Reproducing Kernel Hilbert Spaces

David S. Rosenberg

1 Definition and Basic Properties

Definition 1. [RKHS] A reproducing kernel Hilbert space (RKHS) of functions from $\mathcal{X}$ to $\mathbf{R}$ is a Hilbert Space $\mathcal{H}$ that possesses a reproducing kernel, *i.e.*, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ for which the following properties hold:

1. $k(x, .) \in \mathcal{H}$ for all $x \in \mathcal{X}$, and
2. $\langle f, k(x, .)\rangle_{\mathcal{H}} = f(x)$, for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$.

An *evaluation functional* is a functional that maps a function $f \in \mathcal{H}$ to its evaluation at some fixed point, such as $f \mapsto f(x)$. An RKHS is sometimes defined as a Hilbert space for which all evaluation functionals are continuous. Here we show that this is a consequence of our Definition 1:

Proposition 2. *Any evaluation functional $f \mapsto f(x)$ on an RKHS $\mathcal{H}$ is uniformly continuous.*

Proof. For any $f, g \in \mathcal{H}$, we have

$$\begin{aligned}
|f(x) - g(x)| &= |\langle f, k(x, \cdot)\rangle_{\mathcal{H}} - \langle g, k(x, \cdot)\rangle_{\mathcal{H}}| \\
&= |\langle f - g, k(x, \cdot)\rangle_{\mathcal{H}}| \\
&\leq \|f - g\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}} \text{ by Cauchy Schwartz} \\
&= \|f - g\|_{\mathcal{H}} \sqrt{\langle k(x, \cdot), k(x, \cdot)\rangle_{\mathcal{H}}} \\
&= \|f - g\|_{\mathcal{H}} \sqrt{k(x, x)}
\end{aligned}$$

Thus for all $f, g \in \mathcal{H}$ for which $\|f - g\|_{\mathcal{H}} \leq \varepsilon / \sqrt{k(x, x)}$, we have $|f(x) - g(x)| \leq \varepsilon$. □

Proposition 3. *The reproducing kernel $k(\cdot,\cdot)$ for an RKHS $\mathcal{H}$ is positive definite, which means that for any $n = 1, 2, 3, \ldots$, and any choice of points $x_1, \ldots, x_n \in \mathcal{X}$, the kernel matrix $K = (k(x_i, x_j))_{i,j=1}^n$ is positive semidefinite.*

Proof. Fix any $n = 1, 2, 3, \ldots$ and any choice of points $x_1, \ldots, x_n \in \mathcal{X}$. For any $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbf{R}^n$, we have

$$0 \leq \left\| \sum \alpha_i k(x_i, \cdot) \right\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}' K \boldsymbol{\alpha}$$

Thus K is positive semidefinite. □

Lemma 4. *Let $\mathcal{H}$ be an RKHS of functions mapping $\mathcal{X} \to \mathbf{R}$, having reproducing kernel $k(\cdot,\cdot)$. Fix any $z \in \mathcal{X}$, and let $k_z = k(z, \cdot) \in \mathcal{H}$. Then*

$$\begin{aligned} \sup_{f:\|f\|_{\mathcal{H}} \leq 1} |f(z)| &= \sup_{\|f\|_{\mathcal{H}} \leq 1} |\langle f, k_z \rangle| \\ &= \langle k_z, k_z / \|k_z\|_{\mathcal{H}} \rangle = \frac{1}{\|k_z\|_{\mathcal{H}}} k(z, z) \\ &= \sqrt{k(z, z)} \end{aligned}$$

2 Projections

The **Hilbert projection theorem** states that for any f in a Hilbert space $\mathcal{H}$, and for any closed subspace $\mathcal{L} \subset \mathcal{H}$, there exists a unique element $f_\parallel \in \mathcal{L}$, called the *projection* of f onto $\mathcal{L}$, such that $\|f - f_\parallel\|_{\mathcal{H}} \leq \|f - g\|_{\mathcal{H}}$ for every $g \in \mathcal{L}$. We will denote the projection of f onto $\mathcal{L}$ by $\mathrm{Proj}_{\mathcal{L}} f$. A necessary and sufficient condition for $f_\parallel \in \mathcal{L}$ to be the projection of f onto $\mathcal{L}$ is that $\langle f - f_\parallel, g \rangle = 0$ for every $g \in \mathcal{L}$. A simple corollary of this characterization is that the norm of $f_\parallel$ never exceeds the norm of f:

Lemma 5. *For any f in a Hilbert space $\mathcal{H}$, let $f_\parallel = \mathrm{Proj}_{\mathcal{L}} f$, where $\mathcal{L}$ is a closed subspace. Then*

$$\|f_\parallel\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}}.$$

Proof. We have $f = f_\parallel + (f - f_\parallel)$, and $\langle f_\parallel, f - f_\parallel \rangle = 0$. Thus

$$\|f\|_{\mathcal{H}}^2 = \|f_\parallel\|_{\mathcal{H}}^2 + \|f - f_\parallel\|_{\mathcal{H}}^2 + 2\langle f_\parallel, f - f_\parallel \rangle_{\mathcal{H}} \geq \|f_\parallel\|_{\mathcal{H}}^2.$$

□

When the Hilbert space $\mathcal{H}$ is an RKHS of functions mapping from $\mathcal{X} \to \mathbf{R}$, we have the following interesting result:

Lemma 6. *Let $\mathcal{H}$ be an RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$, and consider any point $x \in \mathcal{X}$. If $\mathcal{L} \subset \mathcal{H}$ is a closed subspace containing $k(x, \cdot)$, then the projection of f onto $\mathcal{L}$ has the same value at x as f does. That is,*

$$f(x) = \left(Proj_{\mathcal{L}} f\right)(x)$$

Proof. By the Hilbert projection theorem, there exists a unique element $f_{\|} = \mathrm{Proj}_{\mathcal{L}} f \in \mathcal{L}$ such that $\left\langle f - f_{\|}, g \right\rangle_{\mathcal{H}} = 0$ for all $g \in \mathcal{L}$. By definition of RKHS, we have $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$. Combining these facts, we get

$$\begin{aligned} f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}} &= \left\langle f_{\|} + f - f_{\|}, k(x,\cdot)\right\rangle = \left\langle f_{\|}, k(x,\cdot)\right\rangle + \left\langle f - f_{\|}, k(x,\cdot)\right\rangle \\ &= \left\langle f_{\|}, k(x,\cdot)\right\rangle = f_{\|}(x) \end{aligned}$$

□

3 The Representer Theorem and the "Span of the Data"

In practice, many RKHS optimization problems refer to a fixed set of points $x_1, \ldots, x_n \in \mathcal{X}$, and the solution to the optimization problem is contained in a special subspace of the RKHS, often referred to informally as the "span of the data." For an RKHS $\mathcal{H}$ with kernel $k(\cdot, \cdot)$, the *span of the data* is the linear subspace

$$\mathcal{L} = \mathrm{span}\left\{k(x_1, \cdot), \ldots, k(x_n, \cdot)\right\}.$$

We now state and prove the representer theorem, which gives some conditions under which the solution to an RKHS optimization problem is contained in the span of the data.

Theorem 7. *[Representer Theorem] Let $\mathcal{H}$ be an RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$. Fix any function $V : \mathbf{R}^n \to \mathbf{R}$ and any nondecreasing function $\Omega : \mathbf{R} \to \mathbf{R}$. Define*

$$J(f) = V\left(f(x_1), \ldots, f(x_n)\right) + \Omega\left(\|f\|_{\mathcal{H}}^2\right)$$

Let $\mathcal{L} = span\left\{k(x_1, \cdot), \ldots, k(x_n, \cdot)\right\}$. Then for any $f \in \mathcal{H}$ we have

$$J(Proj_{\mathcal{L}} f) \leq J(f).$$

Thus if $J^* = \min_{f \in \mathcal{L}} J(f)$ exists, then this minimum is attained for some $f \in \mathcal{L}$. Furthermore, if Ω is strictly increasing, then each minimizer of $J(f)$ over $\mathcal{H}$ is also contained in $\mathcal{L}$.

Proof. Let $f_\parallel = \mathrm{Proj}_{\mathcal{L}} f$. By Lemma 6, $f(x_i) = f_\parallel(x_i)$ for $i = 1, \ldots, n$. Thus

$$V\left(f(x_1), \ldots, f(x_n)\right) = V\left(f_\parallel(x_1), \ldots, f_\parallel(x_n)\right).$$

Since f and $f - f_\parallel$ are orthogonal, we have

$$\|f\|_{\mathcal{H}}^2 = \|f_\parallel\|_{\mathcal{H}}^2 + \|f - f_\parallel\|_{\mathcal{H}}^2 \geq \|f_\parallel\|_{\mathcal{H}}^2.$$

Since Ω is nondecreasing, $\Omega\left(\|f_\parallel\|_{\mathcal{H}}^2\right) \leq \Omega\left(\|f\|_{\mathcal{H}}^2\right)$. Combining these results, we get $J(f_\parallel) \leq J(f)$. If Ω is strictly increasing and $f \neq f_\parallel$ (i.e. $\|f - f_\parallel\|_{\mathcal{H}}^2 > 0$), then

$$\begin{aligned} \|f_\parallel\|_{\mathcal{H}}^2 &< \|f\|_{\mathcal{H}}^2 \\ \implies \Omega\left(\|f_\parallel\|_{\mathcal{H}}^2\right) &< \Omega\left(\|f\|_{\mathcal{H}}^2\right) \\ \implies J(f_\parallel) &< J(f). \end{aligned}$$

Thus every minimizer of $J(f)$ must be contained in $\mathcal{L}$. □

Definition 8. We define the *kernel matrix* for the data $x_1, \ldots, x_n \in \mathcal{X}$ by the matrix $K = (k(x_i, x_j))_{i,j=1}^n$.

This matrix is very useful when dealing with functions that live in the span of the data. The following proposition gives two useful expressions involving the kernel matrix:

Proposition 9. *For any function $f : \mathcal{X} \to \mathbf{R}$ of the form*

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot),$$

for some $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbf{R}^n$, we have

$$\|f\|_{\mathcal{H}}^2 = \left\langle \sum_{i=1}^n \alpha_i k(x_i, \cdot), \sum_{i=1}^n \alpha_i k(x_i, \cdot) \right\rangle_{\mathcal{H}} = \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^T K \boldsymbol{\alpha}.$$

Let $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))^T$ be the column vector of evaluations of f on the data points. Then we also have

$$\boldsymbol{f} = (f(x_j))_{j=1}^n = \left(\sum_{i=1}^n \alpha_i k(x_i, x_j) \right)_{j=1}^n = K\boldsymbol{\alpha}$$

4 Kernel Ridge Regression and the SVM

In this section we present two specific algorithms: kernel ridge regression, also known as regularized least squares (RLS), and the soft-margin support vector machine (SVM). In their most natural forms, the optimization problems associated with these algorithms are optimizations over an RKHS. Below, we show how to use the Representer Theorem to reduce the optimization over a function space to an optimization over a finite dimensional Euclidean space.

Consider an RKHS $\mathcal{H}$ and a loss functional $L : \mathcal{H} \to \mathbf{R}$. We would like to find

$$\arg\min_{f \in \mathcal{H}} \left[L(f) + \lambda \|f\|_{\mathcal{H}}^2 \right], \tag{4.1}$$

for some $\lambda > 0$. Suppose, as is typical in practice, the loss functional takes the form

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

where $(x_1, y_1), \ldots, (x_n, y_n)$ are labeled training examples. Then by the Representer Theorem (Theorem 7), if a minimizer for Equation (4.1) exists, then the minimum is attained by some function of the form $f_{\boldsymbol{\alpha}} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbf{R}^n$. Thus we have reduced the problem of finding the optimal $f \in \mathcal{H}$ to the finite dimensional optimization problem of finding the best $\boldsymbol{\alpha} \in \mathbf{R}^n$.

We now recall some of the notation and facts derived above that will facilitate writing this finite-dimensional optimization problem in a form that can be easily solved on a computer. For any function $f_{\boldsymbol{\alpha}}$, as defined above, define the column vector $\boldsymbol{f}_{\boldsymbol{\alpha}} = (f_{\boldsymbol{\alpha}}(x_1), \ldots, f_{\boldsymbol{\alpha}}(x_n))^T$. Let $K = (k(x_i, x_j))_{i,j=1}^n$ be the $n \times n$ kernel matrix on the data points. Then by Prop. 9, $\boldsymbol{f}_{\boldsymbol{\alpha}} = K\boldsymbol{\alpha}$ and $\|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}}^2 = \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$.

For RLS, we take the loss function is to be $V(\hat{y}, y) = (y - \hat{y})^2$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Then we can write the minimization as

$$\begin{aligned}
\arg\min_{\boldsymbol{\alpha} \in \mathbf{R}^n} \left[L(f_{\boldsymbol{\alpha}}) + \lambda \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}}^2 \right] &= \arg\min_{\boldsymbol{\alpha} \in \mathbf{R}^n} (\boldsymbol{y} - K\boldsymbol{\alpha})^T (\boldsymbol{y} - K\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \\
&= \arg\min_{\boldsymbol{\alpha} \in \mathbf{R}^n} \boldsymbol{\alpha} K^2 \boldsymbol{\alpha} - 2\boldsymbol{y}^T K \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \\
&= \arg\min_{\boldsymbol{\alpha} \in \mathbf{R}^n} \left[\boldsymbol{\alpha} \left(K^2 + \lambda K \right) \boldsymbol{\alpha} - 2\boldsymbol{y}^T K \boldsymbol{\alpha} \right]
\end{aligned}$$

Since we are minimizing over an open set, the minimum must occur at a critical point, so we solve the first order conditions:

$$\begin{aligned} 0 = \partial_{\boldsymbol{\alpha}} \boldsymbol{\alpha} \left(K^2 + \lambda K\right) \boldsymbol{\alpha} - 2\boldsymbol{y}^T K \boldsymbol{\alpha} &= 2\left(K^2 + \lambda K\right) \boldsymbol{\alpha} - 2K\boldsymbol{y} \\ \implies \left(K^2 + \lambda K\right) \boldsymbol{\alpha} &= K\boldsymbol{y} \end{aligned}$$

It is clear that one solution to this equality is $\boldsymbol{\alpha}_* = (K + \lambda I)^{-1} \boldsymbol{y}$, which exists if we assume $\lambda > 0$. Thus the RLS prediction function is $f_{\boldsymbol{\alpha}_*}$.

For the SVM, we take the loss function to be the *hinge loss*, which is defined as

$$V(y, \hat{y}) = (1 - y\hat{y})_+ = \begin{cases} 1 - y\hat{y} & \text{for } 1 - y\hat{y} \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The SVM optimization problem can be written as

$$\min_{\boldsymbol{\alpha} \in \mathbf{R}^n} \sum_{i=1}^{n} (1 - y_i f_{\boldsymbol{\alpha}}(x_i))_+ + \lambda \|f_{\boldsymbol{\alpha}}\|_{\mathcal{H}}^2.$$

Defining $Y = \operatorname{diag}(y_1, \ldots, y_n)$, we can rewrite this optimization problem as

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbf{R}^n} \quad & \sum_{i=1}^{n} \boldsymbol{\beta}_i + \lambda \boldsymbol{\alpha}^T K \boldsymbol{\alpha} \\ \text{subject to} \quad & \boldsymbol{\beta} \succeq 0 \\ & \boldsymbol{\beta} \succeq 1 - YK\boldsymbol{\alpha}, \end{aligned}$$

where $\succeq$ means that the inequality holds component-wise for the vectors. This optimization problem is a quadratic program with linear constraints, which can be solved by many standard numerical computing packages.