

is **strictly convex** and **coercive** (meaning that it grows quickly at the extremes of the space), hence it has exactly one local (and therefore global) minimum.

Both the squared loss and the hinge loss are convex (see Figure 3). On the contrary the 0-1 loss

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

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



(a) Square loss (b) Hinge loss

Figure 3: Two examples of convex loss functions.

6.2 The representer theorem

There is one additional issue to resolve. Because $\mathcal{H}$ is a function space, we note that it may be infinite-dimensional. While this is not a problem in theory, it does pose a computational problem: how can we represent a function with an infinite number of parameters on a computer with a finite

amount of storage? Our solution to Tikhonov regularization could in principle be impossible to write down for this reason, but it is a surprising result that it actually has a very compact representation, as described in the following theorem.

Theorem 11 (The Representer Theorem) *The minimizer over the RKHS $\mathcal{H}$, f_S^λ, 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 $\mathbb{R}^n$.

There are only a finite number n of training set points, so the fact that the minimizer can be written as a linear combination of kernel terms from these points guarantees that we can represent the minimizer as a vector in $\mathbb{R}^n$.

We provide a sketch of the proof for this theorem.

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

$$\mathcal{H}_0 = \{f \in \mathcal{H} \mid f = \sum_{i=1}^{n} \alpha_i K_{x_i}\}.$$

This is the space spanned by the representers of the training set. Let $\mathcal{H}_0^\perp$ be the linear subspace of $\mathcal{H}$ orthogonal to $\mathcal{H}_0$, i.e.

$$\mathcal{H}_0^\perp = \{g \in \mathcal{H} \mid \langle g, f \rangle = 0 \text{ for all } f \in \mathcal{H}_0\}.$$

$\mathcal{H}_0$ is finite-dimensional, hence closed, so we can write $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_0^\perp$. Now we see that every $f \in \mathcal{H}$ can be uniquely decomposed into a component along $\mathcal{H}_0$, denoted by f_0, and a component perpendicular to $\mathcal{H}_0$, given by $f_0^\perp$:

$$f = f_0 + f_0^\perp.$$

By orthogonality

$$\|f_0 + f_0^\perp\|^2 = \|f_0\|^2 + \|f_0^\perp\|^2$$

and by the reproducing property

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

since evaluating $f(x_i) = f_0(x_i) + f_0^\perp(x_i)$ to compute the empirical error requires taking the inner product with the representer K_{x_i}, and doing so nullifies the $f_0^\perp$ term while preserving the f_0 term intact.

Combining these two facts, we see that

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

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

This mechanism for implementing Tikhonov regularization can be applied to regularized least-squares regression and support vector machines, as we will do in the next two classes.