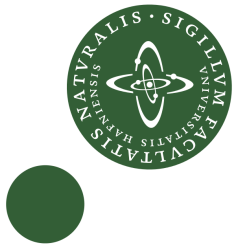
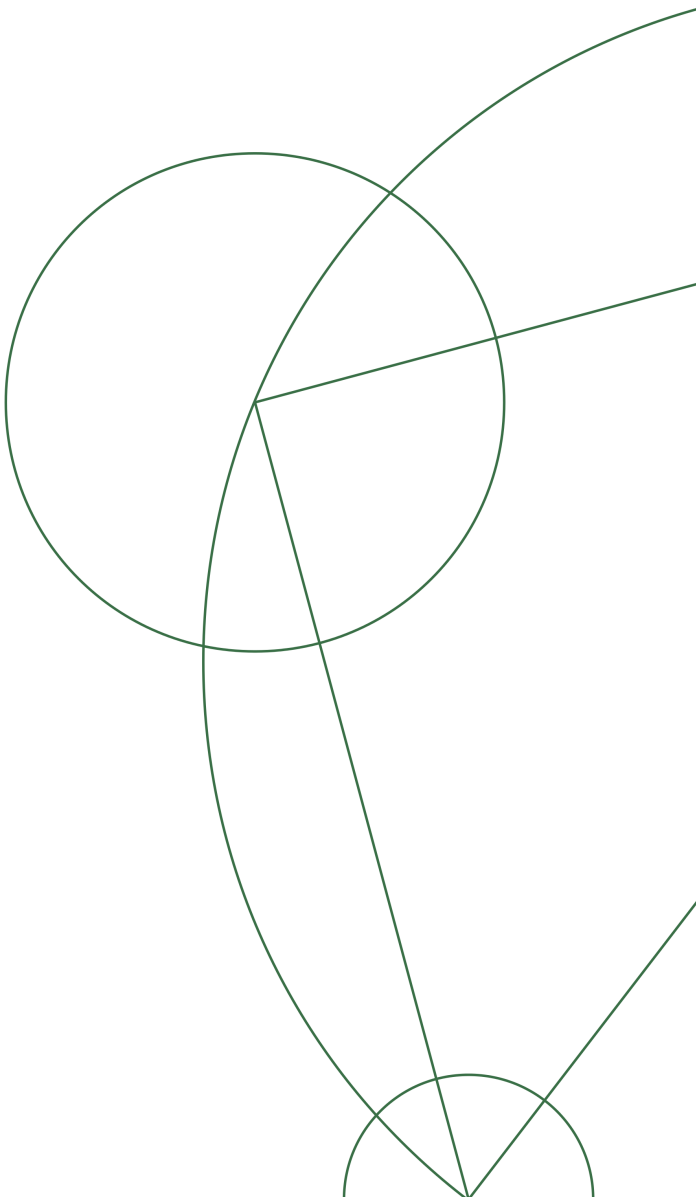
Department of Mathematical Sciences
UNIVERSITY OF COPENHAGEN



Malthe Munk Karbo

A Generalized Representer Theorem



Project in Mathematics
Department of Mathematical Sciences
University of Copenhagen

Advisors
Niels Richard Hansen
Lasse Petersen

July 12, 2018

Introduction

In fields like statistics, machine learning and alike, we are often interested in measuring how 'equal' certain objects are - e.g., observed data, and occasionally this is the goal of a statistician is to find a function which allows him to measure this. There is a rich theory of normed spaces and Hilbert spaces, which aids in this process. We will in this paper introduce and discuss the theory of *Reproducing Kernel Hilbert Spaces*, which is a class of Hilbert spaces stemming from a certain class of functions, called *kernels*, which in a suitable fashion generalizes the notion of an inner product. We then move on to show a result called the 'Representer Theorem', which characterizes the solutions to certain problems in a nice way, and some generalizations of it.

CONTENTS

Kernel theory

In this paper, we will restrict ourselves to working with real Hilbert spaces and real-valued functions, however, it is entirely possible to develop theory discussed in the following for complex Hilbert spaces.

Reproducing Kernels

We will denote Hilbert spaces by $\mathcal{H}$, and by $\mathcal{X}$ we will denote any non-empty set - usually refered to as a sample space. The goal of this section is to characterize a certain kind of Hilbert space of functions through a class of functions called *kernels*:

Definition 1.1. A Hilbert space $\mathcal{H}$ of functions $f\colon \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel Hilbert space* (RHKS) if the evaluation functional $f \mapsto f(x)$ is continuous.

The term *kernel* comes from a connection with a class of certain compact operators on a Hilbert space, called integral operators - which is usually defined pointwise by integration against a function initially refered to as an integral kernel.

By the Riesz Representation theorem, there is a unique vector $ev_x \in \mathcal{H}$ such that the evaluation $f \mapsto f(x)$ is given by $\langle f, ev_x \rangle$ for $f \in \mathcal{H}$, whenever $\mathcal{H}$ is a RKHS. An important distinguishment between general Hilbert spaces and RKHS's is the following:

Lemma 1.2. Let $\mathcal{H}$ be a RKHS in over $\mathcal{X}$, and let $(f_n) \subseteq \mathcal{H}$. If $f_n \to f$ in norm, then $f_n(x) \to f(x)$ for all $x \in \mathcal{H}$.

Proof.
For all $x \in \mathcal{X}$ and f_n, f as above, we get by Cauchy-Schwarz that

$$|f_n(x) - f(x)| = |\langle f_n - f, ev_x \rangle| \leq \|f_n - f\| \, \|ev_x\| \to 0.$$

□

We are interested in a special kind of kernel functions, namely so called positive-definite ones:

Definition 1.3. A symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that given any non-empty finite subset $F \subseteq \mathcal{X}$, the matrix $(k(x,y))_{x,y\in F}$ is positive semi-definite is called a *positive-definite kernel* on $\mathcal{X}$.

Given a symmetric function k on $\mathcal{X}$, then for fixed $y \in \mathcal{X}$ we will use k_y to denote the function $k(\cdot, y)$, whenever this imposes no confusion. This will be convenient later on.

There is a correspondance between positive-definite kernels on $\mathcal{X}$ and reproducing spaces over $\mathcal{X}$, as we will see:

Proposition 1.4. Given a positive-definite kernel k on $\mathcal{X}$, there is a reproducing kernel Hilbert space $\mathcal{H}_k \subseteq \mathbb{R}^{\mathcal{X}}$ such that the evaluation functional is given by $f(x) = \langle f, k_x\rangle$ for $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$.

Proof.
If we set $\mathcal{H}_0 := \operatorname{span}_{\mathbb{R}} \left\{k_x \,\middle|\, x \in \mathcal{X}\right\}$, then we obtain a pre-Hilbert space with respect to the Hermitian form defined by

$$\left\langle \sum_i \alpha_i k(\cdot, y_i), \sum_j \beta_j k(\cdot, x_j) \right\rangle := \sum_{i,j} \alpha_i \beta_j k(y_i, x_j),$$

since k is positive definite and symmetric: It is easy to see that if $f \in \mathcal{H}_0$ - i.e., $f\colon x \mapsto \sum_{i=1}^n \alpha_i k_{x_i}(x)$ for some $x_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$, then for $x \in \mathcal{X}$ we have

$$\langle k(\cdot, x), f\rangle = f(x),$$

and since k is positive definite, we have $k(x,y)^2 \leq k(x,x)k(y,y)$. Hence, for $x \in \mathcal{X}$ we see that $|f(x)|^2 = |\langle k(\cdot,x), f\rangle|^2 \leq |k(x,x)\langle f, f\rangle|$ so that $\langle f, f\rangle = 0 \iff f = 0$.

Let $\mathcal{H}_k$ denote the Cauchy completion of $\mathcal{H}_0$ with respect to $\langle\cdot,\cdot\rangle$, and denote also the inner product extension to $\mathcal{H}_k$ by $\langle\cdot,\cdot\rangle$. We wish to show that $\mathcal{H}_k$ is in fact a RKHS. So let $h \in \mathcal{H}_k$ and pick any Cauchy sequence $(f_n) \subseteq \mathcal{H}_0$ such that $f_n \to f$. For $x \in \mathcal{X}$, let $h(x) := \lim_n f_n(x)$. By Cauchy-Schwarz:

$$|\langle f_n - f_m, k_x\rangle| \leq \|f_n - f_m\| \, \|k(x,x)\| \to 0, \text{ for } x \in \mathcal{X}, \text{ as } n, m \to \infty,$$

so $h(x) \in \mathbb{R}$ is well-defined, and it is clear that $x \mapsto h(x)$ is independent on the choice of Cauchy representant (f_n) of h. Hence we may view $\mathcal{H}_k$ as a subset of $\mathbb{R}^{\mathcal{X}}$. Furthermore, if h and (f_n) is as above, then

$$\langle h, k_x\rangle = \lim_n \langle f_n, k_x\rangle = \lim_{n\to\infty} f_n(x) = h(x), \text{ for } x \in \mathcal{X},$$

and since $k_x \in \mathcal{H}_k$, Riesz Representation theorem says that $\mathcal{H}_k$ is a RKHS. □

If $\mathcal{H}$ is a RKHS over $\mathcal{X}$, then the evaluation functional δ_x is, by the Riesz Representation theorem, given by some element $k_x \in \mathcal{H}$. Then the map $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k(x,y) := k_y(x)$, is clearly symmetric and if $\alpha \in \mathbb{R}^n$ and $x_1, \ldots, x_n \in \mathcal{X}$, then

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \left\| \sum_i \alpha_i k_{x_i} \right\|^2 \geq 0,$$

so it is a positive definite kernel. The correspondance between positive definite kernels and RKHS is in fact one-to-one. This follows from the following:

Lemma 1.5. If $\mathcal{H}_k$ is a RKHS over $\mathcal{X}$, then $\text{span}_{\mathbb{R}} \{k_x \mid x \in \mathcal{X}\}$ is dense in $\mathcal{H}_k$.

Proof.
The orthogonal complement of the span is equal to zero, since $\mathcal{H}_k$ is reproducing since $f \perp k_x \implies f(x) = 0$ for $x \in \mathcal{X}$ and $f \in \mathcal{H}_k$. Hence the span is dense in $\mathcal{H}_k$. □

Proposition 1.6. If $\mathcal{H}_1, \mathcal{H}_2$ are RKHS's over $\mathcal{X}$ such that their kernels k_1 and k_2 are equal, then $\mathcal{H}_1 = \mathcal{H}_2$ and $\|f\|_{\mathcal{H}_1} = \|f\|_{\mathcal{H}_2}$.

Proof.
By Lemma 1.5, we see that $V_i = \text{span}_{\mathbb{R}} \{k_x \in \mathcal{H}_i \mid x \in \mathcal{X}\}$ is dense in H_i, $i = 1, 2$, and $f \in V_i$ is independent of the choice of i, since $f(x) = \sum_k \alpha_k k_{x_k}(x)$ for some $\alpha_j \in \mathbb{R}$ and $x_j \in \mathcal{X}$, so $V_1 = V_2$ in $\mathbb{R}^{\mathcal{X}}$, and if $f = \sum_j \alpha_j k_{x_j} \in V_1$ then

$$\|f\|_{\mathcal{H}_1} = \sum_{i,j} \alpha_j \alpha_i k(x_i, x_j) = \|f\|_{\mathcal{H}_2} .$$

Let $f \in \mathcal{H}_1$, and pick a sequence $f_n \subseteq V_1$ such that $f_n \to f$ in $\mathcal{H}_1$. By the above, f_n is Cauchy in V_2 and hence $\mathcal{H}_2$, with limit $g \in \mathcal{H}_2$. Then, by Lemma 1.2 we have

$$f(x) = \lim_n f_n(x) = g(x), \ x \in \mathcal{X}.$$

By symmetry, we see that $\mathcal{H}_1 = \mathcal{H}_2$, and since the norms agree on a dense set, they agree on the equal closures. □

Hence, we may characterize RKHS's over $\mathcal{X}$ as a pair $(\mathcal{H}_k, k)$ or simply $\mathcal{H}_k$, for some positive-definite kernel k on $\mathcal{X}$, where the space $\mathcal{H}_k$ is the space defined in Proposition 1.4. This space is called the *canonical kernel space*.

It is easy to represent the kernel k of a RKHS $\mathcal{H}_k$ in terms of a ONB:

Theorem 1.7. Let $\mathcal{H}_k$ be a RKHS, and let $(e_i)_{i \in I}$ be an ONB for $\mathcal{H}_k$. Then $k(x, y) = \sum_{i \in I} e_i(y) e_i(x)$, for $x, y \in \mathcal{X}$.

Proof.
It holds that $e_i(x) = \langle k_x, e_i \rangle$, hence expanding into Fourier coefficients we see that $k_y = \sum_{i \in I} e_i(y) e_i$. The theorem follows since norm convergence implies point-wise convergence in a RKHS. □

We may also construct RKHS in a different way: Given a set $\mathcal{X}$, a Hilbert space $\mathcal{H}$ and a map $\varphi \colon \mathcal{X} \to \mathcal{H}$, if we define $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, y) := \langle \varphi(x), \varphi(y) \rangle, \text{ for } x, y \in \mathcal{X},$$

then k is a positive-definite kernel giving rise to a RKHS. While k is uniquely determined by its canonical RKHS $\mathcal{H}_k$, there may be other maps $\psi\colon \mathcal{X} \to \mathcal{H}'$ inducing the same kernel. In this setup, the map φ is called a *feature map* and $\mathcal{H}$ is called a *feature space*. Clearly, given a RKHS $\mathcal{H}_k$, the map $x \mapsto k_x$, is a feature map with feature space $\mathcal{H}_k$. We also say that $\mathcal{H}_k$ is the canonical feature space of k and $\mathcal{X}$.

Examples and constructions of RKHS and kernels

We will now cover a few important examples of both kernels and RKHS's and ways to define new kernels and RKHS's from old ones.

Lemma 1.8. If k_1 and k_2 are kernels on $\mathcal{X}$, then $k = k_1 + k_2$ is a kernel and it's RKHS is the algebraic product $\mathcal{H} = \mathcal{H}_{k_1} + \mathcal{H}_{k_2}$ with norm

$$\|f\|_{\mathcal{H}}^2 = \min_{(f_1,f_2)\in\mathcal{H}_{k_1}\oplus\mathcal{H}_{k_2}} \left\{ \|f_1\|_{\mathcal{H}_{k_1}}^2 + \|f_2\|_{\mathcal{H}_{k_2}}^2 \right\} \tag{1.1}$$

Proof.
Let $\mathcal{H}' = \mathcal{H}_{k_1} \oplus \mathcal{H}_{k_2}$ equipped with the natural inner product and π be the canonical surjection onto $\mathcal{H}$, i.e., $s\colon (f,g) \mapsto f+g$. Then the nullspace of $\mathcal{N}(\pi) = \{f \oplus -f \mid f \in H_{k_1} \cap H_{k_2}\}$ of π is closed, for if $(f_n, -f_n)_n \subseteq \mathcal{N}(s)$ converges in $\mathcal{H}'$ to a pair (f,g), then $f_n \to f$ and $-f_n \to g$ in $\mathcal{H}_{k_1}$ and $\mathcal{H}_{k_2}$, respectively, and therefore also pointwise everywhere, so $f(x) = -g(x)$ for all $x \in \mathcal{X}$. We may then write $\mathcal{H}' = \mathcal{N}(s) \oplus \mathcal{N}(s)^\perp$, and we may identify $\mathcal{H}$ with $\mathcal{N}(s)^\perp$ as Hilbert spaces.

Now, let P denote the orthogonal projection onto $\mathcal{N}(s)^\perp$. Then, by definition of the orthogonal projection, we see that Equation (1.1) holds, and for $f, g \in \mathcal{H}$, we have $\langle f, g\rangle_{\mathcal{H}} = \langle P(f_0,f_1), P(g_0,g_1)\rangle_{\mathcal{H}'}$, where $f = P(f_0,f_1)$ and $g = P(g_0,g_1)$.

It remains to be shown that $\mathcal{H}$ is a RKHS with kernel k. Note that for all x we have $(k_1(\cdot,x), k_2(\cdot,x)) \perp \mathcal{N}(s)$, since the inner product with $(f,-f)$ equals $f(x) - f(x) = 0$. If $f = f_1 + f_2 \mathcal{H}$ then

$$\begin{aligned}\langle f, k_x\rangle_{\mathcal{H}} = \langle P(f_1,f_2), P(k_1(\cdot,x), k_2(\cdot,x))\rangle_{\mathcal{H}'} &= \langle (f_1,f_2), (k_1(\cdot,x), k_2(\cdot,x))\rangle_{\mathcal{H}'} \\ &= \langle f_1, k_1(\cdot,x)\rangle_{\mathcal{H}_1} + \langle f_2, k_2(\cdot,x)\rangle_{\mathcal{H}_2} \\ &= f_1(x) + f_2(x) = f(x),\end{aligned}$$

finishing the proof. □

Lemma 1.9. If k_1 and k_2 are kernels on $\mathcal{X}$, then $k = k_1 \otimes k_2$, given by $\mathcal{X} \times \mathcal{X} \ni (x,y) \mapsto k_1(x,y)k_2(x,y)$, is a kernel.

Proof.
First note that if $\mathcal{H}_1$ and $\mathcal{H}_2$ are the RKHS of k_1 and k_2 respectively, then their Hilbert space tensor product $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ with inner product defined on simple

tensors $g \otimes h, g' \otimes h'$ by $\langle g \otimes h, g' \otimes h' \rangle_{\mathcal{H}} := \langle g, g \rangle_{\mathcal{H}_1} \langle h, h' \rangle_{\mathcal{H}_2}$ is again a RKHS, since for $g \otimes h \in \mathcal{H}$ and $(x, y) \in \mathcal{X} \times \mathcal{X}$, we have

$$g \otimes h(x, y) = g(x)h(y) = \langle g, k_1(\cdot, x) \rangle_{\mathcal{H}_1} \langle h, k_2(\cdot, y) \rangle_{\mathcal{H}_2} = \langle g \otimes h, k_1(\cdot, x)k_2(\cdot, y) \rangle_{\mathcal{H}},$$

and since the inner product is defined pointwise and the simple tensors span a dense subset of $\mathcal{H}$, we obtain the desired. Moreover, k is clearly symmetric and also positive definite, for if $\varphi(x) := k_1(\cdot, x) \otimes k_2(\cdot, x)$, then

$$\begin{aligned} k(x, y) = k_1(x, y)k_2(x, y) &= \langle k_1(\cdot, x), k_1(\cdot, y) \rangle_{\mathcal{H}_1} \langle k_2(\cdot, x), k_2(\cdot, y) \rangle_{\mathcal{H}_2} \\ &= \langle k_1(\cdot, x) \otimes k_2(\cdot, x) \rangle, k_1(\cdot, y) \otimes k_2(\cdot, y) \rangle_{\mathcal{H}} \\ &= \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}. \end{aligned}$$

Alternatively, it follows since $a^* a \otimes b^* b = (a^* \otimes b^*)(a \otimes b) \geq 0$ whenever a, b are bounded linear operators on Hilbert spaces $\mathcal{H}, \mathcal{K}$ respectively, so given any finite set $F \subseteq \mathcal{X}$, the matrix $(k(x, y))_{x,y \in F} = (k_1(x, y))_{x,y \in F} \otimes (k_2(x, y))_{x,y \in F}$, which consequently must be of the form $a^* a \otimes b^* b$ for suitable matrices. □

We end with some examples of canonical kernels:

Example 1.10

1. If $\lambda > 0$ and k is a kernel on $\mathcal{X}$, then λk is a kernel on $\mathcal{X}$.
2. If $(k_n)_{n \in \mathbb{N}}$ is a sequence of kernels on $\mathcal{X}$ such that the limit $k(x, y) := \lim_n k_n(x, y)$ exists for all $x, y \in \mathcal{X}$, then k is a kernel on k.
3. If $f \colon \mathcal{X} \to \mathbb{R}$ is any function and k is a kernel on $\mathcal{X}$, then the function $\tilde{k}(x, y) := f(x)k(x, y)f(y)$ is a kernel on $\mathcal{X}$.
4. If $\mathcal{X} = \mathbb{R}^n$, then $(x, y) \mapsto \langle x, y \rangle^k$ is a kernel on $\mathcal{X}$ called the *polynomial kernel of degree k on $\mathcal{X}$*.
5. If k is a kernel on $\mathcal{X}$, then $k'(x, y) := e^{k(x,y)}$ is a kernel on $\mathcal{X}$ called the *exponential kernel* of k.
6. For $\lambda > 0$ and $\mathcal{X} = \mathbb{R}^n$, then the function $k(x, y) = e^{-\lambda \|x-y\|^2}$ is a kernel called the *radial basis function kernel (RBF kernel)* with bandwidth $\lambda > 0$ (often parametrized with $\lambda = \frac{1}{2\sigma^2}$).

Proof.
1: If φ is a feature map for k, let $\Phi = \sqrt{\lambda}\varphi$, then Φ is a featuremap for λk.

2: Clearly the limit will be symmetric and positive semi-definite.

3: Let φ be a feature map associated to k, and let $\Phi \colon x \mapsto f(x)\varphi(x)$ for $x \in \mathcal{X}$, then $\tilde{k}(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is the desired kernel

4: This follows from repeated use of Lemma 1.9 applied to $k(x, y) = \langle x, y \rangle$.

5: For $N \geq 0$, we see that $\sum_{m=0}^{N} \frac{1}{m!} k(x,y)^m$ defines a kernel by (1), Lemma 1.8 and Lemma 1.9. By (2), we see that

$$e^{k(x,y)} = \sum_{m=0}^{\infty} \frac{1}{m!} k(x,y)^m = \lim_{N\to\infty} \sum_{m=0}^{N} \frac{1}{m!} k(x,y)^m,$$

is a kernel.

6: This follows from the identity

$$k(x,y) = e^{-\lambda\|x-y\|^2} = e^{-\lambda\|x\|^2} e^{2\lambda\langle x,y\rangle} e^{-\lambda\|y\|^2}$$

which is a kernel by (3) and (5) and (1). □

In the following, we let

$$\mathcal{H} = \left\{ f \in \mathrm{AC}([0,1],\mathbb{R}) \mid f' \in L^2([0,1]),\ f(0) = 0 \right\},$$

where $\mathrm{AC}([0,1],\mathbb{R})$ is the space of absolutely continuous functions $f\colon [0,1] \to \mathbb{R}$. Note that we only require f' to be a function satisfying $f(x) = \int_0^x f'(t)\mathrm{d}t$ (i.e., an almost everywhere derivative of f).

Example 1.11
The space $\mathcal{H}$ is a RKHS when given inner product $\langle f, g\rangle_{\mathcal{H}} := \langle f', g'\rangle_{L^2([0,1])}$, and its reproducing kernel is the kernel $k\colon [0,1]\times[0,1] \to [0,1]$, $k(x,y) = \min(x,y)$.

Proof.
For $f \in \mathcal{H}$, it holds that $f(x) = f(0) + \int_0^x f'(t)\mathrm{d}t = \int_0^x f'(t)\mathrm{d}t$ for $x \in [0,1]$, so

$$|f(x)| = \left| \int_0^x f'(t)\mathrm{d}t \right| = |\langle 1_{[0,x]}, f'\rangle_{L^2}| \leq x^{\frac{1}{2}} \|f\|_{\mathcal{H}} \tag{1.2}$$

Thus $f = 0$ in $\mathcal{H}$ implies $f = 0$. The remaining properties making $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ an inner product are immediate from the definition, and $\mathcal{H}$ is clearly a space of functions. It remains to be shown that $\mathcal{H}$ is complete in the norm induced by the inner product: Let (f_n) be a Cauchy-sequence in $\mathcal{H}$ so that (f_n') is Cauchy in $L^2([0,1])$ with limit $g \in L^2([0,1])$. By Equation (1.2) it holds that $f_n(x)$ is Cauchy in $\mathbb{R}$ for all $x \in [0,1]$, hence we may define $f\colon [0,1] \to \mathbb{R}$, $x \mapsto \lim_n f_n(x)$. To see that f is absolutely continuous, note first that $f(0) = \lim_n f_n(0) = 0$, and we calculate for $x \in [0,1]$:

$$f(x) = \lim_n \int_0^x f_n'(t)\mathrm{d}t = \int_0^x g(t)\mathrm{d}t,$$

so $f(x) = f(0) + \int_0^x g(t)\mathrm{d}t$, and g is measurable, hence f is absolutely continuous on $[0,1]$ and has a derivative f' almost everywhere satisfying $f' = g$ almost everywhere, so $f' \in L^2([0,1])$. Hence $f \in \mathcal{H}$ and

$$\|f - f_n\|_{\mathcal{H}} = \lim_n \|f_n' - g\|_{L^2} = 0,$$

so $\mathcal{H}$ is complete, i.e., a Hilbert space. We proceed to show that k is a reproducing kernel for $\mathcal{H}$: Fix $x \in [0,1]$, the function k_x satisfies $k_x(0) = 0$ and is clearly absolutely continuous on $[0,1]$ with a single discontinuity point at x and the almost everywhere defined partial derivative $k_x' = 1_{[0,x]}$ clearly satisfying $k_x' \in L^2([0,1])$, so $k_x \in \mathcal{H}$. For any $f \in \mathcal{H}$, we see that

$$f(x) = \int_0^x f'(t)\mathrm{d}t = \int_0^1 f'(t)k_x'(t)\mathrm{d}t = \langle f, k_x \rangle_{\mathcal{H}},$$

so $\mathcal{H}$ is a reproducing kernel space with kernel k. □

The above example generalizes to a more general form: For $m \geq 1$ let

$$W_0^m := \left\{ f \colon [0,1] \to \mathbb{R} \mid f^{(n)} \in \mathrm{AC}([0,1], \mathbb{R}),\ f^{(n)}(0) = 0 \text{ for } 0 \leq n \leq m-1 \text{ and } f^{(m)} \in L^2([0,1] \right\}.$$

Then W_0^m is a RKHS with norm $\|f\|_{W_0^m} = \left\|f^{(m)}\right\|_{L^2}$. The reproducing kernel of W_0^m can be found by examining the problem:

$$f^{(m)} = g,\ f \in W_0^m.$$

We denote the associated Green's function by G_m, i.e., a function $[0,1]^2 \to \mathbb{R}$ such that $f(x) = \int_0^1 G_m(x,t)g(t)\mathrm{d}t$. It can be shown that

$$G_m(x,y) = 1_{[0,x]}(y)(x-y)^{m-1}((m-1)!)^{-1},$$

and that the function $k_m(x,y) = \int_0^1 G_m(x,t)G_m(y,t)\mathrm{d}t$ is the reproducing kernel for W_0^m. In the case for $m = 1$, we recover the above example.

Let $M(\mathbb{R}^n)$ denote the set of all finite complex-valued Borel measures. For $\mu \in M(\mathbb{R}^n)$, the Fourier transformation of μ is the function $\widehat{\mu} \in C_b(\mathbb{R}^n)$ given by

$$\widehat{\mu}(w) = \int_{\mathbb{R}^n} e^{-i\langle w,t\rangle_{\mathbb{R}^n}} \mathrm{d}\mu(t),\ \text{for } w \in \mathbb{R}^n.$$

This allows us to obtain the following theorem, due to Bochner:

Theorem 1.12 (Bochner's Theorem)**.** For a continuous function $\varphi \colon \mathbb{R}^n \to \mathbb{R}$, the kernel $k(s,t) := \varphi(s-t)$ is positive definite if and only if there is a positive finite Borel measure μ on $\mathbb{R}^n$ such that $\varphi = \widehat{\mu}$.

For proof, see Folland 2016. Bochners thereom allows us to completely characterize the RKHS's of kernels given by $k(s,t) = \varphi(s-t)$ for certain continuous functions $\varphi \colon \mathbb{R}^n \to \mathbb{R}$, for instance the RKHS to the RBF kernel.

Theorem 1.13. Let $\varphi \colon \mathbb{R}^n \to \mathbb{R}$ be a continuous function with $\varphi \in L^1(\mathbb{R}^n)$ and its Fourier transform $\widehat{\varphi} \in L^1(\mathbb{R}^n)$ such that the function $k(s,t) := \varphi(s-t)$ is a positive definite kernel. For $f \in L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$ define the number

$$\|f\|_k^2 := (2\pi)^{-n} \int_{\mathbb{R}^n} \frac{|\hat{f}(w)|^2}{\hat{\varphi}(w)} \mathrm{d}w,$$

Then the RKHS associated to $k(s,t) = \varphi(s-t)$ is the set of functions

$$\mathcal{H} := \cap \left\{ f \in C(\mathbb{R}^n) \cap L^2(\mathbb{R}^n) \mid \|f\|_k < \infty \right\},$$

with inner product defined by

$$\langle f, h \rangle_k := (2\pi)^{-n} \int_{\mathbb{R}^n} \frac{\hat{f}(w)\overline{\hat{h}(w)}}{\hat{\varphi}(w)} \mathrm{d}w.$$

Proof.
Note first that the Plancherel Theorem implies if $f \in \mathcal{H}$ then $\widehat{f} \in L^2(\mathbb{R}^n)$. It is straightforward to check that $\mathcal{H}$ is a pre-Hilbert space of real-valued functions with $\langle \cdot, \cdot \rangle_k$, and the only non-trivial thing to show is completeness: Pick a Cauchy sequence $(f_n) \subseteq \mathcal{H}$. Then $\frac{\widehat{f_n}}{\sqrt{\widehat{\varphi}}} \in L^2(\mathbb{R}^n)$ is Cauchy, hence it has a limit $g \in L^2(\mathbb{R}^n)$ which satisfies

$$\int_{\mathbb{R}^n} |g(x)\sqrt{\widehat{\varphi}}(x)| \mathrm{d}x \leq \|g\|_{L^2} \|\widehat{\varphi}\|_{L^1}^{\frac{1}{2}}$$

so $g \cdot \sqrt{\widehat{\varphi}} \in L^1(\mathbb{R}^n)$ and analogously one sees that $g \cdot \sqrt{\widehat{\varphi}} \in L^2(\mathbb{R}^n)$, since $\widehat{\varphi} \in L^\infty(\mathbb{R}^n)$. We define a candidate function $f(x)$ pointwise by

$$f(x) = (2\pi)^{-n} \int_{\mathbb{R}^n} g(w)\sqrt{\widehat{\varphi}(w)} e^{i\langle x, w \rangle_{\mathbb{R}^n}} \mathrm{d}w,$$

And it is a priori a complex-valued continuous square integrable function satisfying $\frac{\widehat{f}}{\sqrt{\widehat{\varphi}}} = g$. It remains to be shown that f is real-valued: By the Fourier inversion theorem it holds for all $x \in \mathbb{R}^n$ that

$$\begin{aligned}
|f(x) - f_n(x)| &= \left| (2\pi)^{-n} \int_{\mathbb{R}^n} g(w)\sqrt{\widehat{\varphi}(w)} e^{i\langle w, x \rangle_{\mathbb{R}^n}} - \widehat{f_n}(w) e^{i\langle w, x \rangle_{\mathbb{R}^n}} \mathrm{d}w \right| \\
&\leq (2\pi)^{-n} \int_{\mathbb{R}^n} \left| \sqrt{\widehat{\varphi}(w)} (g(w) - \frac{\widehat{f_n}(w)}{\sqrt{\widehat{\varphi}(w)}} \right| \mathrm{d}w \\
&\leq (2\pi)^{-n} \left\| g - \frac{\widehat{f_n}}{\sqrt{\widehat{\varphi}}} \right\|_{L^2} \|\widehat{\varphi}\|_{L^1} \to 0,
\end{aligned}$$

so $f(x) \in \mathbb{R}$ and $f \in \mathcal{H}$. And since $\frac{\widehat{f}}{\sqrt{\widehat{\varphi}}} = g$, we see that $f_n \to f$ in $\mathcal{H}$.

We proceed to show that $\mathcal{H}$ is a RKHS now: We first show that $k_x \in \mathcal{H}$ for all $x \in \mathbb{R}^n$. Let $x \in \mathbb{R}^n$, then the Fourier transform of k_x at $w \in \mathbb{R}^n$ satisfies, by the Fourier inversion theorem (see e.g. Folland 2016), the identity

$$\begin{aligned}
\widehat{k_x}(w) = \int_{\mathbb{R}^n} e^{-i\langle w, t \rangle_{\mathbb{R}^n}} k_x(t) \mathrm{d}t &= \int_{\mathbb{R}^n} e^{-i\langle w, t \rangle_{\mathbb{R}^n}} \varphi(x - t) \mathrm{d}t \\
&= \int_{\mathbb{R}^n} e^{-i\langle w, t + x \rangle_{\mathbb{R}^n}} \varphi(t) \mathrm{d}t \\
&= e^{-i\langle w, x \rangle_{\mathbb{R}^n}} \hat{\varphi}(w).
\end{aligned}$$

Hence we find that

$$\|k_x\|_k^2 = \int_{\mathbb{R}^n} \frac{|\widehat{k_x}(w)|^2}{\hat{\varphi}(w)} \mathrm{d}w \leq \int_{\mathbb{R}^n} |\hat{\varphi}(w)| \mathrm{d}w < \infty,$$

so $k_x \in \mathcal{H}$. Moreover, for $f \in \mathcal{H}$ and $x \in \mathbb{R}^n$:

$$\begin{aligned}
\langle f, k_x \rangle_k &= (2\pi)^{-n} \int_{\mathbb{R}^n} \frac{\widehat{k_x}(w)\overline{\widehat{f}(w)}}{\widehat{\varphi}(w)} \mathrm{d}w \\
&= \int_{\mathbb{R}^n} \overline{\widehat{f}(w)} e^{-i\langle w, x \rangle_{\mathbb{R}^n}} \mathrm{d}w \\
&= \int_{\mathbb{R}^n} \widehat{f}(w) e^{i\langle w, x \rangle_{\mathbb{R}^n}} \mathrm{d}w \\
&= f(x),
\end{aligned}$$

by the inversion theorem and continuity of f. □

The Representer Theorem

Generalized Representer Theorems

In statistics and learning theory, we often set a goal to solve some reguralized minimalization problem associated to some data and a risk function, and one such way is to use a RKHS as the place to look for solutions. However, this could prove difficult, since it might not be computationally feasible if $\mathcal{H}$ is infinite dimensional.

However, as we will see, under certain conditions, we can ensure that the solution is centered around kernels specified by our training data. The following example of this is due to Kimeldorf and Wahba 1970:

Proposition 2.1. Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ be training samples, and assume that $\mathcal{X}$ is equipped with a positive definite kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and where

$$\mathcal{F} = \left\{ f \in \mathbb{R}^{\mathcal{X}} \;\middle|\; f = \sum_{1 \leq i \leq \infty} \beta_i k(\cdot, z_i),\ \beta_i \in \mathbb{R},\ z_i \in \mathcal{X} \text{ and } \|f\|_{\mathcal{H}} < \infty \right\},$$

then if a solution f_* of

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{1 \leq i \leq n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \ \ \lambda > 0,$$

exists, it has the form

$$f_* = \sum_{1 \leq i \leq m} \alpha_i k(\cdot, x_i),$$

for some $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$.

While we will not reproduce the proof, we will show a generalized version of this, where we pose less restrictions on the minimization problem:

Theorem 2.2 (Nonparametric Representer Theorem)**.** Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$ be training samples, and assume that $\mathcal{X}$ is equipped with a positive definite

kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with associated RKHS $\mathcal{H}_k$, and that g is a strictly increasing function on $\mathbb{R}_{\geq 0}$, $C\colon (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$ is any cost function and where

$$\mathcal{F} = \left\{ f \in \mathbb{R}^{\mathcal{X}} \;\middle|\; f = \sum_{1 \leq i \leq \infty} \beta_i k(\cdot, z_i),\ \beta_i \in \mathbb{R},\ z_i \in \mathcal{X} \text{ and } \|f\|_{\mathcal{H}_k} < \infty \right\},$$

then any

$$f_* = \operatorname*{argmin}_{f \in \mathcal{F}} C\left((x_1, y_1, f(x_1)), (x_2, y_2, f(x_2)), \ldots (x_n, y_n, f(x_n))\right) + g(\|f\|_{\mathcal{H}_k}),$$

is of the form

$$f_* = \sum_{1 \leq i \leq m} \alpha_i k(\cdot, x_i),$$

for some $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$.

Proof.
let $\mathcal{H}_k$ denote the canonical feature space of k and $\mathcal{X}$. And let $H_0 = \operatorname{span}_{\mathbb{R}}\{(k(\cdot, x_i))_{i=1}^n\}$. This is a finite dimensional subspace, hence it is closed, and thus we can decompose $\mathcal{H}_k = H_0 \oplus H_0^{\perp}$. It follows that given $f \in \mathcal{F}$, we can write it as $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i) + v$ for some $v \in H_0^{\perp}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Then, since $\mathcal{H}_k$ is RKHS, we see for our samples, $x_1, \ldots, x_n$, that

$$f(x_j) = \langle f, k(\cdot, x_j) \rangle = \left\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), k(\cdot, x_j) \right\rangle + \underbrace{\langle v, k(\cdot, x_j) \rangle}_{=0} = \sum_{i=1}^n \alpha_i k(x_j, x_i).$$

By the Parallelogram law and the strict monotonicity of g, we see that

$$\begin{aligned} g(\|f\|) = g\left(\left\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) + v \right\|_{\mathcal{H}_k} \right) &= g\left(\sqrt{ \left\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) \right\|_{\mathcal{H}_k}^2 + \|v\|_{\mathcal{H}_k}^2 } \right) \\ &\geq g\left(\left\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) \right\|_{\mathcal{H}_k} \right) \end{aligned}$$

We conclude that if f is a solution to the minimization problem then $v = 0$, for else $f - v$ is a better solution, since the first term does not depend on v and the latter is smaller when $v = 0$. □

The above result is very important in various fields - essentially it allows us with numerical precision solve various minimization problems, and often we may deploy various well-researched algorithms to find proper approximations of the desired solutions.

We may extend this result even further, to a semi-parametric setting:

Theorem 2.3 (Semiparamentric Representer Theorem)**.** Assuming the assumptions of Theorem 2.2 together with the additional assumptions that we are given a finite sequence of real-valued functions (φ_n), $n = 1, .., N$, on $\mathcal{X}$ such that the matrix $(\varphi_i(x_j))_{i,j} \in \mathbb{R}^{n\times N}$ has rank N, then any solution $\tilde{f} = f + h$, with $f \in \mathcal{F}$ and $h \in \operatorname{span}_{\mathbb{R}} \{\varphi_j\}$, minimizing

$$C\left((x_1, y_1, \tilde{f}(x_1)), \ldots, (x_n, y_n, \tilde{f}(x_n))\right) + g(\|f\|_{\mathcal{H}_k})$$

is of the form

$$\tilde{f} = \sum_{i=1}^{n} \alpha_i k_{x_i} + \sum_{j=1}^{N} \beta_j \varphi_j,$$

with unique coefficient $\beta_j \in \mathbb{R}$ for $j = 1, \ldots, N$.

Proof.
Analogously as in the proof of Theorem 2.2, we may decompose f in the terms

$$\tilde{f}_* = f + h,$$

and we see that C does not depend on the term which is orthogonal to $H_0 = \operatorname{span}_{\mathbb{R}} \{k_{x_i} \mid i = 1, \ldots, n\}$. Again, the requirement that g is strictly increasing and the Parallelogram law implies that any solution must be of the desired form

$$\tilde{f} = \sum_{i=1}^{n} \alpha_i k_{x_i} + \sum_{j=1}^{N} \beta_j \varphi_j.$$

To see uniqueness of β_i, $i = 1, \ldots, N$, note the rank assumption on the matrix $(\varphi_i(x_j))$ implies that $\{(\varphi_j(x_1), \ldots, \varphi_j(x_n))^T\}_{j=1}^{N}$ is linearly independent, hence the vector $(\beta_1, \ldots, \beta_M)$ must be the unique solution to $\tilde{f} - \sum_{i=1}^{n} \alpha_i k_{x_i} = \sum_{j=1}^{n} \beta_j \varphi_j$ □

Applications and consequences

The Representer Theorem allows for a great range of general problems to be reduced to a form which can be solved either algebraically or numerically. We will examine a few of these including kernelised ridge regression, support vector classification and more. For practical reasons, we introduce some notation:

Given a set of samples $S_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathbb{R})^n$, we define

$$\begin{aligned} \mathbf{k}(x) &:= (k(x_1, x), k(x_2, x), \ldots, k(x_n, x))^T, \\ \mathbf{K} &:= (k(x_i, x_j))_{i,j=1}^n, \\ \mathbf{x} &= (x_1, \ldots, x_n)^T \text{ and} \\ \mathbf{y} &= (y_1, \ldots, y_n)^T. \end{aligned}$$

Kernel Ridge Regression

Let S_n be a set of n samples, and consider the cost function corresponding to the sum of squared errors $\mathcal{R}(S_n, f) = \sum_{i=1}^n (y_i - f(x_i))^2$ for $f \colon \mathcal{X} \to \mathbb{R}$. Given some set of functions F, we could consider the regression problem

$$\min_{f \in F} \mathcal{R}(S_n, f),$$

however, this will often lead to issues such as overfitting, since a function is only being valued based on its empirical fit. Instead, pick a positive definite kernel k with canonical RKHS $\mathcal{H}_k$, $\lambda > 0$ and consider the problem

$$\min_{f \in \mathcal{H}_k} \mathcal{R}(S_n, f) + \lambda \|f\|_{\mathcal{H}_k}^2 . \tag{3.1}$$

This problem is know as a Kernel Ridge Regression problem, and setting $g(x) = \lambda x^2$, we see that Theorem 2.2 forces any solution to be of the form

$$f_* = \sum_{i=1}^n \alpha_i k_{x_i},$$

for some $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$. Note that $(f_*(x_1), f_*(x_2), \ldots, f_*(x_n))^T = \mathbf{K}\boldsymbol{\alpha}$, so that $\|f_*\|^2_{\mathcal{H}_k} = \boldsymbol{\alpha}^t \mathbf{K}\boldsymbol{\alpha}$. Hence we may rewrite 3.1 as the following problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})^T (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha},$$

and we let $J_\lambda(f) := \mathcal{R}(S_n, f) + \lambda \|f\|^2_{\mathcal{H}_k} = (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})^T (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha}$. Note that this is a differentiable function in $\boldsymbol{\alpha}$, and the Hession matrix of $J_\lambda(f)$ with respect to $\boldsymbol{\alpha}$ is positive-definite, since it is $2(\mathbf{K}^2 + \lambda \mathbf{K})$. Therefore $J_\lambda(f)$ is a convex function in $\boldsymbol{\alpha}$, and it will have a minimum when the gradient is 0, and we see that

$$\mathbf{y} = (\mathbf{K} + \lambda)\boldsymbol{\alpha} \Longrightarrow 0 = \nabla 2\mathbf{K}((\mathbf{K} + \lambda)\boldsymbol{\alpha} - \mathbf{y})$$

Since $\mathbf{K}$ is positive semi-definite and $\lambda > 0$, the matrix $(\mathbf{K} + \lambda)$ is invertible, hence a valid solution is of the form $\hat{\alpha} = (\mathbf{K} + \lambda)^{-1}\mathbf{y}$. We include figures showcasing the effect of the penalty parameter λ on the above process applied to some simulated data and the RBF kernel with different σ^2 parameters.

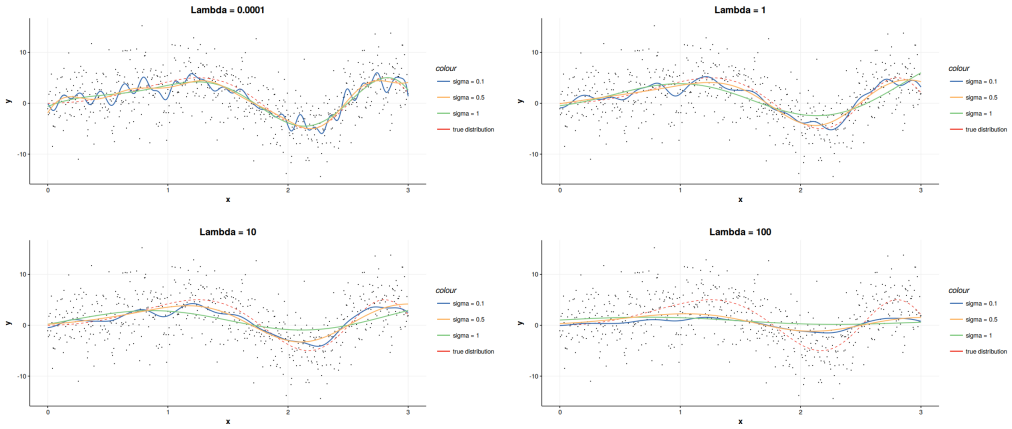


Figure (3.1) Simulated data with respect to $y \sim f(x) + \varepsilon_i$, where $f(x) = 5\sin(x^2)$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. The parameter λ controls how much we penalize lack of "smoothness" of f - very low λ might result in overfitting and very high values results in underfitting.

BOUNDING THE EXPECTED LOSS

In the following, we will assume that all samples in S_n are sampled i.i.d. from an unknown fixed distribution $p(X, Y)$. Let $\ell(y_1, y_2)$ denote a loss function. For a classification hypothesis $h \colon \mathcal{X} \to \mathbb{R}$, we define the *expected risk*, $I(h)$, and the *empirical*

risk, $I_{S_n}(h)$, by

$$I(h) := \int_{\mathcal{X}\times\mathbb{R}} \ell(h(x), y)\mathrm{d}p(x, y)$$
$$I_{S_n}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i).$$

We then define the *generalization error*, $D_{S_n}(h)$, of h to be $D_{S_n}(h) := I(h) - I_{S_n}(h)$.

Definition 3.1. Let H denote a class of hypotheses. A *Generalization Bound* is any inequality such that for $\delta \in (0, 1)$ it that

$$\mathbb{P}(|D_S(h)| \geq B(H, \delta, n, h) \text{ for some } h \in H) \leq \delta,$$

where $B(H, \delta, n, h)$ is a function of h taking values in $(0, \infty)$ for fixed H, δ and sample size n.

If one can choose $H = \mathcal{H}_k$ for some positive definite kernel k and choose a function B such that for fixed $\delta \in (0, 1)$ and $n \geq 1$, the function $g\colon h \mapsto B(H, \delta, n, h)$ is only dependant on $\|h\|_{\mathcal{H}_k}$ and is a strictly increasing, then we can apply the representer theorem with $g(\|h\|_{\mathcal{H}_k}) = B(H, \delta, n, h)$ so that with high probability, we bound the expected risk by reducing the empirical risk $I(h)$ when choosing h.

Applying Vapnik-Chervonenkis theory (see e.g. Abu-Mostafa, Magdon-Ismail, and Lin 2012), one can show that for the RKHS $\mathcal{H}_k$ of linear separators on $\mathbb{R}^n$ that the following holds:

Theorem 3.2. Let $\mathcal{H}_k$ denote the RKHS of linear separators in $\mathbb{R}^n$ and let $\delta \in (0, 1)$. Then

$$\mathbb{P}\left(\exists h = (w, b) \mid |D_S(h)| \geq \sqrt{\frac{8\ln\left(2((2n)^{1+\lceil 1\|w\|^2\rceil} + 1\right)(1 + \lceil\|w\|^2\rceil)\lceil\|w\|^2\rceil\delta^{-1}}{n}}\right) \leq \delta,$$

where the size of $S = n$, where the loss functional ℓ is the $0 - 1$ loss and the target samples $y \in \{\pm 1\}$, and the notation that for $h = (w, b)$ we mean $h(x) = \mathrm{sgn}(\langle w, x\rangle + b)$.

For further reading we refer to Bousquet and Elisseeff 2001, and for a concrete example in the choice of g, we refer to the theory of pattern recognition found in Vapnik 1998.

SUPPORT VECTOR MACHINE FORMULATION

A popular and empirically effective algorithm for classification is the Support Vector Machine (SVM). We now show why the representer theorem ensures that the SVM problems have a "nice" formulation. Let us first state the kernel SVM problem for

some positive definite kernel k: Let $(x_1, y_1), \ldots, (x_n, y_n)$ be given with $y_j \in \{\pm 1\}$ for $j = 1, 2, \ldots, n$. Then the SVM problem formulation without offset is:

$$\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_k}, \tag{3.2}$$

where $\ell_{hinge}(u) = \max(1 - u, 0)$ and $\lambda > 0$. By Theorem 2.2, this is equivalent to

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i \pi_i(\mathbf{K}\boldsymbol{\alpha})) + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha},$$

where π_i denotes the $i'th$ coordinate projection map. This is not a smooth problem by the involvement of ℓ_{hinge}. To deal with this, using Lagrangian theory and forming the Primal formulation of the problem and then its corresponding dual formulation, one can arrive that the equivalent formulation of Equation (3.2) given by:

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad & 2\boldsymbol{\alpha}^T \boldsymbol{y} - \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \\ \text{such that} \quad & 0 \leq y_i \alpha_i \leq (2\lambda n)^{-1}, \ i = 1, 2, \ldots, n. \end{aligned}$$

This formulation is a Quadratic Programming Problem, which can be solved numerically using different algorithms. For the specific proofs regarding the arrival at the Dual formulation, we refer to e.g. Schölkopf and Smola 2002.

Bibliography

[1] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*. Vol. 4. AMLBook New York, NY, USA: 2012.

[2] Olivier Bousquet and André Elisseeff. "Algorithmic stability and generalization performance". In: *Advances in Neural Information Processing Systems*. 2001, pp. 196–202.

[3] Gerald B Folland. *A course in abstract harmonic analysis*. Vol. 29. CRC press, 2016.

[4] George S Kimeldorf and Grace Wahba. "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines". In: *The Annals of Mathematical Statistics* 41.2 (1970), pp. 495–502.

[5] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[6] Vladimir Vapnik. *Statistical learning theory. 1998*. Wiley, New York, 1998.