From Generalized Ridge Regression to Reproducing Kernel Hilbert Spaces

1 A generalized ridge-regression calculation

Consider the linear model

$$Y = X\beta + \varepsilon,$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $\beta \in \mathbb{R}^p$. Ordinary ridge regression replaces least squares by the criterion

$$\|Y - X\beta\|^2 + \lambda \|\beta\|^2 .$$

More generally, let $K \in \mathbb{R}^{p\times p}$ be symmetric and positive semidefinite, and consider

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|^2 + \lambda\beta^T K\beta \right\}, \qquad \lambda > 0. \tag{1}$$

Because K is positive semidefinite, it has a symmetric positive semidefinite square root $K^{1/2}$. Hence

$$\|Y - X\beta\|^2 + \lambda\beta^T K\beta = \left\| \begin{pmatrix} Y \\ 0 \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda}\, K^{1/2} \end{pmatrix} \beta \right\|^2 .$$

Thus (1) is an ordinary least-squares problem for the augmented response and design matrix

$$\widetilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}, \qquad \widetilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda}\, K^{1/2} \end{pmatrix}.$$

Assuming $X^T X + \lambda K$ is invertible, the usual least-squares formula yields

$$\widehat{\beta} = (X^T X + \lambda K)^{-1} X^T Y. \tag{2}$$

Taking $K = I_p$ recovers ordinary ridge regression. Later, a function-estimation problem will reduce to exactly this form, with both the design matrix and the penalty matrix determined by a kernel.

Remark 1.1 (The noninvertible case). If $X^T X + \lambda K$ is not invertible, the minimization problem still has solutions, but the minimizing coefficient vector need not be unique. Every minimizer satisfies

$$(X^T X + \lambda K)\beta = X^T Y.$$

Any two minimizers differ by a vector in $\ker X \cap \ker K$ and consequently give the same fitted values $X\beta$. We will not need a more explicit description.

2 Hilbert spaces

2.1 Inner products and the induced geometry

Throughout these notes all vector spaces are real.

Definition 2.1. An *inner product* on a real vector space H is a map $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$ such that, for all $f, g, h \in H$ and $a, b \in \mathbb{R}$,

(i) $\langle af + bg, h \rangle = a \langle f, h \rangle + b \langle g, h \rangle$;

(ii) $\langle f, g \rangle = \langle g, f \rangle$;

(iii) $\langle f, f \rangle \geq 0$, with equality if and only if $f = 0$.

The induced norm is $\|f\| = \sqrt{\langle f, f \rangle}$. A sequence (f_n) is Cauchy if $\|f_n - f_m\| \to 0$ as $m, n \to \infty$. An inner-product space is a *Hilbert space* if every Cauchy sequence converges to an element of the space.

The basic inequality governing inner-product geometry is Cauchy–Schwarz:

$$|\langle f, g \rangle| \leq \|f\| \, \|g\| . \tag{3}$$

Indeed, if $g \neq 0$, then $0 \leq \|f - tg\|^2$ for every $t \in \mathbb{R}$; choosing $t = \langle f, g \rangle / \|g\|^2$ gives (3). The triangle inequality follows immediately:

$$\|f + g\|^2 = \|f\|^2 + 2 \langle f, g \rangle + \|g\|^2 \leq (\|f\| + \|g\|)^2.$$

If $\langle f, g \rangle = 0$, we write $f \perp g$. In that case the Pythagorean identity is

$$\|f + g\|^2 = \|f\|^2 + \|g\|^2 .$$

2.2 Examples

Example 2.2 (Euclidean space)**.** The space $\mathbb{R}^d$, with $\langle x, y \rangle = \sum_{j=1}^d x_j y_j$, is a Hilbert space. Every finite-dimensional inner-product space is complete.

Example 2.3 (The sequence space ℓ^2)**.** Let

$$\ell^2 = \left\{ x = (x_1, x_2, \dots) : \sum_{j=1}^{\infty} x_j^2 < \infty \right\}, \qquad \langle x, y \rangle = \sum_{j=1}^{\infty} x_j y_j.$$

Cauchy–Schwarz guarantees that the inner-product series converges absolutely. The resulting norm is

$$\|x\|_2 = \left(\sum_{j=1}^{\infty} x_j^2 \right)^{1/2},$$

and ℓ^2 is complete.

Example 2.4 (The function space $L^2(0,1)$)**.** The elements of $L^2([0,1])$ are equivalence classes of measurable functions for which

$$\int_0^1 |f(t)|^2 \, dt < \infty, \qquad \langle f, g \rangle_{L^2} = \int_0^1 f(t) g(t) \, dt.$$

Functions that agree almost everywhere represent the same element. This space is complete. Notice that point evaluation is not well defined on $L^2([0,1])$: changing a function at one point does not change its equivalence class.

Example 2.5 (A Hilbert space defined through derivatives)**.** Let $\mathcal{H}_1$ be the set of functions $f : [0,1] \to \mathbb{R}$ for which there exists an $h \in L^2([0,1])$ satisfying

$$f(x) = \int_0^x h(t)\,dt \qquad \text{for every } x \in [0,1]. \tag{4}$$

The function h is unique up to equality almost everywhere: if the indefinite integral of $h_1 - h_2$ vanishes for every x, then $h_1 = h_2$ almost everywhere. We denote this uniquely determined L^2 element by f'. Equivalently, $\mathcal{H}_1$ consists of the absolutely continuous functions satisfying $f(0) = 0$ and $f' \in L^2([0,1])$.

Define

$$\langle f, g \rangle_{\mathcal{H}_1} = \int_0^1 f'(t) g'(t)\,dt.$$

This is an inner product: if $\|f\|_{\mathcal{H}_1} = 0$, then $f' = 0$ almost everywhere, and (4) gives $f = 0$. Moreover, the map

$$T : L^2([0,1]) \to \mathcal{H}_1, \qquad (Th)(x) = \int_0^x h(t)\,dt,$$

is a linear bijection satisfying $\|Th\|_{\mathcal{H}_1} = \|h\|_{L^2}$. It is therefore an isometric isomorphism, so completeness of $L^2([0,1])$ implies that $\mathcal{H}_1$ is a Hilbert space.

Although $\mathcal{H}_1$ is isometric to $L^2([0,1])$, its elements are genuine continuous functions rather than almost-everywhere equivalence classes. This distinction will become important when we discuss point evaluation.

2.3 Orthogonal projection

For a subset $M \subseteq H$, define

$$M^\perp = \{h \in H : \langle h, m \rangle = 0 \text{ for every } m \in M\}.$$

Theorem 2.6 (Projection theorem)**.** *Let M be a closed linear subspace of a Hilbert space H. Every $f \in H$ has a unique decomposition*

$$f = f_M + f_\perp, \qquad f_M \in M, \quad f_\perp \in M^\perp.$$

Equivalently, for every $f \in H$ there is a unique $f_M \in M$ minimizing $\|f - m\|$ over $m \in M$.

Proof. Let $d = \inf_{m \in M} \|f - m\|$ and choose $m_n \in M$ with $\|f - m_n\| \to d$. Apply the parallelogram identity

$$\|u - v\|^2 + \|u + v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

with $u = f - m_n$ and $v = f - m_k$. Then $u - v = m_k - m_n$, while $u + v = 2(f - (m_n + m_k)/2)$. Therefore

$$\|m_n - m_k\|^2 = 2\|f - m_n\|^2 + 2\|f - m_k\|^2 - 4\left\|f - \frac{m_n + m_k}{2}\right\|^2.$$

Since $(m_n + m_k)/2 \in M$, the definition of d gives $\|f - (m_n + m_k)/2\| \geq d$, and hence

$$\|m_n - m_k\|^2 \leq 2\|f - m_n\|^2 + 2\|f - m_k\|^2 - 4d^2 \longrightarrow 0.$$

Thus (m_n) is Cauchy. Since M is closed, $m_n \to f_M \in M$, and $\|f - f_M\| = d$. For $m \in M$ and $t \in \mathbb{R}$, minimality gives

$$\|f - f_M\|^2 \leq \|f - f_M - tm\|^2 = \|f - f_M\|^2 - 2t\langle f - f_M, m \rangle + t^2\|m\|^2.$$

Since this holds for every t, we must have $\langle f - f_M, m\rangle = 0$. Hence $f_\perp = f - f_M \in M^\perp$.

To prove uniqueness, first observe that $M \cap M^\perp = \{0\}$. Indeed, if $h \in M \cap M^\perp$, then h is orthogonal to every element of M, including itself, and therefore

$$\|h\|^2 = \langle h, h\rangle = 0,$$

so $h = 0$. Now suppose that

$$f = m_1 + r_1 = m_2 + r_2, \qquad m_1, m_2 \in M, \quad r_1, r_2 \in M^\perp.$$

Then

$$m_1 - m_2 = r_2 - r_1.$$

The left-hand side belongs to M, while the right-hand side belongs to $M^\perp$. Hence $m_1 - m_2 \in M \cap M^\perp = \{0\}$, so $m_1 = m_2$ and then $r_1 = r_2$. Thus the decomposition is unique. In particular, if $\widetilde{f}_M \in M$ were another minimizer, the preceding argument would show that $f - \widetilde{f}_M \in M^\perp$, giving a second such decomposition; hence $\widetilde{f}_M = f_M$. □

Finite-dimensional subspaces are closed, so the theorem applies in particular to spans of finitely many vectors.

3 Linear functionals

Definition 3.1. A map $L : H \to \mathbb{R}$ is a *linear functional* if

$$L(af + bg) = aL(f) + bL(g) \qquad (f, g \in H,\ a, b \in \mathbb{R}).$$

It is *bounded* if there is a constant $C < \infty$ such that

$$|L(f)| \le C\,\|f\| \qquad \text{for every } f \in H.$$

The smallest such constant is the operator norm

$$\|L\| = \sup_{\|f\| \le 1} |L(f)|.$$

Proposition 3.2. *A linear functional on a normed vector space is bounded if and only if it is continuous.*

Proof. If L is bounded, then

$$|L(f) - L(g)| = |L(f - g)| \le \|L\|\ \|f - g\|,$$

so L is Lipschitz continuous. Conversely, suppose L is continuous at 0. Choose $\delta > 0$ such that $\|h\| < \delta$ implies $|L(h)| < 1$. For $f \ne 0$, apply this to $h = \delta f/(2\,\|f\|)$ to obtain $|L(f)| < 2\,\|f\|/\delta$. Thus L is bounded. Continuity at any one point is equivalent to continuity at 0 by linearity. □

The next example shows that the qualifier "bounded" is essential. An everywhere-defined linear functional on an infinite-dimensional Hilbert space need not be continuous.

3.1 An everywhere-defined unbounded linear functional on ℓ^2

Let e_j be the jth standard unit vector and let

$$c_{00} = \{x \in \ell^2 : x_j = 0 \text{ for all but finitely many } j\}.$$

The sequence $(e_j)_{j\geq 1}$ is a *Schauder basis* of ℓ^2: every $x \in \ell^2$ has the norm-convergent expansion $x = \sum_{j=1}^{\infty} x_j e_j$. It is not a Hamel basis, because a Hamel basis permits only finite linear combinations, and the finite span of the e_j is merely c_{00}.

Set

$$z = \left(1, \frac{1}{2}, \frac{1}{3}, \ldots\right) \in \ell^2 \qquad \text{and} \qquad M = c_{00} + \operatorname{span}\{z\}.$$

Every element of M has a unique representation $y + tz$, with $y \in c_{00}$ and $t \in \mathbb{R}$. Indeed, $(t - t')z = y' - y \in c_{00}$ forces $t = t'$ because $z \notin c_{00}$. Thus

$$B_0 = \{e_1, e_2, \ldots\} \cup \{z\}$$

is a Hamel basis of M. The subspace M is proper: for example, $w = (j^{-3/4})_{j\geq 1}$ belongs to ℓ^2, but w cannot agree eventually with tz for any fixed t.

Define

$$L_0 : M \to \mathbb{R}, \qquad L_0(y + tz) = t.$$

Then $L_0(e_j) = 0$ for every j, while $L_0(z) = 1$. To see that L_0 is unbounded, let

$$z^{(n)} = \sum_{j=1}^{n} \frac{1}{j} e_j.$$

Since $\left\|z - z^{(n)}\right\|_2^2 = \sum_{j>n} j^{-2} \to 0$ but $L_0(z - z^{(n)}) = 1$, the functional is not continuous at the origin. More explicitly,

$$u_n = \frac{z - z^{(n)}}{\|z - z^{(n)}\|_2} \qquad \Longrightarrow \qquad \|u_n\|_2 = 1, \quad |L_0(u_n)| = \frac{1}{\|z - z^{(n)}\|_2} \to \infty.$$

Thus L_0 is unbounded on the dense subspace M.

It remains to extend L_0 to all of ℓ^2. Consider the collection of all linearly independent subsets of ℓ^2 that contain B_0, ordered by inclusion. The union of any chain is linearly independent, because a linear dependence involves only finitely many vectors and those vectors all lie in one member of the chain. Zorn's lemma therefore supplies a maximal linearly independent set B containing B_0. Maximality implies that B spans ℓ^2, so B is a Hamel basis.

Let $C = B \setminus B_0$ and $N = \operatorname{span}(C)$. Then

$$\ell^2 = M \oplus N$$

as an algebraic direct sum. Define

$$L : \ell^2 \to \mathbb{R}, \qquad L(m + n) = L_0(m), \quad m \in M,\ n \in N.$$

Equivalently, define $L(z) = 1$, $L(e_j) = 0$ for all j, and $L(c) = 0$ for $c \in C$, and extend by finite linearity. The restriction of L to M is L_0, so the sequence (u_n) still shows that L is unbounded.

Theorem 3.3. *There exists an everywhere-defined unbounded linear functional $L : \ell^2 \to \mathbb{R}$. It may be chosen so that*

$$L(e_j) = 0 \quad \textit{for every } j, \qquad L\left(1, \frac{1}{2}, \frac{1}{3}, \dots\right) = 1.$$

This example identifies the precise failure in a tempting coordinate argument. If $x^{(n)} = \sum_{j=1}^{n} x_j e_j$, then $x^{(n)} \to x$ and finite linearity gives

$$L(x^{(n)}) = \sum_{j=1}^{n} x_j L(e_j).$$

Passing to the limit and concluding that $L(x) = \sum_{j=1}^{\infty} x_j L(e_j)$ requires continuity. In the construction above, $L(z^{(n)}) = 0$ for every n, although $z^{(n)} \to z$ and $L(z) = 1$.

The following proposition confirms that an everywhere-convergent coordinate formula would necessarily define a bounded functional.

Proposition 3.4. *Let $(a_j)_{j\geq 1}$ be a real sequence. If $\sum_{j=1}^{\infty} a_j x_j$ converges for every $x \in \ell^2$, then $a = (a_j) \in \ell^2$. Consequently,*

$$T(x) = \sum_{j=1}^{\infty} a_j x_j$$

defines a bounded linear functional and $|T(x)| \leq \|a\|_2 \, \|x\|_2$.

Proof. Suppose instead that $\sum_j a_j^2 = \infty$. Choose $0 = n_0 < n_1 < n_2 < \cdots$ so that

$$S_k = \sum_{j=n_{k-1}+1}^{n_k} a_j^2 \geq 2^k,$$

and set $x_j = a_j / S_k$ when $n_{k-1} < j \leq n_k$. Then

$$\sum_{j=1}^{\infty} x_j^2 = \sum_{k=1}^{\infty} \frac{1}{S_k} \leq \sum_{k=1}^{\infty} 2^{-k} < \infty,$$

so $x \in \ell^2$. At the end of the mth block, however,

$$\sum_{j=1}^{n_m} a_j x_j = \sum_{k=1}^{m} \frac{1}{S_k} \sum_{j=n_{k-1}+1}^{n_k} a_j^2 = m,$$

contradicting convergence. Thus $a \in \ell^2$, and the stated bound follows from Cauchy–Schwarz. □

One might try to extend L_0 by taking the orthogonal complement of M. This does not work. Since $c_{00} \subseteq M$, every $v \in M^{\perp}$ satisfies $v_j = \langle v, e_j \rangle = 0$ for every j, and hence $M^{\perp} = \{0\}$. The point is that M is dense but not closed. The Hilbert-space decomposition is

$$\ell^2 = \overline{M} \oplus M^{\perp},$$

not $\ell^2 = M \oplus M^{\perp}$ for an arbitrary subspace. The complement N above is algebraic, not orthogonal.

Finally, the axiom of choice enters precisely in the use of Zorn's lemma to extend B_0 to a Hamel basis of all of ℓ^2. The explicit construction of z, M, and L_0 requires no choice principle.

4 The Riesz representation theorem

The unbounded example shows that no inner-product representation can hold for arbitrary linear functionals. It does hold for bounded ones.

Theorem 4.1 (Riesz representation theorem)**.** *Let H be a real Hilbert space and let $L : H \to \mathbb{R}$ be a bounded linear functional. There exists a unique $g \in H$ such that*

$$L(f) = \langle f, g \rangle \qquad \text{for every } f \in H. \tag{5}$$

Moreover, $\|L\| = \|g\|$.

Proof. If $L = 0$, take $g = 0$. Suppose $L \neq 0$ and let $M = \ker L$. Since L is continuous, M is closed. Choose $x_0 \in H$ with $L(x_0) = 1$, and use Theorem 2.6 to write $x_0 = m_0 + u$, where $m_0 \in M$ and $u \in M^{\perp}$. Since $L(u) = 1$, we have $u \neq 0$. For arbitrary $f \in H$, the vector $f - L(f)x_0$ belongs to M, so

$$0 = \langle f - L(f)x_0, u \rangle = \langle f, u \rangle - L(f) \langle x_0, u \rangle .$$

Because $x_0 = m_0 + u$ and $u \perp m_0$, $\langle x_0, u \rangle = \|u\|^2$. Hence

$$L(f) = \left\langle f, \frac{u}{\|u\|^2} \right\rangle .$$

This proves existence with $g = u/\|u\|^2$. If both g and h represent L, then $\langle f, g - h \rangle = 0$ for every f; taking $f = g - h$ proves uniqueness. Finally, Cauchy–Schwarz gives $\|L\| \leq \|g\|$, while $f = g/\|g\|$ gives equality when $g \neq 0$. ☐

5 Reproducing Kernel Hilbert Spaces

5.1 Definition and existence of the reproducing kernel

Definition 5.1. Let $\mathcal{X}$ be a set. A Hilbert space H of real-valued functions on $\mathcal{X}$ is a *reproducing kernel Hilbert space* if, for every $x \in \mathcal{X}$, the point-evaluation functional

$$E_x : H \to \mathbb{R}, \qquad E_x(f) = f(x),$$

is continuous.

By Proposition 3.2, continuity of E_x is equivalent to boundedness. The Riesz theorem then produces a representer for each evaluation functional.

Theorem 5.2. *Let H be a reproducing kernel Hilbert space on $\mathcal{X}$. For every $x \in \mathcal{X}$ there is a unique $R_x \in H$ such that*

$$f(x) = \langle f, R_x \rangle_H \qquad \text{for every } f \in H.$$

The function

$$\mathcal{K}(x, y) = \langle R_y, R_x \rangle_H$$

is symmetric and positive semidefinite, $R_x = \mathcal{K}(\cdot, x)$, and

$$f(x) = \langle f, \mathcal{K}(\cdot, x) \rangle_H . \tag{6}$$

The kernel $\mathcal{K}$ is uniquely determined by H.

Proof. Riesz representation gives R_x. Evaluating R_y at x yields

$$R_y(x) = \langle R_y, R_x \rangle_H = \mathcal{K}(x, y),$$

so $R_y = \mathcal{K}(\cdot, y)$ and (6) follows. Symmetry is inherited from the inner product. For $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} a_i a_j \mathcal{K}(x_i, x_j) = \left\| \sum_{i=1}^{n} a_i R_{x_i} \right\|_H^2 \geq 0.$$

Thus $\mathcal{K}$ is positive semidefinite. Uniqueness follows from the uniqueness of each Riesz representer. □

In particular,

$$|f(x)| \leq \|f\|_H \sqrt{\mathcal{K}(x, x)}, \tag{7}$$

because $\|\mathcal{K}(\cdot, x)\|_H^2 = \mathcal{K}(x, x)$.

5.2 The first-order derivative space

Return to the Hilbert space $\mathcal{H}_1$ of Example 2.5. For $f \in \mathcal{H}_1$ and $x \in [0, 1]$,

$$|f(x)| = \left| \int_0^x f'(t)\, dt \right| \leq \sqrt{x}\, \|f\|_{\mathcal{H}_1}.$$

Thus point evaluation is bounded, and $\mathcal{H}_1$ is a reproducing kernel Hilbert space. For fixed x, define

$$R_x(s) = \min\{s, x\}.$$

Then $R_x'(t) = \mathbf{1}_{[0,x]}(t)$ almost everywhere, and therefore

$$\langle f, R_x \rangle_{\mathcal{H}_1} = \int_0^1 f'(t) \mathbf{1}_{[0,x]}(t)\, dt = \int_0^x f'(t)\, dt = f(x).$$

Hence the reproducing kernel is

$$\mathcal{K}_1(x, y) = \min\{x, y\}. \tag{8}$$

5.3 Higher-order derivative spaces

For an integer $\alpha \geq 1$, let $\mathcal{H}_\alpha$ consist of the functions $f \in C^{\alpha-1}([0, 1])$ for which $f^{(\alpha-1)}$ is absolutely continuous, $f^{(\ell)}(0) = 0$ for $\ell = 0, \ldots, \alpha - 1$, and the almost-everywhere derivative $f^{(\alpha)}$ belongs to $L^2([0, 1])$. Define

$$\langle f, g \rangle_{\mathcal{H}_\alpha} = \int_0^1 f^{(\alpha)}(z) g^{(\alpha)}(z)\, dz. \tag{9}$$

Since $L^2([0, 1]) \subset L^1([0, 1])$, the integral form of Taylor's theorem applies:

$$f(x) = \sum_{\ell=0}^{\alpha-1} f^{(\ell)}(0) \frac{x^\ell}{\ell!} + \int_0^1 f^{(\alpha)}(z) \frac{(x - z)_+^{\alpha-1}}{(\alpha - 1)!}\, dz, \qquad (t)_+ = \max\{0, t\}. \tag{10}$$

The boundary conditions remove the finite sum, so every $f \in \mathcal{H}_\alpha$ satisfies

$$f(x) = \int_0^1 f^{(\alpha)}(z) \frac{(x - z)_+^{\alpha-1}}{(\alpha - 1)!}\, dz.$$

The map $f \mapsto f^{(\alpha)}$ is an isometric bijection from $\mathcal{H}_\alpha$ onto $L^2([0,1])$, with inverse given by this integral formula. Thus $\mathcal{H}_\alpha$ is a Hilbert space.

Its reproducing kernel is

$$\mathcal{K}_\alpha(x,y) = \int_0^1 \frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!} \frac{(y-z)_+^{\alpha-1}}{(\alpha-1)!}\, dz. \tag{11}$$

Indeed, for $R_x(\cdot) = \mathcal{K}_\alpha(\cdot, x)$,

$$R_x^{(\alpha)}(z) = \frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!} \quad \text{almost everywhere,}$$

and (10) gives

$$\langle R_x, f\rangle_{\mathcal{H}_\alpha} = \int_0^1 f^{(\alpha)}(z) \frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!}\, dz = f(x).$$

When $\alpha = 1$, formula (11) reduces to $\mathcal{K}_1(x,y) = \min\{x,y\}$.

5.4 Constructing a Hilbert space from a positive-semidefinite kernel

We have shown that every reproducing kernel Hilbert space determines a positive-semidefinite kernel. The converse is also true.

Theorem 5.3 (Kernel construction)**.** *Let $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric and satisfy*

$$\sum_{i,j=1}^{n} a_i a_j \mathcal{K}(x_i, x_j) \geq 0$$

for every n, every $x_1, \ldots, x_n \in \mathcal{X}$, and every $a_1, \ldots, a_n \in \mathbb{R}$. Then there is a reproducing kernel Hilbert space on $\mathcal{X}$ whose reproducing kernel is $\mathcal{K}$.

Proof. Begin with the vector space of finite kernel expansions

$$H_0 = \left\{ \sum_{i=1}^{m} a_i \mathcal{K}(x_i, \cdot) : m \geq 1,\ a_i \in \mathbb{R},\ x_i \in \mathcal{X} \right\}.$$

For

$$f = \sum_{i=1}^{m} a_i \mathcal{K}(x_i, \cdot), \qquad g = \sum_{j=1}^{r} b_j \mathcal{K}(y_j, \cdot),$$

define

$$\langle f, g\rangle_{H_0} = \sum_{i=1}^{m} \sum_{j=1}^{r} a_i b_j \mathcal{K}(x_i, y_j). \tag{12}$$

This definition does not depend on the chosen expansions. Indeed, if f is the zero function, then the right side of (12) equals $\sum_j b_j f(y_j) = 0$ for every g. Positive semidefiniteness gives $\|f\|_{H_0}^2 \geq 0$, and the usual quadratic argument shows that $\|f\|_{H_0} = 0$ implies $f(x) = 0$ for every x. Thus (12) is an inner product on H_0.

For $f \in H_0$,

$$\langle f, \mathcal{K}(x, \cdot)\rangle_{H_0} = f(x), \qquad |f(x)| \leq \|f\|_{H_0} \sqrt{\mathcal{K}(x,x)}.$$

Complete H_0 in its induced norm. The displayed bound allows point evaluation to extend continuously to the completion, and the extension still satisfies the reproducing identity. The resulting Hilbert space therefore has reproducing kernel $\mathcal{K}$. □

Thus one may start with a positive-semidefinite kernel, take finite linear combinations of the sections $\mathcal{K}(x, \cdot)$, and then take their closure in the norm induced by the kernel.

5.5 The multivariate Gaussian kernel

For $x, y \in \mathbb{R}^d$ and $\sigma > 0$, define

$$\mathcal{K}_\sigma(x,y) = \exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\}. \tag{13}$$

We now verify that this kernel is positive semidefinite. Let

$$Z \sim N(0, \sigma^{-2} I_d), \qquad u = x - y, \qquad V = Z^T u.$$

Then V is Gaussian with mean zero and

$$\operatorname{Var}(V) = \sum_{r=1}^d u_r^2 \operatorname{Var}(Z_r) = \frac{1}{\sigma^2}\sum_{r=1}^d u_r^2 = \frac{\|x-y\|^2}{\sigma^2}.$$

To compute its expected cosine, first let $G \sim N(0,1)$ and $\varphi(t) = \mathbb{E}e^{itG}$. If $p(z) = (2\pi)^{-1/2}e^{-z^2/2}$ is the standard Gaussian density, then $p'(z) = -zp(z)$, so integration by parts gives

$$\varphi'(t) = i\int_{\mathbb{R}} z e^{itz} p(z)\, dz = -t\varphi(t), \qquad \varphi(0) = 1.$$

Therefore $\varphi(t) = e^{-t^2/2}$. Since

$$V \overset{d}{=} \frac{\|x-y\|}{\sigma} G,$$

we obtain, on one line,

$$\mathbb{E}\cos(Z^T(x-y)) = \operatorname{Re}\varphi\left(\frac{\|x-y\|}{\sigma}\right) = \exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\} = \mathcal{K}_\sigma(x,y). \tag{14}$$

Now let $x_1, \ldots, x_n \in \mathbb{R}^d$ and $a_1, \ldots, a_n \in \mathbb{R}$. Using $\cos(s-t) = \cos s \cos t + \sin s \sin t$,

$$\begin{aligned}
\sum_{i,j=1}^n a_i a_j \mathcal{K}_\sigma(x_i, x_j) &= \mathbb{E}\sum_{i,j=1}^n a_i a_j \cos(Z^T x_i - Z^T x_j) \\
&= \mathbb{E}\left[\left(\sum_{i=1}^n a_i \cos(Z^T x_i)\right)^2 + \left(\sum_{i=1}^n a_i \sin(Z^T x_i)\right)^2\right] \geq 0.
\end{aligned}$$

Thus the Gaussian kernel is positive semidefinite and generates a reproducing kernel Hilbert space by Theorem 5.3.

Already the finite kernel expansions display a broad variety of shapes. In one dimension, put

$$k_c(x) = \exp\left\{-\frac{(x-c)^2}{2(0.70)^2}\right\}.$$

The four functions in Figure 1 are

$$\begin{aligned}
g_1(x) &= 0.8\{k_{-2}(x) + k_{-1}(x) + k_0(x) + k_1(x) + k_2(x)\}, \\
g_2(x) &= -1.5k_{-2}(x) - 0.8k_{-1}(x) + 0.8k_1(x) + 1.5k_2(x), \\
g_3(x) &= k_{-2}(x) - 1.2k_{-1}(x) + 1.2k_0(x) - 1.2k_1(x) + k_2(x), \\
g_4(x) &= 1.2k_{-2.5}(x) + 0.5k_{-1.3}(x) - 1.5k_{0.2}(x) + 0.9k_{1.7}(x).
\end{aligned}$$

Each is an element of the finite span used in the construction, and hence belongs to the Gaussian reproducing kernel Hilbert space.

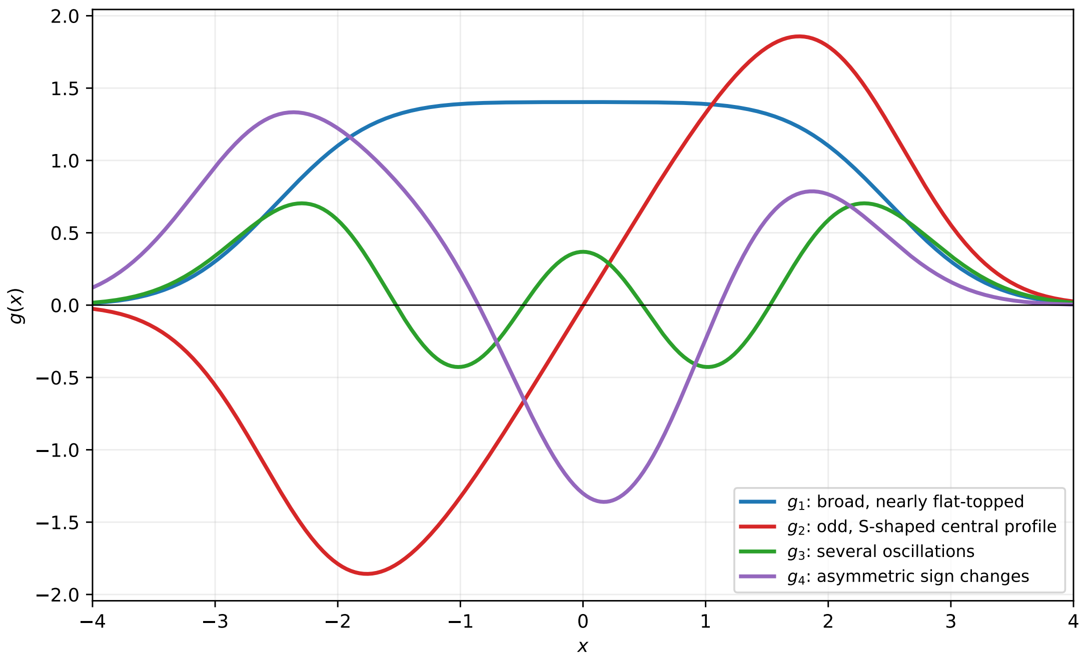

Figure 1: Four finite expansions of one-dimensional Gaussian kernel sections. The curves illustrate a broad nearly flat-topped shape, an odd S-shaped central profile, several oscillations, and asymmetric sign changes.

6 Minimum-norm interpolation

Let H be a reproducing kernel Hilbert space on $\mathcal{X}$ with kernel $\mathcal{K}$. Given sample points $x_1, \ldots, x_n \in \mathcal{X}$ and desired values $y_1, \ldots, y_n \in \mathbb{R}$, consider

$$\underset{f \in H}{\text{minimize}} \ \|f\|_H \qquad \text{subject to} \qquad f(x_i) = y_i, \quad i = 1, \ldots, n. \tag{15}$$

Define the Gram matrix

$$K = \big(\mathcal{K}(x_i, x_j)\big)_{i,j=1}^n$$

and the finite-dimensional subspace

$$M = \operatorname{span}\{\mathcal{K}(\cdot, x_1), \ldots, \mathcal{K}(\cdot, x_n)\}.$$

For $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, write

$$f_\alpha(\cdot) = \sum_{j=1}^n \alpha_j \mathcal{K}(\cdot, x_j).$$

The reproducing property gives, compactly,

$$\big(f_\alpha(x_1), \ldots, f_\alpha(x_n)\big)^T = K\alpha, \qquad \|f_\alpha\|_H^2 = \alpha^T K \alpha. \tag{16}$$

Theorem 6.1 (Minimum-norm interpolation)**.** *The interpolation constraints in* (15) *are feasible if and only if $y = (y_1, \ldots, y_n)^T$ belongs to* $\operatorname{Range}(K)$*. Whenever they are feasible, there is a unique minimum-norm interpolating function, and it belongs to M. Thus it has the form*

$$\widehat{f}(\cdot) = \sum_{j=1}^n \widehat{\alpha}_j \mathcal{K}(\cdot, x_j), \qquad K\widehat{\alpha} = y.$$

If K is invertible, then $\widehat{\alpha} = K^{-1}y$.

Proof. Since M is finite dimensional, it is closed. Decompose any $f \in H$ as $f = f_M + f_\perp$, where $f_M \in M$ and $f_\perp \perp M$. For every sample point,

$$f_\perp(x_i) = \langle f_\perp, \mathcal{K}(\cdot, x_i)\rangle_H = 0,$$

so f and f_M have exactly the same sample values. Since $f_M = f_\alpha$ for some α, the attainable sample-value vectors are precisely the vectors $K\alpha$. This proves the feasibility statement.

If f interpolates the data, then so does f_M, and Pythagoras gives

$$\|f\|_H^2 = \|f_M\|_H^2 + \|f_\perp\|_H^2 \geq \|f_M\|_H^2 .$$

Hence every minimum-norm interpolant lies in M. The minimum-norm function is unique: if two different functions were minimizers, their midpoint would also interpolate and would have strictly smaller squared norm. When K is singular, the representing coefficient vector need not be unique, but any two solutions of $K\alpha = y$ determine the same function. □

7 Penalized fitting and the representer theorem

Exact interpolation is often undesirable when the observations contain noise. Instead, for $\lambda > 0$, consider

$$\widehat{f} = \arg\min_{f \in H} \left\{ \sum_{i=1}^{n} \left(f(x_i) - y_i\right)^2 + \lambda \|f\|_H^2 \right\}. \tag{17}$$

Theorem 7.1 (Finite-sample representer theorem)**.** *The unique minimizer of* (17) *belongs to*

$$M = \operatorname{span}\{\mathcal{K}(\cdot, x_1), \ldots, \mathcal{K}(\cdot, x_n)\}.$$

Consequently,

$$\widehat{f}(\cdot) = \sum_{j=1}^{n} \widehat{\alpha}_j \mathcal{K}(\cdot, x_j)$$

for a suitable coefficient vector $\widehat{\alpha} \in \mathbb{R}^n$.

Proof. Write $f = f_M + f_\perp$ as above. The orthogonal component vanishes at every sample point, while Pythagoras gives $\|f\|_H^2 = \|f_M\|_H^2 + \|f_\perp\|_H^2$. Therefore the objective in (17) satisfies

$$J(f_M + f_\perp) = J(f_M) + \lambda \|f_\perp\|_H^2 \geq J(f_M),$$

with equality only when $f_\perp = 0$. Thus the minimizer belongs to M. □

Using (16), the infinite-dimensional problem reduces to

$$\operatorname*{minimize}_{\alpha \in \mathbb{R}^n} \left\{ \|K\alpha - y\|^2 + \lambda \alpha^T K \alpha \right\}. \tag{18}$$

This is the generalized ridge-regression problem from Section 1, with design matrix K and penalty matrix K. If K is invertible, the normal equation is

$$K(K\alpha - y) + \lambda K \alpha = 0,$$

so

$$\widehat{\alpha} = (K + \lambda I_n)^{-1} y. \tag{19}$$

Accordingly,

$$\widehat{f}(\cdot) = \sum_{j=1}^{n} \big[(K + \lambda I_n)^{-1} y\big]_j \mathcal{K}(\cdot, x_j), \tag{20}$$

and the fitted values at the sample points are

$$\widehat{y} = K(K + \lambda I_n)^{-1} y. \tag{21}$$

Even when K is singular, $K + \lambda I_n$ is invertible and the coefficient vector in (19) gives a minimizing representation. Other coefficient vectors may represent the same minimizing function, differing only by vectors in $\ker K$.

The essential point is that the optimization began over an infinite-dimensional Hilbert space of functions, but the reproducing property and orthogonal projection reduced it to an n-dimensional ridge-regression problem.

7.1 A data-driven choice of the regularization parameter

The amount of regularization can be chosen from the data. For a fixed value of λ and an index i, let $\widehat{f}_\lambda^{(-i)}$ denote the kernel-ridge fit obtained after deleting the observation (x_i, y_i). The leave-one-out cross-validation score is

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \widehat{f}_\lambda^{(-i)}(x_i) \Big)^2. \tag{22}$$

One chooses $\widehat{\lambda}$ to minimize this score over a specified collection of candidate values, and then refits the model to all n observations using $\widehat{\lambda}$.

As an illustration, take $n = 28$ equally spaced points in $[-3, 3]$ and generate

$$y_i = f_0(x_i) + \varepsilon_i, \qquad f_0(x) = 0.8 \sin(1.7x) + 0.35 \cos(0.65x) + 0.55 e^{-2.8(x-1.15)^2}, \tag{23}$$

where the errors are independent $N(0, 0.18^2)$ random variables. We use the one-dimensional Gaussian kernel

$$\mathcal{K}_\sigma(x, y) = \exp\left\{ -\frac{(x-y)^2}{2\sigma^2} \right\} \qquad \text{with } \sigma = 0.55.$$

Evaluating (22) at 141 logarithmically spaced values between 10^{-6} and 10 gives

$$\widehat{\lambda} = 0.0794.$$

After fitting once more to the complete data set, the resulting curve is

$$\widehat{f}(x) = k(x)^T (K + \widehat{\lambda} I_n)^{-1} y, \qquad k(x) = \big(\mathcal{K}_\sigma(x, x_1), \ldots, \mathcal{K}_\sigma(x, x_n)\big)^T. \tag{24}$$

Figure 2 compares the fitted curve with the function that generated the data.

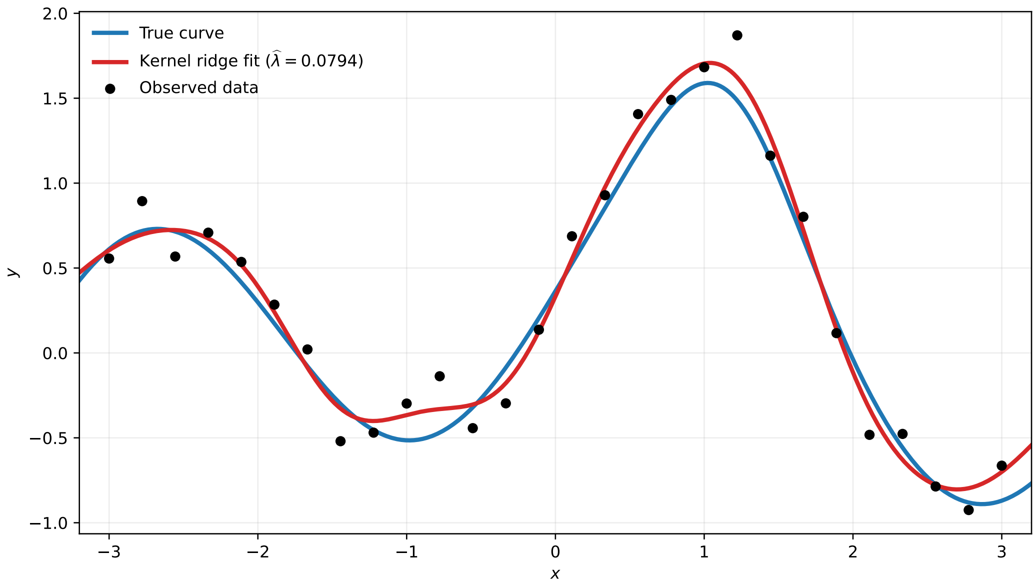

Figure 2: A Gaussian-kernel ridge fit with the regularization parameter selected by leave-one-out cross-validation. The blue curve is the true function f_0, the black points are the observations, and the red curve is the fit obtained using $\widehat{\lambda} = 0.0794$.