The Application of Reproducing Kernel Hilbert Spaces to Regularization in Machine Learning

MATH 0710 Thesis

Toby Weed

May 17, 2022

Abstract

Reproducing Kernel Hilbert Spaces (RKHS) are spaces of functions with properties that make them useful in a number of applied settings. In this paper I derive the central results of RKHS theory starting from a low level, along the way drawing connections to geometry and providing simple examples in Euclidean space. I then offer a brief overview of the statistical learning theory framework for machine learning and motivate the technique of regularization via a penalty term. Finally, I describe the connection between RKHS and regularization, and show how RKHS theory can reduce regularized empirical risk minimization problems over infinite dimensional function spaces to tractable, finite dimensional problems via the representer theorem.

Contents

1 Introduction **3**

2 Preliminaries **3**
2.1 Hilbert Spaces 4
2.2 Orthogonality 4
2.3 Linear Functionals 5
2.3.1 The Reisz Representation Theorem 8

3 Reproducing Kernel Hilbert Spaces **9**
3.1 Reproducing Kernels 10
3.2 Positive Definite Functions 11
3.3 Kernel Functions 12
3.4 Finite-dimensional RKHS 12
3.5 Examples 14

4 Statistical Learning Theory **16**
4.1 The Learning Problem 16
4.2 Regularization 17
4.2.1 Ill-Posed Problems 18
4.2.2 Adding a Penalty Term 20

5 Regularization and RKHS **20**
5.1 RKHS and Smoothness 20
5.2 Regularization Operators 21
5.3 Representer Theorem 22

6 Conclusion **23**

A Support Vector Machines and Feature Maps **24**
A.1 Support Vector Machines 24
A.2 Feature Maps 27

1 Introduction

This thesis is an exploration of Reproducing Kernel Hilbert Spaces and their application via kernel methods to machine learning. Section 2 offers a concise overview of RKHS' mathematical context, including Hilbert spaces, orthogonality, and linear functionals. This discussion culminates in the Reisz representation theorem, a fundamental functional analysis result which is the precursor to RKHS. The discussion is geared to readers with courses in linear algebra and real analysis under their belts, along with a bit of knowledge about metric spaces, normed spaces, inner product spaces, and so on.

In Section 3, I introduce RKHS and derive the core results which characterize them. Section 3.4 is a discussion of RKHS in finite dimensions intended to make these results intuitively accessible. Connections are made to geometry. Proofs, notation, and examples are chosen to make the chapter as transparent as possible.

Section 4 introduces the framework of statistical learning theory, along with such concepts as Empirical Risk Minimization (ERM), instability, and regularization. No results from Statistical Learning Theory are explored in depth, and the treatment of instability and regularization is confined to intuitive motivation rather than formal derivation from the ERM framework; the purpose is to lay just enough groundwork to give the reader an understanding of how the class of penalty regularized ERM problems arise, why they're important, and how they may be posed naturally in the RKHS context.

Finally, Section 5 makes this connection explicit. First I discuss the formal relation between regularization terms and RKHS norms and state the regularized ERM problem in RKHS. To cap off the paper I prove the celebrated Representer Theorem, an indispensable theoretical tool for machine learning which reduces the problem with which we're concerned from an intractable search through infinite dimensional function space to optimization over a finite number of coefficients.

Appendix A attempts to ground the theory of the paper in a practical setting, explaining notation and terminology which is often used in the machine learning literature due to kernel methods' historical development.

2 Preliminaries

In order to understand RKHS, we need to develop some tools. We'll assume that the reader has some familiarity with metric spaces, normed spaces, and inner product spaces, along with basic topological concepts like continuity and completeness. We'll very briefly introduce Hilbert spaces and orthogonality. The purpose of this chapter is to exhibit the character of Hilbert spaces' duals and to build a bit of intuition by relating them to familiar concepts from linear algebra.

2.1 Hilbert Spaces

Definition 2.1. An **inner product space** is a vector space V on a field F equipped with an **inner product** $\langle \cdot, \cdot \rangle : V \times V \to F$ satisfying the following three properties for all $x, y, z \in V$ and $a, b \in F$:

- $\langle x, y \rangle = \overline{\langle y, x \rangle}$ (*conjugate symmetry*),
- $\langle x, x \rangle > 0$ for all $x \neq 0$ (*positive definiteness*),
- and $\langle ax + by, x \rangle = a\langle x, z \rangle + b\langle y, z \rangle$ (*linearity*).

The canonical example of an inner product space is Euclidean space $\mathbb{R}^n$ equipped with the dot product $\langle \vec{x}, \vec{y} \rangle = \sum_{i=0}^{n} x_i y_i$. Just as the dot product measures how much one vector "lies along" another, inner products are often thought of as measuring the similarity, in one sense or another, between two vectors.

All inner product spaces are normed linear spaces. The **norm induced by an inner product** on a space V is $||v||_V = \sqrt{\langle v, v \rangle_V}$.

Definition 2.2. A **Hilbert Space** is an inner product space that is complete with respect to the norm induced by the inner product.

Hilbert spaces allow us to generalize geometrically-informed methods from linear algebra and calculus from finite Euclidean space to potentially infinite-dimensional spaces, for instance spaces of functions.

2.2 Orthogonality

One of the most useful concepts arising in the study of inner product spaces is that of orthogonality between vectors, which generalizes the Euclidean notion of perpendicularity.

Definition 2.3. Two vectors u, v in an inner product space are said to be **orthogonal** if $\langle u, v \rangle = 0$.

We will often want to talk about orthogonality between sets.

Definition 2.4. The **orthogonal complement** $A^\perp$ of a set A in an inner product space V is the set of vectors which are orthogonal to each vector in A:

$$A^\perp \equiv \{v \in V : \langle v, a \rangle = 0 \text{ for all } a \in A\}.$$

Colloquially, $A^\perp$ is said to "take A to zero." A fundamental fact about Hilbert spaces which we will use in several proofs throughout this paper is that any element can be written as the sum of one element from a closed subspace and one from its orthogonal complement. This is called *orthogonal decomposition.*

Theorem 2.1 (Orthogonal Decomposition Theorem)**.** *If $\mathcal{H}$ is a Hilbert space and S is a closed subspace of $\mathcal{H}$, then*

$$H = S \oplus S^{\perp 1}$$

A full treatment of very rich topic of orthogonality in Hilbert spaces is not the purpose of this thesis, but a good discussion can be found (along with a proof of Theorem 2.1) in section 7.2 of [2].

2.3 Linear Functionals

The Hilbert spaces that we're interested in are spaces of functions, so we'll frequently need to consider maps in which the domain is a function space ("functions of functions"). These are often referred to as **operators**, although this term can be used more generally to refer to functions with any kind of vector domain. When an operator's output is a scalar–the codomain is $\mathbb{C}$ or $\mathbb{R}$–we call that operator a **functional**.

Definition 2.5. A map $\Lambda : X \to \mathbb{R}$ on a normed vector space X is called a **linear functional** if $\Lambda(\alpha f + \beta g) = \alpha\Lambda(f) + \beta\Lambda(g)$ for all scalars α, β and $f, g \in X$.

The space of all linear functionals on a normed vector space X is referred to as the **dual space** of X, sometimes denoted X^*. The elements of X^* might be called "covectors" – sound familiar?

Example 2.1. Consider our favorite old vector space: $\mathbb{R}^n$. Recall from linear algebra that any element $\vec{x} \in \mathbb{R}^n$ may be represented as a linear combination of basis vectors e_i,

$$\vec{x} = \sum_{i=1}^{n} a_i e_i.$$

So for any linear functional $f : \mathbb{R}^n \to \mathbb{R}$,

$$f(\vec{x}) = f\left(\sum_{i=1}^{n} a_i e_i\right) = \sum_{i=1}^{n} a_i f(e_i).$$

But each $f(e_i)$ is just a real number, and the sum on the right is the dot product of $\vec{x}$ with the vector $\vec{y} = (f(e_1), f(e_2), ..., f(e_3))$. We have shown that every linear functional on $\mathbb{R}^n$ – every element of the dual space $\mathbb{R}^{n*}$– corresponds to a vector $\vec{x}' \in \mathbb{R}^n$, that is, for every $f \in \mathbb{R}^{n*}$,

$$f(\vec{x}) = \vec{x} \cdot \vec{y}$$

[1] $\oplus$ is the "direct sum," which here really just means that $S \cap S^{\perp} = \emptyset$ and any element in $\mathcal{H}$ may be written as a sum of one from S and one from $S^{\perp}$.

for all $\vec{x} \in \mathbb{R}^n$. This result might seem trivial, but the underlying concept ends up being quite profound when extended to more exotic vector spaces. The Reisz Representation Theorem (2.3) is a generalization of this result to Hilbert spaces in general.

Definition 2.6. A linear functional is **bounded** if there exists a finite constant C such that $|\Lambda(f)| \leq C||f||_X$ for all $f \in X$.

This is equivalent to saying that $\frac{|\Lambda(f)|}{||f||} \leq C$; the ratio of the norm of $\Lambda(f)$ to the norm of f in X is bounded by a constant. This ensures that the value of the functional for some function is in some sense "anchored" to the norm of the function.

Example 2.2. Consider the Lesbesgue space $L^2[0,1]$ and let $\Lambda_x(f)$ be the *evaluation functional* of $x \in [0,1]$, so that $\Lambda_x(f) = f(x)$.[2] Consider the sequence of functions defined by

$$g_n(x) = \begin{cases} a & 0 \leq x < \frac{1}{n} \\ 0 & \frac{1}{n} \leq x \leq 1 \end{cases}$$

for some $a \neq 0$. Notice that $L_0(g_n) = a$ for all $n \in \mathbb{N}$, but $C||g_n||_X = C(\int g_n^2)^{\frac{1}{2}} \to 0$. Another way to say this is that $\frac{|L_0(g_n)|}{||g_n||_X}$ is unbounded as $n \to \infty$. L_0 is *not* bounded.

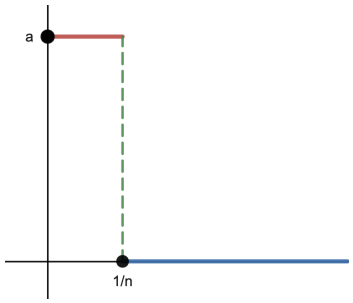

Figure 1: $g_n(x)$.

Example 2.3. Let $\vec{B}(\vec{x})$ be any linear operator between finite-dimensional normed spaces, $B : N \to M$, where $\dim N = n$ and $\dim M = m$. As in Example 2.1, any $\vec{x} \in N$ may be written as a linear combination of basis vectors $\vec{x} = \sum_{i=1}^{i=n} a_i \vec{e_i}$, so by the linearity of B,

[2]Since the elements of L^2 are in fact equivalence classes of functions rather than functions themselves, defining point evaluation functionals doesn't really make sense. However, this example nicely demonstrates the intuition behind what it means for point evaluation to be unbounded relative to a norm, so we'll just ignore this incoherence for the sake of pedagogy.

$$||\vec{B}(\vec{x})||_M = \left|\left|\vec{B}\left(\sum_{i=1}^{i=n} a_i\vec{e_i}\right)\right|\right|_M$$

$$\leq \sum_{i=1}^{n} |a_i| \, ||\vec{B}(\vec{e_i})||_M.$$

Let $\mathcal{M} = \max_{i\in 1,\ldots,n}\{||\vec{B}(\vec{e_i})||_M\}$. Then

$$||\vec{B}(\vec{x})||_M \leq \mathcal{M} \sum_{i=1}^{i=n} |a_i|$$

The quantity $\sum_{i=1}^{i=n} |a_i|$ is the L^1-norm of $\vec{x}$ relative to the particular basis we have chosen. Using without proof the fact that all norms on a finite-dimensional vector space are equivalent [23], we get

$$||\vec{B}(\vec{x})||_M \leq \mathcal{M}'||\vec{x}||_N.$$

B is a bounded operator. □

This notion of boundedness is slightly different in character and behavior from the one we're used to. In the context of linear functionals, boundedness is identical to continuity.

Proposition 2.2. *For a linear functional $\Lambda : X \to \mathbb{C}$ on a normed vector space X the following are equivalent:*

1. *Λ is continuous*
2. *Λ is continuous at 0*
3. *Λ is bounded*

Proof. **(1) ⇒ (2):** by definition.

(2) ⇒ (3): Since $\Lambda(\alpha 0 + \beta y) = \alpha\Lambda(0) + \Lambda(\beta y) = \Lambda(\beta y)$ for all α, $\Lambda(0) = 0$. If Λ is continuous, then there exists a $\delta > 0$ such that if $||x|| \leq \delta$, $|\Lambda(x)| < 1$. Then,

$$|\Lambda(x)| = \left|\Lambda\left(\tfrac{||x||}{\delta} \cdot \tfrac{\delta}{||x||} \cdot x\right)\right| = \left|\tfrac{||x||}{\delta} \cdot \Lambda\left(\tfrac{\delta}{||x||} \cdot x\right)\right|.$$

Notice that $\left|\left|\tfrac{\delta}{||x||} \cdot x\right|\right| = \delta \cdot \tfrac{||x||}{||x||} = \delta \leq \delta$, so $\Lambda\left(\tfrac{\delta}{||x||} \cdot x\right) < 1$ and

$$|\Lambda(x)| \leq \frac{||x||}{\delta}.$$

This is the definition of boundedness with $C = 1/\delta$.

(3) $\Rightarrow$ (1): Let Λ be bounded, so that $|\Lambda(x)| \leq C||x||$ for all $x \in X$. Let $\epsilon > 0$.

Assume that $||x - x'|| < \delta = \epsilon/C$. Then by boundedness, $|\Lambda(x - x')| \leq C||x - x'|| < C \cdot \delta = \epsilon$. But by linearity $|\Lambda(x - x')| = |\Lambda(x) - \Lambda(x')|$, so $|\Lambda(x) - \Lambda(x')| < \epsilon$.

□

2.3.1 The Reisz Representation Theorem

We have hinted that there is a correspondence between a Hilbert space $\mathcal{H}$ and the corresponding dual space of linear functionals defined on $\mathcal{H}$. This correspondence is critical for our purposes, and partly captured by the following theorem.

Theorem 2.3 (**Reisz Representation Theorem**, [1], p. 17)**.** *Every bounded linear functional L on a Hilbert space $\mathcal{H}$ is associated with a unique element $k \in \mathcal{H}$ such that*

$$L(f) = \langle f, k \rangle$$

for all $f \in \mathcal{H}$.

Proof. If $L = 0$ everywhere in $\mathcal{H}$, then we can simply set $k = 0$. Otherwise, let

$$Z \equiv \{f \in \mathcal{H} : L(f) = 0\} = L^{-1}(0).$$

A few observations:

(i) Z is a vector subspace of $\mathcal{H}$ by L's linearity.

(ii) The topological characterization of continuity implies that $Z = L^{-1}(0)$ is closed since the point 0 is compact.[3]

(iii) $Z \neq \mathcal{H}$ since $L \neq 0$.

We're going to construct the k we're looking for out of an element from Z's orthogonal complement. Z is a closed subspace of $\mathcal{H}$ by (i) and (ii), so Theorem 2.1 asserts that $\mathcal{H} = Z \oplus Z^{\perp}$. Together with (iii), this means that $Z^{\perp} \neq \emptyset$.

We can thus start by choosing some nonzero $z \in Z^{\perp}$. We're free to assume that $L(z) = 1$ since we can always rescale the bounded linear operator L. Now consider an arbitrary $f \in \mathcal{H}$ and let $v = L(f)z - f$. Notice that $L(v) = L(L(f)z - f) = L(f)L(z) - L(f) = 0$, so $v \in Z$, and is thus orthogonal to z (remember, $z \in Z^{\perp}$). We have shown that

$$\begin{aligned} & L(f)z - f \perp z \\ \Rightarrow \quad & \langle L(f)z - f, z \rangle = 0. \end{aligned}$$

[3]Recall that a continuous function may be defined as one for which $f^{-1}(O)$ is open whenever O is open (and thus $f^{-1}(C)$ is closed whenever C is closed) [2].

This is starting to look good! We can rearrange $\langle L(f)z - f, z\rangle = 0$ to

$$\begin{aligned} & L(f)\langle z, z\rangle - \langle f, z\rangle = 0 \\ \Rightarrow \quad & L(f)||z||^2 = \langle f, z\rangle \\ \Rightarrow \quad & L(f) = \tfrac{\langle f,z\rangle}{||z||^2} = \langle f, \tfrac{z}{||z||^2}\rangle. \end{aligned}$$

Then the element k we are looking for is given by $k = \frac{z}{||z||^2}$. To show that k is unique, suppose k' satisfies $L(f) = \langle f, k'\rangle$ for all $f \in \mathcal{H}$. Then

$$\begin{aligned} & \langle f, k\rangle = \langle f, k'\rangle \\ \Rightarrow \quad & \langle f, k - k'\rangle = 0. \end{aligned}$$

Since $\langle \cdot, \cdot \rangle$ is positive definite, $k - k' = 0$ and $k' = k$.

□

Take a look back at Example 2.1. There, we said that, given any linear function f which maps $\mathbb{R}^n$ to $\mathbb{R}$, we can find a vector $\vec{x}'$ whose dot product imitates the action of f: $f(\vec{x}) = \vec{x} \cdot \vec{x}'$. This is a simple case of the basic linear algebra theorem that any linear transformation can be represented as multiplication by some matrix.

Here, we're saying basically the same thing. Given any linear function L which maps $\mathcal{H}$ to $\mathbb{R}$, we can find a vector $k \in \mathcal{H}$ whose inner product reproduces the action of L: $L(f) = \langle f, k\rangle$. This **reproducing property** is the key to the reproducing kernels which we're after.

3 Reproducing Kernel Hilbert Spaces

Now we're ready to introduce the headliner.

Definition 3.1. A **Reproducing Kernel Hilbert Space (RKHS)** is a Hilbert space of functions in which the evaluation functional is continuous (or bounded, equivalently).

Specifically, let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Then for each $x \in \mathcal{X}$ the evaluation functional $\mathcal{L}_x : \mathcal{H} \to \mathbb{R}$ is defined by

$$\mathcal{L}_x(f) = f(x),$$

and $\mathcal{H}$ is a **RKHS** if $\mathcal{L}_x(f)$ is pointwise continuous at every $f \in \mathcal{H}$ for every $x \in \mathcal{X}$.

Colloquially, a Hilbert space of functions is an RKHS if point evaluation is a continuous linear functional. This means that if two functions $f, g \in \mathcal{H}$ are close in norm ($||f - g||_{\mathcal{H}}$ is small) then they must be close at every point in their shared domain $\mathcal{X}$ ($|f(x) - g(x)|$ is small $\forall x \in \mathcal{X}$).

As mentioned in Example 2.2, since the elements of L^2 spaces are actually equivalence classes of function modulo differences on sets of measure zero, it is not coherent to talk about "point evaluation" as defined above in the context of L^2 spaces. They are not RKHS.

3.1 Reproducing Kernels

Definition 3.2. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **reproducing kernel** of $\mathcal{H}$ if:

1. For every $y \in \mathcal{X}$, $K(\cdot, y) \equiv K_y(\cdot) \in \mathcal{H}$, and
2. For all $f \in \mathcal{H}$ and $x \in \mathcal{X}$,
$$f(x) = \langle f, K_x \rangle. \tag{1}$$

Property 2 above is the reproducing property we talked about in the context of the Reisz representation theorem (Theorem 2.3) for the special case where the linear operator is the evaluation functional of $\mathcal{H}$. That is to say, K "reproduces" the action of the evaluation functional. The Reisz representation theorem thus ensures that Hilbert spaces with continuous evaluation functionals–RKHS–have reproducing kernels. The converse is also true, admitting an alternative definition for RKHS. This is the definition for which they are named.

Proposition 3.1. *Let $\mathcal{H}$ be a Hilbert space of functions on some set $\mathcal{X}$. The following are equivalent:*

- *$\mathcal{H}$ has a unique reproducing kernel.*
- *Point evaluation is a continuous linear functional in $\mathcal{H}$, as stated in Definition 3.1.*

Proof. Choose arbitrary $y \in \mathcal{X}$. Then if K is the reproducing kernel for $\mathcal{H}$,

$$|f(y)| = |\langle f, K_y \rangle|,$$

which by Cauchy-Schwarz is

$$\begin{aligned} |f(y)| &\leq \langle f, f \rangle^{\frac{1}{2}} \langle K_y, K_y \rangle^{\frac{1}{2}} \\ &= ||f||_{\mathcal{H}} K_y(y)^{\frac{1}{2}} = ||f||_{\mathcal{H}} K(y, y)^{\frac{1}{2}}. \end{aligned}$$

$K(y, y)^{\frac{1}{2}}$ is just a number, so point evaluation is a bounded functional.

Now, suppose that $\mathcal{L}_x$ is continuous. Then the Reisz representation theorem gives us a $K_x \in \mathcal{H}$ such that $\mathcal{L}_x(f) = f(x) = \langle f, K_x \rangle$. Since K_x is itself a function in $\mathcal{H}$, if we consider the evaluation functional of some other $z \in \mathcal{H}$, we can define

$$K(x, z) \equiv \mathcal{L}_x(K_z) = \langle K_z, K_x \rangle = K_z(x),$$

the reproducing kernel we were looking for.

In both directions of the proof, the uniqueness of K follows from the uniqueness of the reproducing k in Theorem 2.3. □

3.2 Positive Definite Functions

Definition 3.3. A function $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is **positive definite** if it is symmetric and for all $N \in \mathbb{N}$, $x_1, x_2, ..., x_N \in \mathcal{X}$, and $a_1, a_2, ..., a_N \in \mathbb{R}$,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j F(x_i, x_j) \geq 0,$$

where the equality only holds if $a_1 = a_2 = ... = a_N = 0$.

Note 3.1 (Notation)**.** Following the usage in Definition 3.2, throughout this paper we will write, for example, $K_x = K_x(\cdot) \equiv K(\cdot, x)$. This hints at the conceptual relationship between the function $K_x(\cdot)$ when $\mathcal{X}$ has infinite cardinality and the column vector $\vec{K_x}$ when $\mathcal{X}$ has finite cardinality (in which case K may be thought of simply as a matrix), a topic which will be discussed in Section 3.4.

Positive definite functions have a bijective correspondence with RKHS. This is often viewed as the key fact of RKHS theory, leading to its characterization as a transform theory.

Theorem 3.2 (**Moore-Aronszajn Theorem**, ([4], p. 344), ([5], p. 2-3))**.** *To every RKHS there corresponds a unique positive definite kernel, and every positive definite function is the unique reproducing kernel for some RKHS.*

Proof. Given a RKHS $\mathcal{H}$, Proposition 3.1 tells us that $\mathcal{H}$ possesses a unique reproducing kernel K. This kernel is positive definite, because

$$\begin{aligned} \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(x_i, x_j) &= \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \langle K_{x_i}, K_{x_j} \rangle_{\mathcal{H}} \\ &= \left\langle \textstyle\sum_{i=1}^{N} a_i K_{x_i}, \sum_{j=1}^{N} a_j K_{x_j} \right\rangle_{\mathcal{H}} = \left|\left| \textstyle\sum_{i=1}^{N} a_i K_{x_i} \right|\right|_{\mathcal{H}}^2 \end{aligned}$$

and inner products are positive definite.

More profoundly, *every* positive definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the unique reproducing kernel for some corresponding RKHS $\mathcal{H}_K$ of functions $f : \mathcal{X} \to \mathbb{R}$. To construct this space, we'll start with the set $\mathcal{F}$ of all functions of the form

$$f(\cdot) = \sum_i a_i K(\cdot, x_i), \tag{2}$$

that is, arbitrary linear combinations of K. Since K is positive definite, we can define the inner product

$$\langle f, g \rangle_{\mathcal{F}} = \left\langle \sum_i a_i K(\cdot, x_i), \sum_j b_j K(\cdot, x_j) \right\rangle_{\mathcal{F}} \equiv \sum_i \sum_j a_i b_j K(x_i, x_j),$$

which induces the norm

$$||f||_{\mathcal{F}} = \left|\left|\sum_i a_i K(\cdot, x_i)\right|\right|_{\mathcal{F}}^2 \equiv \sum_i \sum_j a_i a_j K(x_i, x_j)$$

and forms an inner product space. The positive definiteness and symmetry of K makes these expressions satisfy the definitions of inner product and norm.

While this set of functions[4] is not yet a Hilbert space, we can complete it by adding to $\mathcal{F}$ the pointwise limits of Cauchy sequences contained in $\mathcal{F}$. The set already satisfies the reproducing property by construction, which we can use to show that norm convergence of any sequence $f_n \in \mathcal{F}$ implies pointwise convergence:

$$|f_n(x) - f_m(x)| = |\langle f_n - f_m, K_x \rangle_{\mathcal{F}}| \leq ||f_n - f_m||_{\mathcal{F}} ||K_x||_{\mathcal{F}},$$

which is basically the boundedness of point evaluation. Thus the limit of any Cauchy sequence will be well-defined as a pointwise limit, which also guarantees that the reproducing property carries over to the completion.

□

3.3 Kernel Functions

Let us summarize our discussion of RKHS theory to this point. We have established a bijective correspondence between spaces with continuous evaluation functionals and spaces with reproducing kernels. We have also established that positive definite functions and reproducing kernels are one and the same thing; every positive definite function is the unique reproducing kernel for some RKHS, and every reproducing kernel is a positive definite function. For the rest of this paper, then, we might refer to *kernel functions*: the bivariate functions which form the core characterization of RKHS, which we know to be positive definite, symmetric, and in possession of the reproducing property.

The connection between a RKHS and its kernel function is quite deep. As we saw in the proof of Theorem 3.2, every RKHS is the completion of the span of the kernel. Therefore, any function in an RKHS may be expressed as a linear combination of kernels or the pointwise limit of a sequence of such linear combinations.

3.4 Finite-dimensional RKHS

In order to understand the preceding discussion of RKHS theory, it will help to consider the case in which $\mathcal{H} \subseteq \mathbb{R}^n$. As summarized in Section 3.3, the theory we've developed so far describes a correspondence between positive definite kernel functions and RKHS. How does this correspondence look in finite dimensions?

[4] Specifically, $\mathcal{F}$ is a linear manifold [5].

First we have to decide what the underlying domain $\mathcal{X}$ even is and figure out how we can think of elements of $\mathbb{R}^n$ as functions. There's an easy, canonical correspondence between $\vec{v} = (v_1, v_2, ..., v_n)^T \in \mathbb{R}^n$ and functions $f : \{1, 2, ..., n\} \to \mathbb{R}$. Simply let $f(i) = v_i$. So we may think of $\mathbb{R}^n$ as a Hilbert space of real-valued functions on $\mathcal{X} = \{1, 2, ..., n\}$.

Remark 3.1. If $\mathcal{X}$ is any set of finite cardinality n, then the set $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ may be thought of as a subpsace of $\mathbb{R}^n$ in the same style. So while we use the language of Euclidean space, this discussion can actually be generalized to all spaces of functions $f : \mathcal{X} \to \mathbb{R}$ with finite $\mathcal{X}$.

Now, consider an inner product space $V \subseteq \mathbb{R}^n$. Since all finite dimensional normed spaces are complete, V is a Hilbert space. In fact, since elements of V are vectors $v \in \mathbb{R}^n$, the evaluation functionals from Definition 3.1 just index these vectors: $\mathcal{L}_i(v) = v_i$. $\mathcal{L}_i$ is clearly continuous (for a proof see Example 2.3), so V is also a RKHS. In $\mathbb{R}^n$, inner product spaces, Hilbert spaces, and RKHS are all the same thing.

This equivalence is illustrated by the following proposition.

Proposition 3.3. *A function $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an inner product for a space $V \subseteq \mathbb{R}^n$ if and only if there exists a symmetric, positive definite matrix M such that $\langle v_1, v_2 \rangle = v_1^T M v_2$ for all $v_1, v_2 \in V$.*

This follows from the fact that linear transformations in Euclidean space can be represented by matrices. Therefore, the linear transformation from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ represented by an inner product corresponds to a matrix, which must be symmetric and positive definite since inner products are symmetric and positive definite.

How does the M corresponding to the inner product relate to the kernel of the space? Applying Definition 3.2 to the finite case,

Definition 3.4 (Kernels of Finite RKHS)**.** The **kernel** of an inner product space $V \subseteq \mathbb{R}^n$ is the unique $n \times n$ matrix K with columns K_i such that:

- For all $i \in [n] = \{1, ..., n\}$, $K_i \in V$ and
- for all $v \in V$ and $i \in [n]$, $\langle v, K_i \rangle = v_i$.

The reproducing property can then be written

$$\begin{aligned} & \langle v, K_i \rangle = v^T M K_i = v_i \\ \Rightarrow \quad & M^{-1}(v^T)^{-1} v_i = K_i \\ \Rightarrow \quad & M_i^{-1} = K_i, \end{aligned}$$

which is to say that $K = M^{-1}$. The idea is that the kernel function is intrinsically connected to, and arising from, the inner product of the space. This is true in infinite dimensions, as well: property 2 of Definition 3.2 encodes this fact. The explicit story is a bit more complicated, but the connection persists.

Example 3.1 ([17], p. 12)**.** Let $K = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. Every vector in the inner product space V_K corresponding to K may be written

$$\begin{aligned} v &= \sum_i a_i K(\cdot, x_i) \\ &= \sum_i a_i K_1 + \sum_i a_i K_2 \\ &= \sum_i a_i \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \sum_i a_i \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned}$$

$V_K = \text{span}([1,0]^T, [0,0]^T)$ with inner product $\langle v_1, v_2 \rangle_{V_K} = v_1^T \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} v_2$, making V_K the plane in $\mathbb{R}^2$ with $v_2 = 0$.

What we have been discussing is a bijective correspondence between positive definite matrices and inner product spaces in $\mathbb{R}^n$. Basic linear algebra tells us that two vector subspaces U and V of $\mathbb{R}^n$ having the same dimension are isomorphic, that is, there exists a structure-preserving bijection between the two spaces. However, if the inner products of U and V differ, then they will have different kernels. This difference tells us about the particular ways U and V are embedded within larger Euclidean space $\mathbb{R}^n$, that is, their *extrinsic geometry*. From one perspective, what's valuable about kernels is that they completely encode information about an inner product space's intrinsic *and* extrinsic properties. If we know a vector $v \in V$ where $V \subseteq \mathbb{R}^n$ is an inner product space, then RKHS theory allows us to write v as a linear combination of kernels–that is, as a vector in $\mathbb{R}^n$, capturing v's extrinsic properties. For more discussion of this topic, see [17]. The characterization of RKHS arising from the kernel captures completely and elegantly properties of how RKHS are embedded in a larger function space, and it may be argued that this is in some sense the "proper" way to view RKHS.

3.5 Examples

Example 3.2 (Euclidean Space)**.** Consider $\mathcal{H} = \mathbb{R}^n$. There are exactly n evaluation functionals, $\mathcal{L}_1(\vec{x}) = x_1$, $\mathcal{L}_2(\vec{x}) = x_2$, and so on. To find the kernel, we need there to be $\vec{K_i} \in \mathbb{R}^n$ such that

$$\mathcal{L}_i(\vec{x}) = x_i = \langle \vec{x}, \vec{K_i} \rangle.$$

This is cleary satisfied by the canonical basis vector $\vec{K_i} = \vec{e_i}$, the vector with zeros in every index but i. The reproducing kernel of $\mathbb{R}^n$ satisfies $K(i,j) = 1$ if $i = j$ and $K(i,j) = 0$ if $i \neq j$. It's the identity matrix.

Example 3.3 (Bergman Space)**.** A Bergman space is a particular kind of subspace of a Lebesgue space with complex domain.[5] For example, as described in [1], the Bergman

[5]Specifically, the domain must be an open and connected subset of $\mathbb{C}$ [22].

space $L^2_a(\mathbb{D})$ on the unit disk $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ consists of all analytic functions on $\mathbb{D}$ such that

$$\int_{\mathbb{D}} |f(z)|^2 dA < \infty.$$

That is, $L^2_a(\mathbb{D})$ consists of the square-integrable analytic functions on $\mathbb{D}$. That makes it the subset of $L^2(\mathbb{D})$ which contains only analytic functions, hence the notation.

Theorem 3.4. *$L^2_a(\mathbb{D})$ is a reproducing kernel Hilbert space.*

We won't offer a proof of this fact, but we'll sketch out a few results which are necessary to the proof and show why the evaluation functionals of this space are bounded.

One key result from complex analysis known as Cauchy's integral formula, a good exposition of which can be found in [3], can be extended relatively easily[6] to provide an analogy of the mean value theorem on the real line for analytic functions on a disk:

Proposition 3.5. *For an analytic function f in a closed disk $\overline{B(a,R)}$,*

$$f(a) = \frac{1}{\pi R^2} \int_{\overline{B(a,R)}} f \, dA$$

This asserts that the value of f at the center of the disk is proportional to the area integral of f on the disk. It also equips us to to place a relative upper bound on the pointwise evaluation of a function $f \in L^2_a(\mathbb{D})$.

Corollary 3.5.1. *Let $x \in \mathbb{D}$ and $f \in L^2(\mathbb{D})$. Then*

$$|f(x)| \leq \frac{1}{1-|x|} ||f||_{L^2_a(\mathbb{D})}$$

for every $f \in \mathbb{D}$.

The details of the proof are also in [1]; the thrust is to start with Proposition 3.5 and use Hölder's inequality to relate $\left|\int_{\overline{B(x,R)}} f \, dA\right|$ with $\left(\int_{\mathbb{D}} |f|^2 dA\right)^2 = ||f||_{L^2_a}(\mathbb{D})$. This result is necessary to prove that $L^2_a(\mathbb{D})$ is a Hilbert space at all, but it's important to us because we've just placed an bound on the evaluation functional $\mathcal{L}_x$ of $L^2_a(\mathbb{D})$ with respect to $||\cdot||_{L^2_a(\mathbb{D})}$, proving that the evaluation functional is bounded. If $L^2_a(\mathbb{D})$ is a Hilbert space (a fact which we'll state without proof), then, it's also a RKHS.

The reproducing kernel for $L^2(\mathbb{D})$ is given by $K_x(y) = \frac{1}{(1-y\overline{x})^2}$ [20], which can be verified by showing that $\langle f, K_x \rangle = \int_{\mathbb{D}} \frac{f(y)}{(1-\overline{y}x)^2} \frac{dA}{\pi} = f(x)$ using Proposition 3.5.

[6] See page 9 of [1].

4 Statistical Learning Theory

4.1 The Learning Problem

Suppose there is a set of variables X associated probabilistically with some variable Y. Each element in X is not associated uniquely with an element of Y; rather, each element of X determines a probability distribution on Y. We have a probability distribution $P(z)$ on $Z = X \times Y$. We can write $P(z) = P(x, y) = P(x)P(y|x)$, that is, $P(z)$ is the product of the marginal probability of x with the conditional probability of y given x. Since we're posing a learning problem, we'll take this $P(z)$ to be unknown. We'd like to be able to "predict" values in Y given data in X.

Now, imagine we have a set S of samples drawn independently from this underlying distribution. Given any $x \in X$, we'd like to be able to predict the value of $y \in Y$ most likely associated with that x. This is the problem of **supervised learning**. A learning algorithm looks at the training set S and determines a function which can be used to make a prediction $\hat{y}$ given some new, previously unseen data x_{new}:

$$\hat{y} = f_S(x_{new}).$$

Hopefully, we can "train" our function on our training data S to produce an *estimator* function f_S which is good at predicting values it hasn't seen before: an f_S which *generalizes* well.

We pick our function f_S from a set of candidates which we'll call a **hypothesis space** $\mathcal{H}$. In order to choose f_S, we need some way of measuring how good any arbitrary function $f \in \mathcal{H}$ is at predicting y given x. We do this by comparing the predictions $f(x)$ with the true values y using a **loss function** $V(f(x), y)$. Throughout this paper, we'll assume that $V : \mathbb{R}^2 \to \mathbb{R}$. This is a common assumption, but need not always be the case, e.g. for categorical, nonbinary Y. V is similar in concept to a metric; for example, it might be the L_2-loss,

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

or the L_1-loss,

$$V(f(x), y) = |f(x) - y|.$$

Now, we're interested in minimizing the "distance," as measured by our loss function, between the relationship described by our estimator and the true relationship $P(x, y)$. This is called the *expected risk*:

$$I(f) \equiv \int_{X,Y} V(f(x), y)P(x, y)dxdy.$$

The ideal estimator f_{ideal} minimizes the expected risk:

$$f_{ideal}(x) = \arg\min_{f \in \mathcal{H}} I(f).^{7}$$

The expression for $I(f)$ contains $P(x, y)$, which we've assumed was undefined. Thus, we are unable to find f_{ideal} in practice. Instead, we'll approximate the expected risk using our training data S. If $S = z_1, z_2, ..., z_N = (x_1, y_1), ..., (x_N, y_N)$, then the *empirical risk* functional is:

$$I_S(f) \equiv \frac{1}{N} \sum_{i=1}^{N} V(f(x_i), y_i), \tag{3}$$

and the empirical risk minimization (ERM) problem is to find:

$$\arg\min_{f \in \mathcal{H}} I_S(f). \tag{4}$$

Much of statistical learning theory seeks to answer the question of when the minimizer of the empirical risk is a good approximation for the minimizer of expected risk. Formally, under what conditions does

$$\lim_{N \to \infty} I_S(\hat{f}_S) = \lim_{N \to \infty} I(\hat{f}_S) = I(f_{ideal}),$$

where $\hat{f}_S$ is the minimizer of the empirical risk on the training set S. Good answers to this question have been found under pretty general conditions, and constitute much of the work of Vapnik in developing the modern field of statistical learning theory [10]. A good, more recent overview can be found at [9]. Treating this material properly is not within the scope of this paper; instead, we'll introduce a class of models which arises from the theory of ERM, and consider what happens when we consider these models in the context of RKHS.

4.2 Regularization

This is the most informal section of the paper, but it serves the important purpose of building intuition as to why we're interested in a particular class of problems (to be introduced in Section 4.2.2) which modify Equation 4. A formal discussion of regularization would take us too far afield and is too tangentially related to the RKHS material we're interested in to be included. Instead, we'll discuss these issues just enough to build an intuitive picture of how the class of optimization problems we're interested in arises.

[7] The "arg min" function indicates that we're seeking the argument f which minimizes the functional I_S.

4.2.1 Ill-Posed Problems

A problem is considered *well-posed* [19] if it satisfies three conditions:

1. The solution exists.
2. The solution is unique.
3. The solution is *stable*, that is, it depends continuously on the initial conditions.

In a general hypothesis space, the ERM problem (4) is ill-posed [24]. We won't go into all the ways these conditions can be violated, nor will we rigorously describe the conditions under which they are satisfied. Instead, we'll give an example drawn from [24] of a case where condition 3 is violated.

Consider a training set of 10 points as shown on the left in figure 2. An ERM algorithm restricted to polynomials of degree nine might return the estimator function shown on the right, with zero empirical error.

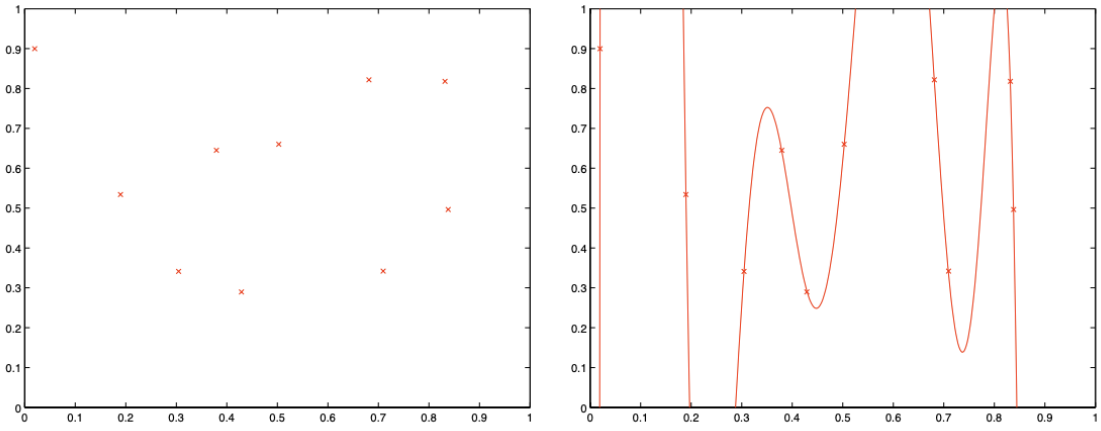

Figure 2: Left: a training set of 10 observations. Right: a degree-9 polynomial estimator.

However, if we perturb the training data just slightly, in order to fit a polynomial with zero empirical error the algorithm will have to return the function shown by the dotted blue line on the right of Figure 3.

In this case, ERM reduced our empirical risk to zero. However, tiny changes in our input data caused the estimator to change quite dramatically. The solutions to ERM are *unstable* if our hypothesis space consists of polynomials of degree 9. This instability corresponds to discontinuous dependence of the minimizer $\hat{f}_S$ of the functional $I_S(f)$ on the training set S. For stability, we need the linear functional $I_S : \mathcal{X} \times Y \to \mathcal{H}$ to be bounded. In this case, it is not, and the ERM problem (4) is ill-posed.

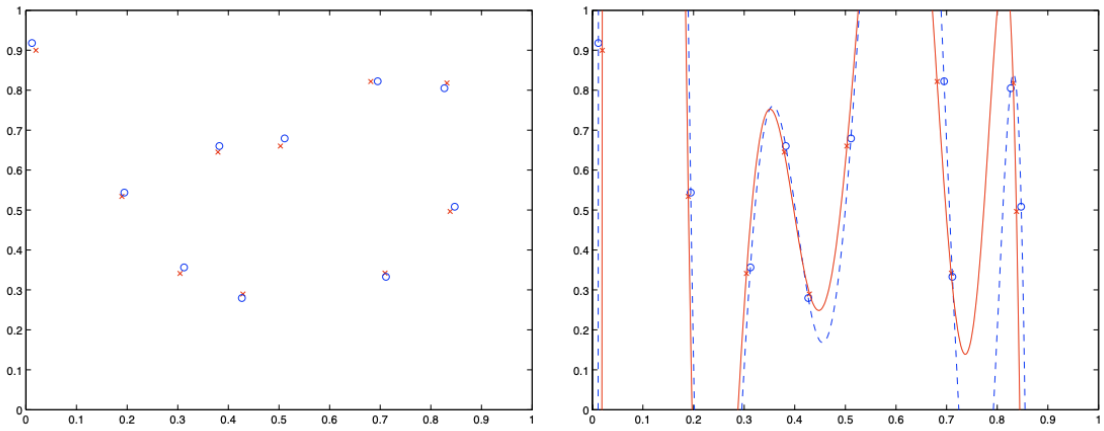

Figure 3: Left: the training set slightly perturbed. Right: a much different fit.

In Figure 4, we manually restrict our hypothesis space to polynomials of degree 2. ERM is no longer able to fit a curve with zero empirical risk, but we solve the problem of instability.

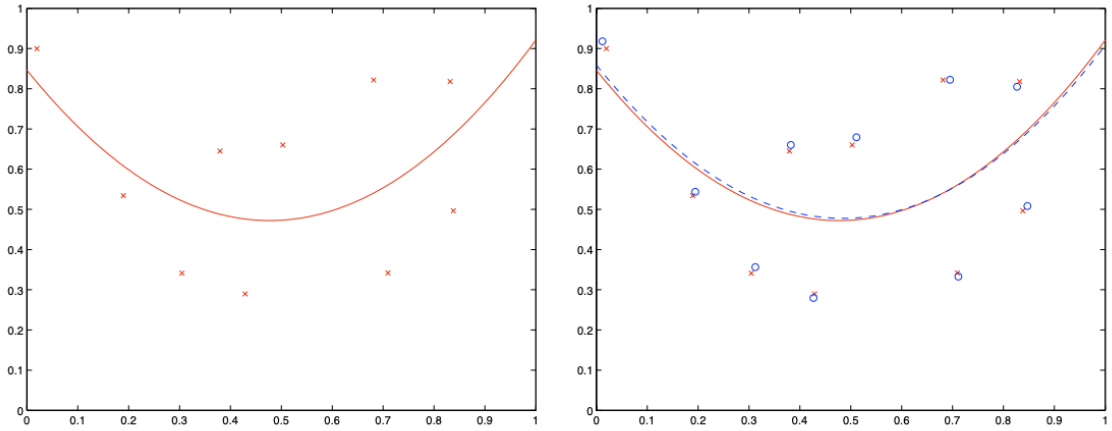

Figure 4: Left: a quadratic fit to the training set. Right: a quadratic fit to the perturbed training set.

There's a fundamental trade-off between model *bias*, how well a model matches the training set, and model *variance*, how much the model changes with the training data. Formally, balancing bias and variance corresponds to ERM on a stable problem.

4.2.2 Adding a Penalty Term

The example in the last section may also be thought of as a trade-off between model complexity and model stability. What we saw is that if we restrict our attention to a less-complex subspace (the space of second degree polynomials) of the full hypothesis space, then we can sidestep the issue of instability.

Instead of solving the original ERM problem (Equation 4), we can add a "cost" for complexity, represented by a penalty term $P(f)$. Our new optimization problem is:

$$\arg\min_{f \in \mathcal{H}} \left[\frac{1}{N} \sum_{i=1}^{N} V(f(x_i), y_i) + \lambda P(f) \right], \tag{5}$$

where $\mathcal{H}$ is the space of functions for which $P(f)$ is defined. Equation (5) turns out to be useful in a very wide range of settings [14]. While we're stopping short of an in-depth or rigorous discussion here, what's key is that the introduction of the penalty term limits the possible minimizers (which we'll take as our estimator function), and therefore can avoid the kind of instability that we saw in Figure 3–what is sometimes called the problem of overfitting. A rigorous theoretical justification of regularized ERM as seen in Equation 5, how problems of this form come to be well-posed, may be found at [9].

5 Regularization and RKHS

5.1 RKHS and Smoothness

As discussed in the last section, the goal of regularization is ultimately to state empirical risk minimization as a well-posed problem. In particular, we're focused on ensuring that the minimizer found depends continuously on the training data. One way to accomplish this is by explictly restricting the hypothesis space, and another (more popular) way is to add a penalty term which encodes a preference for functions which are "smoother" or "less complex," though these ideas may be defined in a number of different ways [6].

No matter how we're formalizing our notion of smoothness,[8] however, it always has to do with point evaluation, since our training set S consists of particular points in $\mathcal{X} \times Y$. Every definition of smoothness must relate directly to the values of the candidate functions f at each point in $\mathcal{X}$, or else it'll be meaningless in the ERM setting. Also, in order to induce a preference for some functions in $\mathcal{H}$ we need to be able to compare them and evaluate their similarity, which we accomplish via an inner product and its norm. All of this is to say, we need point evaluation $f(x)$ of functions $f \in \mathcal{H}$ to be bounded with regard to their norm $||f||_{\mathcal{H}}$. This is exactly the requirement that point evaluation be continuous in Definition

[8] Throughout this discussion we'll use the word "smoothness" to refer to the quality penalized by $J(f)$. Note that this property does not always correspond to smoothness in the formal sense, and might therefore be referred to as, for example, "flatness" or "complexity" in other contexts.

3.1; which is to say that *the appropriate space in which to conduct regularization via a loss-penalty formulation is a RKHS*. The penalty term would not be able to meaningfully compare the smoothness of functions across their entire domain without an inner product and the continuity of point evaluation functionals with respect to this inner product's norm. By the same token, RKHS may be viewed as function spaces in which smoothness is well-defined, an idea which can be formalized [6].

5.2 Regularization Operators

With this in mind, it turns out that for most of the problems we're interested in, the penalty $P(f)$ may be written as the square norm of a linear operator with its codomain in an inner product space: $P(f) = ||R(f)||^2$ [8] [9]. Unsurprisingly, if R is a positive definite operator, we can find a specific kernel K corresponding to it [8].

Example 5.1 (Regularization Operators and Kernels in $\mathbb{R}^n$ ([13], p. 3273-3274))**.** In finite dimensions, where operators are matrices, the correspondence between K and R is very straightforward. Given R, we may define $K = (R^T R)^{-1}$. Then by Proposition 3.3, there's an inner product space $V_K \subseteq \mathbb{R}^n$ such that for $v \in V_K$, $||v||^2_{V_K} = \langle v, v \rangle_{V_K} = v^T K^{-1} v = v^T R^T R v = ||Rv||_{\mathbb{R}^n}$.

This story is somewhat more involved in the general setting, but it draws attention to a fact whose significance should not be missed: given a particular regularization problem, there is a *correct* kernel to use. Likewise, we can choose any positive definite kernel, along with a loss function, and pose a problem in the form of Equation 5.

From here on out, we'll take the penalty term to be the square norm in a RKHS, $P(f) = ||f||^2_{\mathcal{H}_K}$. Equation 5 becomes:

$$\arg\min_{f \in \mathcal{H}_K} \left[\frac{1}{N} \sum_{i=1}^{N} V(f(x_i), y_i) + \lambda ||f||^2_{\mathcal{H}_K} \right], \tag{6}$$

Recall from the construction of $\mathcal{H}_K$ from K offered in the proof of Theorem 3.2 that $||f||_{\mathcal{H}_K}$ is defined by $||f(x)||^2_{\mathcal{H}_K} = ||\sum_k a_k K(x, y_k)||^2_{\mathcal{H}_K} \equiv \sum_i \sum_j a_i a_j K(y_i, y_j)$ for some countable set of coefficients a_i and points $y_i \in \mathcal{X}$ (we can see how this quantity might correspond to the "complexity" or "smoothness" of the function f). Likewise $f(x)$ is the limit $\sum_i a_i K(x, y_i)$, so we can write Equation 6 in terms of the kernel:

$$\arg\min_{\{a_1, a_2, \ldots\} \in \mathbb{R}^N, \{y_1, y_2, \ldots\} \in \mathcal{X}} \left[\frac{1}{N} \sum_{i=1}^{N} V \left(\sum_j a_j K(x_i, y_j), y_i \right) + \lambda \sum_j \sum_k a_j a_k K(y_j, y_k) \right]. \tag{7}$$

We have posed the class of loss-penalty ERM problems in terms of kernels. Equation 7 still isn't tractable, however; we have no way of finding the infinite number of coefficients a_i nor the points y_i which minimize the functional.

5.3 Representer Theorem

The following theorem is the core theorem of this paper. While Section 5.1 discusses why it is natural to restrict machine learning problems' hypothesis spaces to RKHS, the representer theorem is the primary theoretical reason for using kernel representations explicitly in practice. It shows that the function which minimizes Equations 6-7 may be written as a *finite* linear combination of the reproducing kernels K_{x_i} of the training points, where arbitrary functions in $\mathcal{H}_K$ might be represented by infinite series of reproducing kernels of any points in $\mathcal{X}$. Its first version appeared in Wahba's 1990 classic *Spline Models for Observational Data* [5], but the following form comes from Scholkopf and Smola [7].

Theorem 5.1 (**Representer Theorem**, ([7], p. 90-91))**.** *Let Y be a set and $V : Y \times Y \to \{\mathbb{R} \cup \infty\}$ be an arbitrary loss function. Then each minimizer $\hat{f}$ of the the regularized empirical risk*

$$\sum_{i=1}^{N} \frac{1}{N} V(f(x_i), y_i) + \lambda ||f||^2_{\mathcal{H}_K} \tag{8}$$

may be represented as a finite linear combination of kernels K:

$$\hat{f}(x) = \sum_{i=1}^{N} a_i K(x, x_i). \tag{9}$$

Proof. First, any function $f \in \mathcal{H}_K$ may be written as a sum of the part contained in the span of the kernel functions on the training points $x_1, ..., x_m$ and the part contained in its orthogonal complement:

$$f(x) = \sum_{i=1}^{N} a_i K(x, x_i) + f^{\perp}(x),$$

By the definition of orthogonal complement, $\langle f^{\perp}(x), K(x, x_i) \rangle = 0$ for all $i \in [N] \equiv \{1, ..., N\}$, $x \in X$.

By the reproducing property (Equation 1), $f(x_i)$ for each training point $i \in [N]$ may be written in terms of reproducing kernels as

$$\begin{aligned} f(x_i) &= \langle f(\cdot), K(\cdot, x_i) \rangle = \left\langle \textstyle\sum_{j=1}^{N} a_j K(\cdot, x_j), K(\cdot, x_i) \right\rangle + \langle f^{\perp}(\cdot), K(\cdot, x_i) \rangle \\ &= \textstyle\sum_{j=1}^{N} a_j K(x_i, x_j). \end{aligned}$$

That is, *all* functions in $\mathcal{H}_K$, when evaluated at a training point (which may be any point in the domain X!), are equal to a linear combination of the kernel evaluated at the training points.

Also, notice that for any function $f \in \mathcal{H}_K$,

$$||f||^2_{\mathcal{H}_K} = \left|\left|\sum_{i=1}^{N} a_i K(\cdot, x_i)\right|\right|^2_{\mathcal{H}_K} + ||f^{\perp}||^2_{\mathcal{H}_K} \geq \left|\left|\sum_{i=1}^{N} a_i K(\cdot, x_i)\right|\right|^2_{\mathcal{H}_K}.$$

To summarize, we can write Equation 8 in the following form:

$$\frac{1}{N}\sum_{i=1}^{N} V\left(\sum_{j=1}^{N} a_j K(x_i, x_j), y_i\right) + \left|\left|\sum_{i=1}^{N} a_i K(\cdot, x_i)\right|\right|^2_{\mathcal{H}_K} + ||f^{\perp}||^2_{\mathcal{H}_K},$$

which is clearly minimized when $f^{\perp} = 0$. Any minimizer $\hat{f}(x)$ of Equation 8 thus has the form $\hat{f}(x) = \sum_{i=1}^{N} a_i K(x, x_i)$.

□

We've proven that the minimizer of Equations 6-7 is a sum of kernels *on the training points*. Now we can finally rephrase the empirical risk functional as a finite-dimensional optimization problem! Following the treatment in [14] (pp. 169) let $\mathbf{K}$ be the matrix representation of K on the training set, that is, the $N \times N$ matrix with entries $K_{ij} = K(x_i, x_j)$. Likewise, let $\mathbf{y} \in \mathbb{R}^N$ be the vector with entries $(y_1, ..., y_N)$, and $\mathbf{a} \in \mathbb{R}^N$ be the vector with entries $(a_1, ..., a_N)$. Equation 7 can be rewritten:

$$\arg\min_{\mathbf{a}} V(\mathbf{y}, \mathbf{Ka}) + \lambda \mathbf{a}^T \mathbf{Ka}.$$

This is a problem which can be optimized using standard numerical algorithms. We've done what we set out to do–converted an intractable regularization problem into an optimization problem over finite dimensions.

Remark 5.1. A message of this theorem is that arbitrarily complex functions cannot be estimated using ERM on a finite training set; the estimated function is always confined to a subspace of the original hypothesis space with maximum dimension N. This actually holds whether we're using a penalty term or not. By Theorem 5.1, the complexity of the $\hat{f}$ selected is restricted by N, regardless of what $\mathcal{X}$ and $\mathcal{H}$ look like.

6 Conclusion

The purpose of this paper is to bridge the gap between RKHS and their usage in machine learning in an accessible, yet relatively thorough, way. To this end, in certain places I have sacrificed rigor to avoid large detours, and throughout I have omitted significant results, classes of examples, and so on. The connection between RKHS and statistical learning runs quite deep; the interested reader will have no trouble extending this inquiry.

A Support Vector Machines and Feature Maps

The narrative constructed in Sections 3-5 does not represent the historical development of the theory relating RKHS and machine learning. Because of this, the literature on the subject is quite scattered, and it can be difficult to relate the relatively succinct and elegant mathematical theory of RKHS to the way kernel methods are deployed in practice. This appendix attempts to bridge this gap via particular examples that appear in the machine learning literature.

A.1 Support Vector Machines

Kernel methods became part of the machine learning literature as part of the development of a class of supervised learning techniques called Support Vector Machines. In the simplest case, SVMs perform binary classification via separating hyperplanes–they separate categories of observations by drawing a line between them (Figure 5).

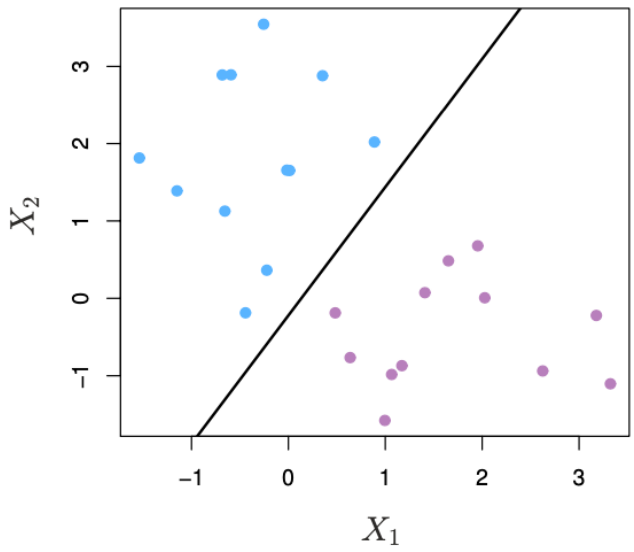

Figure 5: A separating hyperplane between blue and purple points in two dimenions [15].

Supposing that the input data is a vector of real numbers, $\mathcal{X} = \mathbb{R}^n$, this separating hyperplane takes the form

$$f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle + b,$$

where $\boldsymbol{a}$ is a vector of coefficients. That is to say, $f(\boldsymbol{x})$ is a line.

We want to choose the separating hyperplane that maximizes the margin between the nearest points on either side of the line. If the data is not perfectly linearly separable, we may want to use a "soft margin" which allows some points to lie on the wrong side of the line. Without going into the details, this corresponds to estimating f via ERM on the regularized risk functional

$$\frac{1}{N}\sum_{i=1}^{N} V(f(\boldsymbol{x_i}), y_i) + \lambda||\boldsymbol{a}||^2, \tag{10}$$

where $y_i \in \{-1, 1\}$ represents the color of the point and the loss function V is the hinge loss

$$V(f(\boldsymbol{x_i}), y_i) = \max[0, 1 - y_i(\langle \boldsymbol{a}, \boldsymbol{x_i} \rangle + b)],$$

pictured in Figure 6.

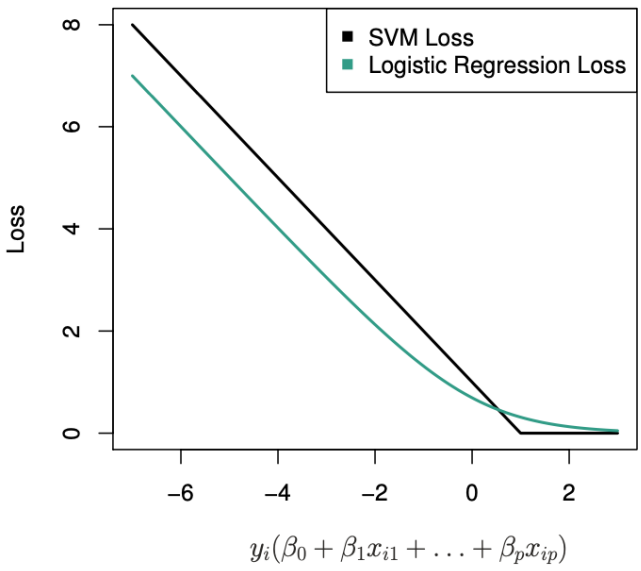

Figure 6: The hinge loss function plotted against the logistic regression loss function. [15].

Of course, this classification technique has the major weakness that it only works on data that is approximately linearly separable. The earliest way around this was by preprocessing the data into some **feature space** $\mathcal{F}$, for example, if $\mathcal{X} = \mathbb{R}^2$, via a **feature map** $\phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$\phi((x_1, x_2)^T) = (x_1, x_1^2 + x_2^2, x_2)^T.$$

Figure 7 shows the data as it appears in $\mathcal{F}$–significantly, it's much more linearly separable. We've chosen the feature map above to make the plot look good, but the same result can be achieved via the feature map $\phi((x_1, x_2)^T) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)^T$.

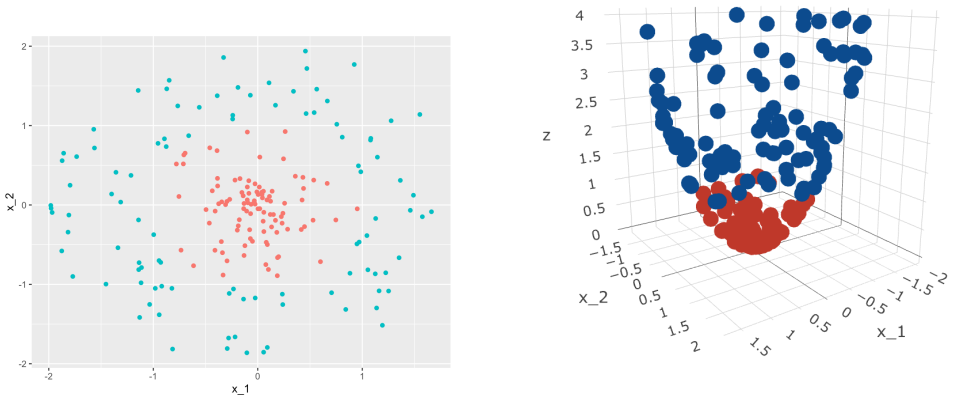

Figure 7: Left: the data in the input space. Right: the data in the feature space.

Now, a separating hyperplane in this six-dimensional feature space can be written

$$f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{\phi}(\boldsymbol{x_i}) \rangle + b,$$

where $\boldsymbol{a}$ is now in $\mathbb{R}^6$ rather than $\mathbb{R}^2$. Likewise Equation 10 is now written

$$\frac{1}{N} \sum_{i=1}^{N} \max[0, 1 - y_i(\langle \boldsymbol{a}, \boldsymbol{\phi}(\boldsymbol{x_i}) \rangle + b)] + \lambda ||\boldsymbol{a}||^2. \tag{11}$$

To this point, we've been carrying out our feature mapping explicitly as a preprocessing step, and that is the way this technique was invented. However, at this point we can define a bivariate function

$$\begin{aligned} K(x, y) &\equiv (1 + \textstyle\sum_{j=1}^{2} x_j y_j)^2 \\ &= 1 + 2\textstyle\sum_{j=1}^{2} x_j y_j + (\sum_{j=1}^{2} x_j y_j)^2 \\ &= 1 + 2x_1y_1 + 2x_2y_2 + (x_1y_1)^2 + (x_2y_2)^2 + 2(x_1y_1)(x_2y_2) \\ &= \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}y_1y_2 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \\ 1 \end{bmatrix} = \langle \phi(x), \phi(y) \rangle_{\mathbb{R}^6}. \end{aligned}$$

Using this clever correspondence, we can replace the inner product in Equation 11 with a bivariate function that is much cheaper to compute. This is an example of what is called

the **kernel trick** in the machine learning literature, and, historically, this is roughly how kernel methods made their entrance into the methods of ML practicioners.

A.2 Feature Maps

Given our knowledge of RKHS, we make the observation that $K(x, y) = (1 + \sum_{j=1}^{p} x_j y_j)^d$ is a positive definite functon, generally called the polynomial kernel. It corresponds to an RKHS $\mathcal{H}_K$ via $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_K}$. In fact, for any RKHS $\mathcal{H}_K$, we may define the feature map $\phi : \mathcal{X} \to \mathcal{H}_K$ with $\phi : x \mapsto K(\cdot, x)$. This leads to yet another way of understanding RKHS.

Proposition A.1 ([12], p. 2)**.** *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ for all $x, y \in \mathcal{X}$.*

RKHS may *always* be seen as mapping $\mathcal{X}$ into some higher-dimensional feature space of functions of the original inputs. These features may or may not have an easily discernable meaning in terms of the original problem, and in fact we often only reference them implicitly via kernel representations.

In the case of the polynomial kernel $K(x, y) = (1 + \sum_{j=1}^{p} x_j y_j)^d$, $\mathcal{H}_K$ is the space of polynomials of degree d in $\mathbb{R}^p$. This space happens to have $\binom{p+d}{d}$ eigenfunctions, given in Section A.1 by the ith element of the vector ϕ, which we can derive explicitly ([14], pp. 171).

Not all RKHS have such a basis, but many do (see [18]). Some sources simply assume that our kernel has a representation in terms of linearly independent functions ϕ_i:

$$K(x, y) \equiv \sum_{i=1}^{\infty} \gamma_i \phi_i(x) \phi_i(y), \tag{12}$$

where $\gamma_i \geq 0$ and $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$. A function thus defined is indeed positive definite. Then each function $f(x) \in \mathcal{H}$ can be written in terms of the eigenfunctions $\phi_i(x)$:

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x),$$

and the norm is

$$||f||_{\mathcal{H}_K} = \sum_{i=1}^{\infty} c_i^2 / \gamma_i.$$

Then Equations 6-7 become:

$$
\begin{aligned}
& \arg\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{N} V(f(x_i), y_i) + \lambda ||f||_{\mathcal{H}_K} \right] \\
= \; & \arg\min_{\{c_i\}_1^\infty} \left[\sum_{i=1}^{N} V\left(\sum_{i=1}^{\infty} c_i \phi_i(x), y_i \right) + \lambda \textstyle\sum_{i=1}^{\infty} c_i^2 / \gamma_i \right].
\end{aligned}
\tag{13}
$$

Example A.1 ([6], Appendix A.1)**.** Expression 12 presents a way to construct kernels and corresponding RKHS. We'll briefly sketch an example.

Let $\mathcal{X} = [0, 2\pi]$, and let $K(x) = \sum_{i=0}^{\infty} \lambda_i \cos(ix)$, the Fourier series of a continuous, symmetric, periodic function. Then we can use a trig identity and define

$$
K(x, y) \equiv K(x - y) = 1 + \sum_{i=1}^{\infty} \lambda_i \sin(ix) \sin(iy) + \sum_{i=1}^{\infty} \lambda_i \cos(ix) \cos(iy).
$$

This puts $K(x, y)$ in the form of Equation 12, where

$$
\{\phi_i(x)\}_{i=0}^{\infty} = (1, \sin(x), \cos(x), \sin(2x), \cos(2x), ...),
$$

so we can define a RKHS $\mathcal{H}_K$ with an inner product in terms of the Fourier coefficients of functions in $\mathcal{H}_K$. If developed more fully, the connections with harmonic analysis give a very strong connection between the norm in this RKHS and the smoothness properties of its functions.

References

[1] Barbara MacCluer. *Graduate Texts in Mathematics: Elementary Functional Analysis*. Springer, New York, NY, 2009.

[2] Sergei Ovchinnikov. *Functional Analysis: An Introductory Course*. Springer, New York, NY, 2018.

[3] Edward B Saff and Arthur David Snider. *Fundamentals of complex analysis for mathematics, science and engineering*, 2nd ed. Pearson, February 1993.

[4] Nachman Aronszajn. *Theory of Reproducing Kernels*. Trans. Amer. Math. Soc., 686:337-404, 1950.

[5] Grace Wahba. *Spline Models for Observational Data*. Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990. Pp. 1-20.

[6] Federico Girosi, *An Equivalence Between Sparse Approximation and Support Vector Machines*. A.I. Memo No. 1606, Artificial Intelligence Laboratory, Massachusetts Institution of Technology, 1997.

[7] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 1998.

[8] Alexander J. Smola, Bernhard Schölkopf, and Klaus-Robert Müller, *The connection between regularization operators and support vector kernels*. Neural Networks, Volume 11, Elsevier, Amsterdam, Netherlands, 1998.

[9] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. *Regularization Networks and Support Vector Machines*. Advances in Computational Mathematics, Springer, New York, NY, 1999.

[10] Vladimir Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed.. Springer, New York, NY, 2000.

[11] Alexander J. Smola and Bernhard Schölkopf, *A Tutorial on Support Vector Regression*. Statistics and Computing, Volume 14, Kluwer Academic Publishers, Amsterdam, Netherlands, 2003.

[12] Matthias Hein and Oliver Bousquet, *Kernels, Associated Structures and Generalizations*. Technical Report 127, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2004.

[13] Florian Steinke and Bernard Schölkopf, *Kernels, regularization and differential equations*. Pattern Recognition, Volume 41, Elsevier, Amsterdam, Netherlands, 2008.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Section 5.8. Springer, New York, NY, February 2009.

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning: With applications in R*. 2nd ed., Springer, New York, NY, 2021.

[16] Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe De Nicolao, and Lennart Ljung, *Regularized System Identification: Learning Dynamic Models from Data.* Communications and Control Engineering, Springer, New York, NY, 2022.

[17] Jonathan H. Manton and Pierre-Olivier Amblard, *A Primer on Reproducing Kernel Hilbert Spaces*. Foundations and Trends in Signal Processing, now Publishers Inc., 2015.

[18] Frigyes Reisz and Béla Sz.-Nagy. *Functional Analysis*. Frederick Ungar Publishing Co., New York, NY, 1955.

[19] Jacques Hadamard. *Sur les problèmes aux dérivées partielles et leur signification physique*. Princeton University Bulletin, 1902. Pp. 49–52.

[20] Peter Duren and Alexander Schuster, *Bergman Spaces*. Mathematical Surveys and Monographs, Volume 100, American Mathematical Society, 2004.

[21] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*. MIT Press, Boston, MA, 2016. Retrieved from `http://www.deeplearningbook.org`.

[22] Vern Paulsen. An Introduction to the Theory of Reproducing Kernel Hilbert Spaces. University of Houston lecture notes, Spring 2006. Retrieved from `https://www.math.uh.edu/ vern/rkhs.pdf`.

[23] Steven Johnson. Introduction to Numerical Methods. Massachusetts Institute of Technology lecture notes, Fall 2020. Retrieved from `https://math.mit.edu/ stevenj/18.335/norm-equivalence.pdf`

[24] Tomaso Poggio, Lorenzo Rosasco, Alexander Rakhlin, and Andrzej Banburski. Statistical Learning Theory and Applications, Fall 2019. Retrieved from `http://www.mit.edu/ 9.520/scribe-notes/class02`$_s crb_s ara.pdf$