Function Spaces

A function space is a set of functions $\mathcal{F}$ that has some structure. Often a nonparametric regression function or classifier is chosen to lie in some function space, where the assumed structure is exploited by algorithms and theoretical analysis. Here we review some basic facts about function spaces.

As motivation, consider nonparametric regression. We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we want to estimate $m(x) = \mathbb{E}(Y|X = x)$. We cannot simply choose m to minimize the training error $\sum_i (Y_i - m(X_i))^2$ as this will lead to interpolating the data. One approach is to minimize $\sum_i (Y_i - m(X_i))^2$ while restricting m to be in a well behaved function space.

1 Hilbert Spaces

Let V be a vector space. A *norm* is a mapping $\|\cdot\| : V \to [0, \infty)$ that satisfies

1. $\|x + y\| \leq \|x\| + \|y\|$.
2. $\|ax\| = a\|x\|$ for all $a \in \mathbb{R}$.
3. $\|x\| = 0$ implies that $x = 0$.

An example of a norm on $V = \mathbb{R}^k$ is the *Euclidean norm* $\|x\| = \sqrt{\sum_i x_i^2}$. A sequence $x_1, x_2, \ldots$ in a normed space is a *Cauchy sequence* if $\|x_m - x_n\| \to 0$ as $m, n \to \infty$. The space is *complete* if every Cauchy sequence converges to a limit. A complete, normed space is called a *Banach space*.

An *inner product* is a mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies, for all $x, y, z \in V$ and $a \in \mathbb{R}$:

1. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$
2. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$
3. $\langle x, ay \rangle = a\langle x, y \rangle$
4. $\langle x, y \rangle = \langle y, x \rangle$

An example of an inner product on $V = \mathbb{R}^k$ is $\langle x, y \rangle = \sum_i x_i y_i$. Two vectors x and y are *orthogonal* if $\langle x, y \rangle = 0$. An inner product defines a norm $\|v\| = \sqrt{\langle v, v \rangle}$. We then have the Cauchy-Schwartz inequality

$$|\langle x, y \rangle| \leq \|x\| \ \|y\|. \tag{1}$$

A *Hilbert space* is a complete, inner product space. Every Hilbert space is a Banach space but the reverse is not true in general. In a Hilbert space, we write $f_n \to f$ to mean that $||f_n - f|| \to 0$ as $n \to \infty$. Note that $||f_n - f|| \to 0$ does NOT imply that $f_n(x) \to f(x)$. For this to be true, we need the space to be a reproducing kernel Hilbert space which we discuss later.

If V is a Hilbert space and L is a closed subspace then for any $v \in V$ there is a unique $y \in L$, called the *projection* of v onto L, which minimizes $\|v - z\|$ over $z \in L$. The set of elements orthogonal to every $z \in L$ is denoted by $L^\perp$. Every $v \in V$ can be written uniquely as $v = w + z$ where z is the projection of v onto L and $w \in L^\perp$. In general, if L and M are subspaces such that every $\ell \in L$ is orthogonal to every $m \in M$ then we define the *orthogonal sum* (or *direct sum*) as

$$L \oplus M = \{\ell + m : \ell \in L, m \in M\}. \tag{2}$$

A set of vectors $\{e_t, t \in T\}$ is *orthonormal* if $\langle e_s, e_t \rangle = 0$ when $s \neq t$ and $\|e_t\| = 1$ for all $t \in T$. If $\{e_t, t \in T\}$ are orthonormal, and the only vector orthogonal to each e_t is the zero vector, then $\{e_t, t \in T\}$ is called an *orthonormal basis*. Every Hilbert space has an orthonormal basis. A Hilbert space is *separable* if there exists a countable orthonormal basis.

Theorem 1 *Let V be a separable Hilbert space with countable orthonormal basis $\{e_1, e_2, \ldots\}$. Then, for any $x \in V$, we have $x = \sum_{j=1}^\infty \theta_j e_j$ where $\theta_j = \langle x, e_j \rangle$. Furthermore, $\|x\|^2 = \sum_{j=1}^\infty \theta_j^2$, which is known as* Parseval's identity.

The coefficients $\theta_j = \langle x, e_j \rangle$ are called *Fourier coefficients*.

The set $\mathbb{R}^d$ with inner product $\langle v, w \rangle = \sum_j v_j w_j$ is a Hilbert space. Another example of a Hilbert space is the set of functions $f : [a, b] \to \mathbb{R}$ such that $\int_a^b f^2(x)dx < \infty$ with inner product $\int f(x)g(x)dx$. This space is denoted by $L_2(a, b)$.

2 L_p Spaces

Let $\mathcal{F}$ be a collection of functions taking $[a, b]$ into $\mathbb{R}$. The L_p norm on $\mathcal{F}$ is defined by

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{1/p} \tag{3}$$

where $0 < p < \infty$. For $p = \infty$ we define

$$\|f\|_\infty = \sup_x |f(x)|. \tag{4}$$

Sometimes we write $\|f\|_2$ simply as $\|f\|$. The space $L_p(a,b)$ is defined as follows:

$$L_p(a,b) = \left\{ f : [a,b] \to \mathbb{R} : \ \|f\|_p < \infty \right\}. \tag{5}$$

Every L_p is a Banach space. Some useful inequalities are:

Cauchy-Schwartz	$\left(\int f(x)g(x)dx\right)^2 \leq \int f^2(x)dx \int g^2(x)dx$
Minkowski	$\|f+g\|_p \leq \|f\|_p + \|g\|_p$ where $p > 1$
Hölder	$\|fg\|_1 \leq \|f\|_p \|g\|_q$ where $(1/p) + (1/q) = 1$.

Special Properties of L_2. As we mentioned earlier, the space $L_2(a,b)$ is a Hilbert space. The *inner product* between two functions f and g in $L_2(a,b)$ is $\int_a^b f(x)g(x)dx$ and the *norm* of f is $\|f\|^2 = \int_a^b f^2(x)\,dx$. With this inner product, $L_2(a,b)$ is a separable Hilbert space. Thus we can find a countable orthonormal basis $\phi_1, \phi_2, \ldots$; that is, $\|\phi_j\| = 1$ for all j, $\int_a^b \phi_i(x)\,\phi_j(x)dx = 0$ for $i \neq j$ and the only function that is orthogonal to each ϕ_j is the zero function. (In fact, there are many such bases.) It follows that if $f \in L_2(a,b)$ then

$$f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x) \tag{6}$$

where

$$\theta_j = \int_a^b f(x)\,\phi_j(x)\,dx \tag{7}$$

are the coefficients. Also, recall Parseval's identity

$$\int_a^b f^2(x)dx = \sum_{j=1}^{\infty} \theta_j^2. \tag{8}$$

The set of functions

$$\left\{ \sum_{j=1}^{n} a_j \phi_j(x) : \ a_1, \ldots, a_n \in \mathbb{R} \right\} \tag{9}$$

is the called the *span* of $\{\phi_1, \ldots, \phi_n\}$. The projection of $f = \sum_{j=1}^{\infty} \theta_j \phi_j(x)$ onto the span of $\{\phi_1, \ldots, \phi_n\}$ is $f_n = \sum_{j=1}^{n} \theta_j \phi_j(x)$. We call f_n the n-term *linear approximation* of f. Let Λ_n denote all functions of the form $g = \sum_{j=1}^{\infty} a_j \phi_j(x)$ such that at most n of the a_j's are non-zero. Note that Λ_n is not a linear space, since if $g_1, g_2 \in \Lambda_n$ it does not follow that $g_1 + g_2$ is in Λ_n. The best approximation to f in Λ_n is $f_n = \sum_{j \in A_n} \theta_j\,\phi_j(x)$ where A_n are the n indices corresponding to the n largest $|\theta_j|$'s. We call f_n the n-term *nonlinear approximation* of f.

The *Fourier basis* on $[0,1]$ is defined by setting $\phi_1(x) = 1$ and

$$\phi_{2j}(x) = \frac{1}{\sqrt{2}}\cos(2j\pi x), \quad \phi_{2j+1}(x) = \frac{1}{\sqrt{2}}\sin(2j\pi x), \quad j = 1, 2, \ldots \tag{10}$$

The *cosine basis* on $[0,1]$ is defined by

$$\phi_0(x) = 1, \qquad \phi_j(x) = \sqrt{2}\cos(2\pi j x), \qquad j = 1, 2, \ldots. \tag{11}$$

The *Legendre basis* on $(-1,1)$ is defined by

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{1}{2}(5x^3 - 3x), \ \ldots \tag{12}$$

These polynomials are defined by the relation

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2 - 1)^n. \tag{13}$$

The Legendre polynomials are orthogonal but not orthonormal, since

$$\int_{-1}^{1} P_n^2(x)dx = \frac{2}{2n+1}. \tag{14}$$

However, we can define modified Legendre polynomials $Q_n(x) = \sqrt{(2n+1)/2}\ P_n(x)$ which then form an orthonormal basis for $L_2(-1,1)$.

The *Haar basis* on [0,1] consists of functions

$$\left\{ \phi(x), \psi_{jk}(x) : \ j = 0, 1, \ldots, \ k = 0, 1, \ldots, 2^j - 1 \right\} \tag{15}$$

where

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \le x < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ and

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \le x \le \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} < x \le 1. \end{cases} \tag{17}$$

This is a doubly indexed set of functions so when f is expanded in this basis we write

$$f(x) = \alpha\phi(x) + \sum_{j=1}^{\infty}\sum_{k=1}^{2^j-1}\beta_{jk}\psi_{jk}(x) \tag{18}$$

where $\alpha = \int_0^1 f(x)\,\phi(x)\,dx$ and $\beta_{jk} = \int_0^1 f(x)\,\psi_{jk}(x)\,dx$. The Haar basis is an example of a *wavelet basis*.

Let $[a,b]^d = [a,b] \times \cdots \times [a,b]$ be the d-dimensional cube and define

$$L_2\left([a,b]^d\right) = \left\{ f : [a,b]^d \to \mathbb{R} : \int_{[a,b]^d} f^2(x_1,\ldots,x_d)\, dx_1 \ldots dx_d < \infty \right\}. \tag{19}$$

Suppose that $\mathcal{B} = \{\phi_1, \phi_2, \ldots\}$ is an orthonormal basis for $L_2([a,b])$. Then the set of functions

$$\mathcal{B}^d = \mathcal{B} \otimes \cdots \otimes \mathcal{B} = \left\{ \phi_{i_1}(x_1)\, \phi_{i_2}(x_2) \cdots \phi_{i_d}(x_d) : \; i_1, i_2, \ldots, i_d \in \{1,2,\ldots,\} \right\}, \tag{20}$$

is called the *tensor product* of $\mathcal{B}$, and forms an orthonormal basis for $L_2([a,b]^d)$.

3 Hölder Spaces

Let β be a positive integer.[1] Let $T \subset \mathbb{R}$. The Holder space $H(\beta, L)$ is the set of functions $g : T \to \mathbb{R}$ such that

$$|g^{(\beta-1)}(y) - g^{(\beta-1)}(x)| \leq L|x-y|, \quad \text{for all } x, y \in T. \tag{21}$$

The special case $\beta = 1$ is sometimes called the Lipschitz space. If $\beta = 2$ then we have

$$|g'(x) - g'(y)| \leq L\,|x-y|, \quad \text{for all } x, y.$$

Roughly speaking, this means that the functions have bounded second derivatives.

There is also a multivariate version of Holder spaces. Let $T \subset \mathbb{R}^d$. Given a vector $s = (s_1, \ldots, s_d)$, define $|s| = s_1 + \cdots + s_d$, $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$ and

$$D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

The Hölder class $H(\beta, L)$ is the set of functions $g : T \to \mathbb{R}$ such that

$$|D^s g(x) - D^s g(y)| \leq L\|x-y\|^{\beta - |s|} \tag{22}$$

for all x, y and all s such that $|s| = \beta - 1$.

If $g \in H(\beta, L)$ then $g(x)$ is close to its Taylor series approximation:

$$|g(u) - g_{x,\beta}(u)| \leq L\|u-x\|^\beta \tag{23}$$

where

$$g_{x,\beta}(u) = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(u-x)^s}{s!} D^s g(x). \tag{24}$$

[1] It is possible to define Holder spaces for non-integers but we will not need this generalization.

In the case of $\beta = 2$, this means that

$$|g(u) - [g(x) + (x-u)^T \nabla g(x)]| \leq L||x-u||^2.$$

We will see that in function estimation, the optimal rate of convergence over $H(\beta, L)$ under L_2 loss is $O(n^{-2\beta/(2\beta+d)})$.

4 Sobolev Spaces

Let f be integrable on every bounded interval. Then f is *weakly differentiable* if there exists a function f' that is integrable on every bounded interval, such that $\int_x^y f'(s)ds = f(y) - f(x)$ whenever $x \leq y$. We call f' the *weak derivative* of f. Let $D^j f$ denote the j^{th} weak derivative of f.

The *Sobolev space of order* m is defined by

$$W_{m,p} = \Big\{ f \in L_p(0,1) : \ \|D^m f\| \in L_p(0,1) \Big\}. \tag{25}$$

The *Sobolev ball of order* m *and radius* c is defined by

$$W_{m,p}(c) = \Big\{ f : \ f \in W_{m,p}, \ \|D^m f\|_p \leq c \Big\}. \tag{26}$$

For the rest of this section we take $p = 2$ and write W_m instead of $W_{m,2}$

Theorem 2 *The Sobolev space* W_m *is a Hilbert space under the inner product*

$$\langle f, g \rangle = \sum_{k=0}^{m-1} f^{(k)}(0) g^{(k)}(0) + \int_0^1 f^{(k)}(x) g^{(k)}(x)\, dx. \tag{27}$$

Define

$$K(x,y) = \sum_{k=1}^{m-1} \frac{1}{k!} x^k y^k + \int_0^{x \wedge y} \frac{(x-u)^{m-1}(y-u)^{m-1}}{(m-1)!^2} du. \tag{28}$$

Then, for each $f \in W_m$ we have

$$f(y) = \langle f, K(\cdot, y) \rangle \tag{29}$$

and

$$K(x,y) = \langle K(\cdot, x), K(\cdot, y) \rangle. \tag{30}$$

We say that K is a kernel for the space and that W_m is a *reproducing kernel Hilbert space* or *RKHS*. See Section 7 for more on reproducing kernel Hilbert spaces.

It follows from Mercer's theorem (Theorem 4) that there is an orthonormal basis $\{e_1, e_2, \ldots, \}$ for $L_2(a, b)$ and real numbers $\lambda_1, \lambda_2, \ldots$ such that

$$K(x, y) = \sum_{j=1}^{\infty} \lambda_j \, e_j(x) \, e_j(y). \tag{31}$$

The functions e_j are eigenfunctions of K and the λ_j's are the corresponding eigenvalues,

$$\int K(x, y) \, e_j(y) \, dy = \lambda_j e_j(x). \tag{32}$$

Hence, the inner product defined in (27) can be written as

$$\langle f, g \rangle = \sum_{j=0}^{\infty} \frac{\theta_j \, \beta_j}{\lambda_j} \tag{33}$$

where $f(x) = \sum_{j=0}^{\infty} \theta_j e_j(x)$ and $g(x) = \sum_{j=0}^{\infty} \beta_j e_j(x)$.

Next we discuss how the functions in a Sobolev space can be parameterized by using another convenient basis. An *ellipsoid* is a set of the form

$$\Theta = \left\{ \theta : \; \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq c^2 \right\} \tag{34}$$

where a_j is a sequence of numbers such that $a_j \to \infty$ as $j \to \infty$. If Θ is an ellipsoid and if $a_j^2 \sim (\pi j)^{2m}$ as $j \to \infty$, we call Θ a *Sobolev ellipsoid* and we denote it by $\Theta_m(c)$.

Theorem 3 *Let* $\{\phi_j, j = 0, 1, \ldots\}$ *be the* Fourier basis*:*

$$\phi_1(x) = 1, \quad \phi_{2j}(x) = \frac{1}{\sqrt{2}} \cos(2j\pi x), \quad \phi_{2j+1}(x) = \frac{1}{\sqrt{2}} \sin(2j\pi x), \quad j = 1, 2, \ldots \tag{35}$$

Then,

$$W_m(c) = \left\{ f : \; f = \sum_{j=1}^{\infty} \theta_j \phi_j, \; \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq c^2 \right\} \tag{36}$$

where $a_j = (\pi j)^m$ *for* j *even and* $a_j = (\pi(j-1))^m$ *for* j *odd. Thus, a Sobolev space corresponds to a Sobolev ellipsoid with* $a_j \sim (\pi j)^{2m}$*.*

Note that (36) allows us to define the Sobolev space W_m for fractional values of m as well as integer values. A multivariate version of Sobolev spaces can be defined as follows. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be non-negative integers and define $|\alpha| = \alpha_1 + \cdots + \alpha_d$. Given $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ write $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ and

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}. \tag{37}$$

Then the Sobolev space is defined by

$$W_{m,p} = \left\{ f \in L_p\left([a,b]^d\right) : \ D^\alpha f \in L_p([a,b]^d) \text{ for all } |\alpha| \leq m \right\}. \tag{38}$$

We will see that in function estimation, the optimal rate of convergence over $W_{\beta,2}$ under L_2 loss is $O(n^{-2\beta/(2\beta+d)})$.

5 Besov Spaces*

Functions in Sobolev spaces are homogeneous, meaning that their smoothness does not vary substantially across the domain of the function. Besov spaces are richer classes of functions that include inhomogeneous functions.

Let

$$\Delta_h^{(r)} f(x) = \sum_{k=0}^{r} (-1)^k \binom{r}{k} f(x + kh). \tag{39}$$

Thus, $\Delta_h^{(0)} f(x) = f(x)$ and

$$\Delta_h^{(r)} f(x) = \Delta_h^{(r-1)} f(x+h) - \Delta_h^{(r-1)} f(x). \tag{40}$$

Next define

$$w_{r,p}(f;t) = \sup_{|h| \leq t} \|\Delta_h^{(r)} f\|_p \tag{41}$$

where $\|g\|_p = \left\{ \int |g(x)|^p dx \right\}^{1/p}$. Given (p, q, ς), let r be such that $r - 1 \leq \varsigma \leq r$. The *Besov seminorm* is defined by

$$\|f\|_{p,q}^{\varsigma} = \left[\int_0^\infty (h^{-\varsigma} w_{r,p}(f;h))^q \frac{dh}{h} \right]^{1/q}. \tag{42}$$

For $q = \infty$ we define

$$\|f\|_{p,\infty}^{\varsigma} = \sup_{0<h<1} \frac{w_{r,p}(f;h)}{h^{\varsigma}}. \tag{43}$$

The *Besov space* $B_{p,q}^{\varsigma}(c)$ is defined to be the set of functions f mapping $[0,1]$ into $\mathbb{R}$ such that $\int |f|^p < \infty$ and $\|f\|_{p,q}^{\varsigma} \leq c$.

Besov spaces include a wide range of familiar function spaces. The Sobolev space $W_{m,2}$ corresponds to the Besov ball $B_{2,2}^m$. The generalized Sobolev space $W_{m,p}$ which uses an L_p norm on the m^{th} derivative is almost a Besov space in the sense that $B_{p,1}^m \subset W_p(m) \subset B_{p,\infty}^m$. The Hölder space H_α with $\alpha = k + \beta$ is equivalent to $B_{\infty,\infty}^{k+\beta}$, and the set T consisting of functions of bounded variation satisfies $B_{1,1}^1 \subset T \subset B_{1,\infty}^1$.

6 Entropy and Dimension

Given a norm $\|\cdot\|$ on a function space $\mathcal{F}$, a sphere of radius ϵ is a set of the form $\{f \in \mathcal{F} : \|f - g\| \leq \epsilon\}$ for some g. A set of spheres *covers* $\mathcal{F}$ if $\mathcal{F}$ is contained in their union. The *covering number* $N(\epsilon, \|\cdot\|)$ is the smallest number of spheres of radius ϵ required to cover $\mathcal{F}$. We drop the dependence on the norm $\|\cdot\|$ when it is understood from context. The *metric entropy* of $\mathcal{F}$ is $H(\epsilon) = \log N(\epsilon)$. The class $\mathcal{F}$ has dimension d if, for all small ϵ, $N(\epsilon) = c(1/\epsilon)^d$ for some constant c.

A finite set $\{f_1, \ldots, f_k\}$ is an *ϵ-net* if $\|f_i - f_j\| > \epsilon$ for all $i \neq j$. The *packing number* $M(\epsilon)$ is the size of the largest ϵ-net, and the *packing entropy* is $V(\epsilon) = \log M(\epsilon)$. The packing entropy and metric entropy are related by

$$M(2\epsilon) \leq H(\epsilon) \leq M(\epsilon). \tag{44}$$

Here are some common spaces and their entropies:

Space	$H(\epsilon)$
Sobolev $W_{m,p}$	$\epsilon^{-d/m}$
Besov B^{ς}_{pq}	$\epsilon^{-d/\varsigma}$
Hölder H_α	$\epsilon^{-d/\alpha}$

7 Mercer Kernels and Reproducing Kernel Hilbert Spaces

Intuitively, a reproducing kernel Hilbert space (RKHS) is a class of smooth functions defined by an object called a Mercer kernel. Here are the details.

Mercer Kernels. A Mercer kernel is a continuous function $K : [a, b] \times [a, b] \to \mathbb{R}$ such that $K(x, y) = K(y, x)$, and such that K is *positive semidefinite*, meaning that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i, x_j)c_i c_j \geq 0 \tag{45}$$

for all finite sets of points $x_1, \ldots, x_n \in [a, b]$ and all real numbers $c_1, \ldots, c_n$. The function

$$K(x, y) = \sum_{k=1}^{m-1} \frac{1}{k!}x^k y^k + \int_0^{x\wedge y} \frac{(x-u)^{m-1}(y-u)^{m-1}}{(m-1)!^2}du \tag{46}$$

introduced in the Section 4 on Sobolev spaces is an example of a Mercer kernel. The most commonly used kernel is the Gaussian kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{\sigma^2}}.$$

Theorem 4 (Mercer's theorem) *Suppose that $K : \mathcal{X} \times X \to \mathbb{R}$ is symmetric and satisfies $\sup_{x,y} K(x,y) < \infty$, and define*

$$T_K f(x) = \int_{\mathcal{X}} K(x,y)\, f(y)\, dy \tag{47}$$

suppose that $T_k : L^2(\mathcal{X}) \to L^2(X)$ is positive semidefinite; thus,

$$\int_{\mathcal{X}} \int_{\mathcal{X}} K(x,y)\, f(x)\, f(y)\, dx\, dy \geq 0 \tag{48}$$

for any $f \in L^2(\mathcal{X})$. Let λ_i, Ψ_i be the eigenfunctions and eigenvectors of T_K, with

$$\int_{\mathcal{X}} K(x,y)\Psi_i(y)\, dy = \lambda_i \Psi_i(x). \tag{49}$$

Then $\sum_i \lambda_i < \infty$, $\sup_x \Psi_i(x) < \infty$, and

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \Psi_i(x)\, \Psi_j(y), \tag{50}$$

where the convergence is uniform in x, y.

This gives the mapping into feature space as

$$x \mapsto \Phi(x) = \left(\sqrt{\lambda_1}\Psi_1(x), \sqrt{\lambda_2}\Psi_2(x), \ldots\right)^\top \tag{51}$$

The positive semidefinite requirement for Mercer kernels is generally difficult to verify. But the following basic results show how one can build up kernels in pieces.

If $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are Mercer kernels then so are the following:

$$K(x,y) = K_1(x,y) + K_2(x,y) \tag{52}$$

$$K(x,y) = c\, K_1(x,y) + K_2(x,y) \quad \text{for } c \in \mathbb{R}_+ \tag{53}$$

$$K(x,y) = K_1(x,y) + c \quad \text{for } c \in \mathbb{R}_+ \tag{54}$$

$$K(x,y) = K_1(x,y)\, K_2(x,y) \tag{55}$$

$$K(x,y) = f(x)\, f(y) \quad \text{for } f : \mathcal{X} \longrightarrow \mathbb{R} \tag{56}$$

$$K(x,y) = (K_1(x,y) + c)^d \quad \text{for } \theta_1 \in \mathbb{R}_+ \text{ and } d \in \mathbb{N} \tag{57}$$

$$K(x,y) = \exp\left(K_1(x,y)/\sigma^2\right) \quad \text{for } \sigma \in \mathbb{R} \tag{58}$$

$$K(x,y) = \exp\left(-(K_1(x,x) - 2K_1(x,y) + K_1(y,y))/2\sigma^2\right) \tag{59}$$

$$K(x,y) = K_1(x,y)/\sqrt{K_1(x,x)\, K_1(y,y)} \tag{60}$$

RKHS. Given a kernel K, let $K_x(\cdot)$ be the function obtained by fixing the first coordinate. That is, $K_x(y) = K(x, y)$. For the Gaussian kernel, K_x is a Normal, centered at x. We can create functions by taking liner combinations of the kernel:

$$f(x) = \sum_{j=1}^{k} \alpha_j K_{x_j}(x).$$

Let $\mathcal{H}_0$ denote all such functions:

$$\mathcal{H}_0 = \left\{ f : \sum_{j=1}^{k} \alpha_j K_{x_j}(x) \right\}.$$

Given two such functions $f(x) = \sum_{j=1}^{k} \alpha_j K_{x_j}(x)$ and $g(x) = \sum_{j=1}^{m} \beta_j K_{y_j}(x)$ we define an inner product

$$\langle f, g \rangle = \langle f, g \rangle_K = \sum_i \sum_j \alpha_i \beta_j K(x_i, y_j).$$

In general, f (and g) might be representable in more than one way. You can check that $\langle f, g \rangle_K$ is independent of how f (or g) is represented. The inner product defines a norm:

$$||f||_K = \sqrt{\langle f, f, \rangle} = \sqrt{\sum_j \sum_k \alpha_j \alpha_k K(x_j, x_k)} = \sqrt{\alpha^T \mathbb{K} \alpha}$$

where $\alpha = (\alpha_1, \ldots, \alpha_k)^T$ and $\mathbb{K}$ is the $k \times k$ matrix with $\mathbb{K}_{jk} = K(x_j, x_k)$.

The Reproducing Property. Let $f(x) = \sum_i \alpha_i K_{x_i}(x)$. Note the following crucial property:

$$\langle f, K_x \rangle = \sum_i \alpha_i K(x_i, x) = f(x).$$

This follows from the definition of $\langle f, g \rangle$ where we take $g = K_x$. This implies that

$$\langle K_x, K_x \rangle = K(x, x).$$

This is called the reproducing property. It also implies that K_x is the **representer** of the evaluation functional.

The completion of $\mathcal{H}_0$ with respect to $|| \cdot ||_K$ is denoted by $\mathcal{H}_K$ and is called the RKHS generated by K.

To verify that this is a well-defined Hilbert space, you should check that the following properties hold:

$$\begin{aligned} \langle f, g \rangle &= \langle g, f \rangle \\ \langle cf + dg, h \rangle &= c\langle f, h \rangle + c\langle g, h \rangle \\ \langle f, f \rangle = 0 \quad &\text{iff} \quad f = 0. \end{aligned}$$

The last one is not obvious so let us verify it here. It is easy to see that $f = 0$ implies that $\langle f, f \rangle = 0$. Now we must show that $\langle f, f \rangle = 0$ implies that $f(x) = 0$. So suppose that $\langle f, f \rangle = 0$. Pick any x. Then

$$\begin{aligned} 0 &\leq f^2(x) = \langle f, K_x \rangle^2 = \langle f, K_x \rangle \, \langle f, K_x \rangle \\ &\leq ||f||^2 \, ||K_x||^2 = \langle f, f \rangle^2 \, ||K_x||^2 = 0 \end{aligned}$$

where we used Cauchy-Schwartz. So $0 \leq f^2(x) \leq 0$ which means that $f(x) = 0$.

Evaluation Functionals. A key property of RKHS's is the behavior of the evaluation functional. The *evaluation functional* δ_x assigns a real number to each function. It is defined by $\delta_x f = f(x)$. In general, the evaluation functional is not continuous. This means we can have $f_n \to f$ but $\delta_x f_n$ does not converge to $\delta_x f$. For example, let $f(x) = 0$ and $f_n(x) = \sqrt{n} I(x < 1/n^2)$. Then $||f_n - f|| = 1/\sqrt{n} \to 0$. But $\delta_0 f_n = \sqrt{n}$ which does not converge to $\delta_0 f = 0$. Intuitively, this is because Hilbert spaces can contain very unsmooth functions.

But in an RKHS, the evaluation functional is continuous. Intuitively, this means that the functions in the space are well-behaved. To see this, suppose that $f_n \to f$. Then

$$\delta_x f_n = \langle f_n K_x \rangle \to \langle f K_x \rangle = f(x) = \delta_x f$$

so the evaluation functional is continuous. In fact:

A Hilbert space is a RKHS if and only if the evaluation functionals are continuous.

Examples. Here are some examples of RKHS's.

Example 5 *Let $\mathcal{H}$ be all functions f on $\mathbb{R}$ such that the support of the Fourier transform of f is contained in $[-a, a]$. Then*

$$K(x, y) = \frac{\sin(a(y - x))}{a(y - x)}$$

and

$$\langle f, g \rangle = \int fg.$$

Example 6 *Let $\mathcal{H}$ be all functions f on $(0, 1)$ such that*

$$\int_0^1 (f^2(x) + (f'(x))^2) x^2 dx < \infty.$$

Then

$$K(x,y) = (xy)^{-1}\left(e^{-x}\sinh(y)I(0<x\leq y) + e^{-y}\sinh(x)I(0<y\leq x)\right)$$

and

$$||f||^2 = \int_0^1 (f^2(x) + (f'(x))^2)x^2dx.$$

Example 7 *The Sobolev space of order m is (roughly speaking) the set of functions f such that $\int(f^{(m)})^2 < \infty$. For $m = 1$ and $\mathcal{X} = [0,1]$ the kernel is*

$$K(x,y) = \begin{cases} 1 + xy + \frac{xy^2}{2} - \frac{y^3}{6} & 0 \leq y \leq x \leq 1 \\ 1 + xy + \frac{yx^2}{2} - \frac{x^3}{6} & 0 \leq x \leq y \leq 1 \end{cases}$$

and

$$||f||_K^2 = f^2(0) + f'(0)^2 + \int_0^1 (f''(x))^2dx.$$

Spectral Representation. Suppose that $\sup_{x,y} K(x,y) < \infty$. Define eigenvalues λ_j and orthonormal eigenfunctions ψ_j by

$$\int K(x,y)\psi_j(y)dy = \lambda_j\psi_j(x).$$

Then $\sum_j \lambda_j < \infty$ and $\sup_x |\psi_j(x)| < \infty$. Also,

$$K(x,y) = \sum_{j=1}^{\infty} \lambda_j\psi_j(x)\psi_j(y).$$

Define the **feature map** Φ by

$$\Phi(x) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots).$$

We can expand f either in terms of K or in terms of the basis $\psi_1, \psi_2, \ldots$:

$$f(x) = \sum_i \alpha_i K(x_i, x) = \sum_{j=1}^{\infty} \beta_j\psi_j(x).$$

Furthermore, if $f(x) = \sum_j a_j\psi_j(x)$ and $g(x) = \sum_j b_j\psi_j(x)$, then

$$\langle f, g\rangle = \sum_{j=1}^{\infty} \frac{a_jb_j}{\lambda_j}.$$

Roughly speaking, when $||f||_K$ is small, then f is smooth.

Representer Theorem. Let ℓ be a loss function depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$ and on $f(X_1), \ldots, f(X_n)$. Let $\widehat{f}$ minimize

$$\ell + g(||f||_K^2)$$

where g is any monotone increasing function. Then $\widehat{f}$ has the form

$$\widehat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

for some $\alpha_1, \ldots, \alpha_n$.

RKHS Regression. Define $\widehat{m}$ to minimize

$$R = \sum_i (Y_i - m(X_i))^2 + \lambda ||m||_K^2.$$

By the representer theorem, $\widehat{m}(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$. Plug this into R and we get

$$R = ||Y - \mathbb{K}\alpha||^2 + \lambda \alpha^T \mathbb{K} \alpha$$

where $\mathbb{K}_{jk} = K(X_j, X_k)$ is the Gram matrix. The minimizer over α is

$$\widehat{\alpha} = (\mathbb{K} + \lambda I)^{-1} Y$$

and $\widehat{m}(x) = \sum_j \widehat{\alpha}_j K(X_i, x)$. The fitted values are

$$\widehat{Y} = \mathbb{K}\widehat{\alpha} = \mathbb{K}(\mathbb{K} + \lambda I)^{-1} Y = LY.$$

So this is a linear smoother. We will discuss this in detail later.

Support Vector Machines. Suppose $Y_i \in \{-1, +1\}$. Recall the the linear SVM minimizes the penalized hinge loss:

$$J = \sum_i [1 - Y_i(\beta_0 + \beta^T X_i)]_+ + \frac{\lambda}{2} ||\beta||_2^2.$$

The dual is to maximize

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle$$

subject to $0 \leq \alpha_i \leq C$.

The RKHS version is to minimize

$$J = \sum_i [1 - Y_i f(X_i)]_+ + \frac{\lambda}{2} ||f||_K^2.$$

The dual is the same except that $\langle X_i, X_j \rangle$ is replaced with $K(X_i, X_j)$. This is called the kernel trick.

The Kernel Trick. This is a fairly general trick. In many algorithms you can replace $\langle x_i, x_j \rangle$ with $K(x_i, x_j)$ and get a nonlinear version of the algorithm. This is equivalent to replacing x with $\Phi(x)$ and replacing $\langle x_i, x_j \rangle$ with $\langle \Phi(x_i), \Phi(x_j) \rangle$. However, $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ and $K(x_i, x_j)$ is much easier to compute.

In summary, by replacing $\langle x_i, x_j \rangle$ with $K(x_i, x_j)$ we turn a linear procedure into a nonlinear procedure without adding much computation.

Hidden Tuning Parameters. There are hidden tuning parameters in the RKHS. Consider the Gaussian kernel

$$K(x, y) = e^{-\frac{||x-y||^2}{\sigma^2}}.$$

For nonparametric regression we minimize $\sum_i (Y_i - m(X_i))^2$ subject to $||m||_K \leq L$. We control the bias variance tradeoff by doing cross-validation over L. But what about σ?

This parameter seems to get mostly ignored. Suppose we have a uniform distribution on a circle. The eigenfunctions of $K(x, y)$ are the sines and cosines. The eigenvalues λ_k die off like $(1/\sigma)^{2k}$. So σ affects the bias-variance tradeoff since it weights things towards lower order Fourier functions. In principle we can compensate for this by varying L. But clearly there is some interaction between L and σ. The practical effect is not well understood.

Now consider the polynomial kernel $K(x, y) = (1 + \langle x, y \rangle)^d$. This kernel has the same eigenfunctions but the eigenvalues decay at a polynomial rate depending on d. So there is an interaction between L, d and, the choice of kernel itself.