

# 2

## Model Construction

The two basic components of a statistical model, the deterministic part and the stochastic part, are well separated in the penalized likelihood score  $L(f) + (\lambda/2)J(f)$  of (1.3). The deterministic part is specified via  $J(f)$ , which defines the notion of smoothness for functions on domain  $\mathcal{X}$ . The stochastic part is characterized by  $L(f)$ , which reflects the sampling structure of the data.

In this chapter, we are mainly concerned with the construction of  $J(f)$  for use in  $L(f) + (\lambda/2)J(f)$ . At the foundation of the construction is some elementary theory of reproducing kernel Hilbert spaces, of which a brief self-contained introduction is given in §2.1. Illustrations of the construction are presented on the domain  $\{1, \dots, K\}$  through shrinkage estimates (§2.2) and on the domain  $[0, 1]$  through polynomial smoothing splines (§2.3); the discrete case also provides insights into the entities in a reproducing kernel Hilbert space through those in a standard vector space. The construction of models on product domains with the ANOVA structure of §1.3.2 built in is discussed in §2.4, with detailed examples on domains  $\{1, \dots, K_1\} \times \{1, \dots, K_2\}$ ,  $[0, 1]^2$ , and  $\{1, \dots, K\} \times [0, 1]$ .

Also included in this chapter are some general properties of the penalized likelihood score  $L(f) + (\lambda/2)J(f)$  that are largely independent of  $L(f)$ . One such property is the fact that a quadratic functional  $J(f)$  acts like the minus log likelihood of a Gaussian process prior for  $f$ , which leads to the Bayes model discussed in §2.5. Other important properties include the existence of the minimizer of  $L(f) + (\lambda/2)J(f)$  and the equivalence of penalized minimization and constrained minimization (§2.6).

The definitions of numerous technical terms are embedded in the text. For convenient back reference, the terms are set in boldface at the point of definition.

Mathematically more sophisticated constructions, such as the thin-plate splines on  $(-\infty, \infty)^d$ , are deferred to Chap. 4.

## 2.1 Reproducing Kernel Hilbert Spaces

By adding a roughness penalty  $J(f)$  to the minus log likelihood  $L(f)$ , one considers only smooth functions in the space  $\{f : J(f) < \infty\}$  or a subspace therein. To assist analysis and computation, one needs a metric and a geometry in the space, and the score  $L(f) + (\lambda/2)J(f)$  to be continuous in  $f$  under the metric. The so-called reproducing kernel Hilbert space, of which a brief introduction is presented here, is adequately equipped for the purpose.

We start with the definition of Hilbert space and some of its elementary properties. The discussion is followed by the Riesz representation theorem, which provides the technical foundation for the notion of a reproducing kernel. The definition of reproducing kernel Hilbert space comes next and it is shown that a reproducing kernel Hilbert space is uniquely determined by its reproducing kernel, for which any non-negative definite function qualifies.

### 2.1.1 Hilbert Spaces and Linear Subspaces

As abstract generalizations of the familiar vector spaces, Hilbert spaces inherit many of the structures of the vector spaces. To provide insights into the technical concepts introduced here, abstract materials are followed by vector space examples set in *italic*.

For elements  $f, g, h, \dots$ , define the operation of **addition** satisfying the following properties: (i)  $f+g = g+f$ , (ii)  $(f+g)+h = f+(g+h)$ , and (iii) for any two elements  $f$  and  $g$ , there exists an element  $h$  such that  $f+h = g$ . The third property implies the existence of an element  $0$  satisfying  $f+0 = f$ . Further, define the operation of **scalar multiplication** satisfying  $\alpha(f+g) = \alpha f + \alpha g$ ,  $(\alpha+\beta)f = \alpha f + \beta f$ ,  $1f = f$ , and  $0f = 0$ , where  $\alpha$  and  $\beta$  are real numbers. A set  $\mathcal{L}$  of such elements form a **linear space** if  $f, g \in \mathcal{L}$  implies that  $f+g \in \mathcal{L}$  and  $\alpha f \in \mathcal{L}$ . A set of elements  $f_i \in \mathcal{L}$  are said to be **linearly independent** if  $\sum_i \alpha_i f_i = 0$  holds only for  $\alpha_i = 0, \forall i$ . The maximum number of elements in  $\mathcal{L}$  that can be linearly independent defines its **dimension**.

*Take real vectors of a given length as the elements; the standard vector addition and scalar-vector multiplication satisfy the conditions specified for*

the operations of addition and scalar multiplication. The notions of linear space, linear independence, and dimension reduce to those in standard vector spaces.

A **functional** in a linear space  $\mathcal{L}$  operates on an element  $f \in \mathcal{L}$  and returns a real number as its value. A **linear functional**  $L$  in  $\mathcal{L}$  satisfies  $L(f + g) = Lf + Lg$ ,  $L(\alpha f) = \alpha Lf$ ,  $f, g \in \mathcal{L}$ ,  $\alpha$  real. A **bilinear form**  $J(f, g)$  in a linear space  $\mathcal{L}$  takes  $f, g \in \mathcal{L}$  as arguments and returns a real value and satisfies  $J(\alpha f + \beta g, h) = \alpha J(f, h) + \beta J(g, h)$ ,  $J(f, \alpha g + \beta h) = \alpha J(f, g) + \beta J(f, h)$ ,  $f, g, h \in \mathcal{L}$ ,  $\alpha, \beta$  real. Fixing one argument in a bilinear form, one gets a linear functional in the other argument. A bilinear form  $J(\cdot, \cdot)$  is **symmetric** if  $J(f, g) = J(g, f)$ . A symmetric bilinear form is **non-negative definite** if  $J(f, f) \geq 0$ ,  $\forall f \in \mathcal{L}$ , and it is **positive definite** if the equality holds only for  $f = 0$ . For  $J(\cdot, \cdot)$  non-negative definite,  $J(f) = J(f, f)$  is called a **quadratic functional**.

Consider the linear space of all real vectors of a given length. A functional in such a space is simply a multivariate function with the coordinates of the vector as its arguments. A linear functional in such a space can be written as a dot product,  $Lf = g_L^T f$ , where  $g_L$  is a vector “representing”  $L$ . A bilinear form can be written as  $J(f, g) = f^T B_J g$  with  $B_J$  a square matrix, and  $J(f, g)$  is symmetric, non-negative definite, or positive definite when  $B_J$  is symmetric, non-negative definite, or positive definite. A quadratic functional  $J(f) = f^T B_J f$  is better known as a quadratic form in the classical linear model theory.

A linear space is often equipped with an **inner product**, a positive definite bilinear form with a notation  $(\cdot, \cdot)$ . An inner product defines a **norm** in the linear space,  $\|f\| = \sqrt{(f, f)}$ , which induces a metric to measure the **distance** between elements in the space,  $D[f, g] = \|f - g\|$ . The Cauchy-Schwarz inequality,

$$|(f, g)| \leq \|f\| \|g\|, \quad (2.1)$$

with equality if and only if  $f = \alpha g$ , and the triangle inequality,

$$\|f + g\| \leq \|f\| + \|g\|, \quad (2.2)$$

with equality if and only if  $f = \alpha g$  for some  $\alpha > 0$ , hold in such a linear space; see Problems 2.1 and 2.2.

Equip the linear space of all real vectors of a given length with an inner product  $(f, g) = f^T g$ ; one obtains the Euclidean space. The Euclidean norm  $\|f\| = \sqrt{f^T f}$  induces the familiar Euclidean distance between vectors. The Cauchy-Schwarz inequality and the triangle inequality are familiar results in a Euclidean space.

When  $\lim_{n \rightarrow \infty} \|f_n - f\| = 0$  for a sequence of elements  $f_n$ , the sequence is said to **converge** to its **limit point**  $f$ , with a notation  $\lim_{n \rightarrow \infty} f_n = f$  or  $f_n \rightarrow f$ . A functional  $L$  is **continuous** if  $\lim_{n \rightarrow \infty} Lf_n = Lf$  whenever  $\lim_{n \rightarrow \infty} f_n = f$ . By the Cauchy-Schwarz inequality,  $(f, g)$  is continuous in  $f$  or  $g$  when the other argument is fixed.

In the Euclidean space, a functional is a multivariate function in the coordinates of the vector, and the definition of continuity reduces to the definition found in standard multivariate calculus.

A sequence satisfying  $\lim_{n,m \rightarrow \infty} \|f_n - f_m\| = 0$  is called a **Cauchy sequence**. A linear space  $\mathcal{L}$  is **complete** if every Cauchy sequence in  $\mathcal{L}$  converges to an element in  $\mathcal{L}$ . An element is a **limit point of a set**  $A$  if it is the limit point of a sequence in  $A$ . A set  $A$  is **closed** if it contains all of its own limit points.

The Euclidean space is complete. In the two-dimensional Euclidean space,  $(-\infty, \infty) \times \{0\}$  is a closed set, so is  $[a_1, b_1] \times [a_2, b_2]$ , where  $-\infty < a_i \leq b_i < \infty$ ,  $i = 1, 2$ .

A **Hilbert space**  $\mathcal{H}$  is a complete inner product linear space. A closed linear subspace of  $\mathcal{H}$  is itself a Hilbert space. The **distance** between a point  $f \in \mathcal{H}$  and a closed linear subspace  $\mathcal{G} \subset \mathcal{H}$  is defined by  $D[f, \mathcal{G}] = \inf_{g \in \mathcal{G}} \|f - g\|$ . By the closedness of  $\mathcal{G}$ , there exists an  $f_{\mathcal{G}} \in \mathcal{G}$ , called the **projection** of  $f$  in  $\mathcal{G}$ , such that  $\|f - f_{\mathcal{G}}\| = D[f, \mathcal{G}]$ . Such an  $f_{\mathcal{G}}$  is unique by the triangle inequality. See Problem 2.3.

In the two-dimensional Euclidean space,  $\mathcal{G} = \{f : f = (a, 0)^T, a \text{ real}\}$  is a closed linear subspace. The distance between  $f = (a_f, b_f)^T$  and  $\mathcal{G}$  is  $D[f, \mathcal{G}] = |b_f|$ , and the projection of  $f$  in  $\mathcal{G}$  is  $f_{\mathcal{G}} = (a_f, 0)^T$ .

**Proposition 2.1** Let  $f_{\mathcal{G}}$  be the projection of  $f \in \mathcal{H}$  in a closed linear subspace  $\mathcal{G} \subset \mathcal{H}$ . Then,  $(f - f_{\mathcal{G}}, g) = 0, \forall g \in \mathcal{G}$ .

*Proof:* We prove by negation. Suppose  $(f - f_{\mathcal{G}}, h) = \alpha \neq 0, h \in \mathcal{G}$ . Write  $\beta = (h, h)$  and take  $g = f_{\mathcal{G}} + (\alpha/\beta)h \in \mathcal{G}$ . It is easy to compute

$$\|f - g\|^2 = \|f - f_{\mathcal{G}}\|^2 - \alpha^2/\beta < \|f - f_{\mathcal{G}}\|^2,$$

a contradiction.  $\square$

The linear subspace  $\mathcal{G}^c = \{f : (f, g) = 0, \forall g \in \mathcal{G}\}$  is called the **orthogonal complement** of  $\mathcal{G}$ . By the continuity of  $(f, g)$ ,  $\mathcal{G}^c$  is closed. Using Proposition 2.1, it is easy to verify that

$$\begin{aligned} \|f - f_{\mathcal{G}} - f_{\mathcal{G}^c}\|^2 &= (f - f_{\mathcal{G}} - f_{\mathcal{G}^c}, f - f_{\mathcal{G}} - f_{\mathcal{G}^c}) \\ &= (f - f_{\mathcal{G}}, f - f_{\mathcal{G}^c}) - (f - f_{\mathcal{G}}, f_{\mathcal{G}}) \\ &\quad - (f_{\mathcal{G}^c}, f - f_{\mathcal{G}^c}) + (f_{\mathcal{G}^c}, f_{\mathcal{G}}) \\ &= 0, \end{aligned}$$

where  $f_{\mathcal{G}} \in \mathcal{G}$  and  $f_{\mathcal{G}^c} \in \mathcal{G}^c$  are the projections of  $f$  in  $\mathcal{G}$  and  $\mathcal{G}^c$ , respectively. Hence, there exists a unique decomposition  $f = f_{\mathcal{G}} + f_{\mathcal{G}^c}$  for every  $f \in \mathcal{H}$ . It is clear now that  $(\mathcal{G}^c)^c = \mathcal{G}$ . The decomposition  $f = f_{\mathcal{G}} + f_{\mathcal{G}^c}$  is called a **tensor sum decomposition** and is denoted by  $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^c$ ,  $\mathcal{G}^c = \mathcal{H} \ominus \mathcal{G}$ , or  $\mathcal{G} = \mathcal{H} \ominus \mathcal{G}^c$ . Multiple-term tensor sum decompositions can be defined recursively.

In the two-dimensional Euclidean space, the orthogonal complement of  $\mathcal{G} = \{f : f = (a, 0)^T, a \text{ real}\}$  is  $\mathcal{G}^c = \{f : f = (0, b)^T, b \text{ real}\}$ .

Consider linear subspaces  $\mathcal{H}_0$  and  $\mathcal{H}_1$  of a linear space  $\mathcal{L}$ , equipped with inner products  $(\cdot, \cdot)_0$  and  $(\cdot, \cdot)_1$ , respectively. Assume the completeness of  $\mathcal{H}_0$  and  $\mathcal{H}_1$  so that they are Hilbert spaces. If  $\mathcal{H}_0$  and  $\mathcal{H}_1$  have only one common element 0, then one may define a tensor sum Hilbert space  $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$  with elements  $f = f_0 + f_1$  and  $g = g_0 + g_1$ , where  $f_0, g_0 \in \mathcal{H}_0$  and  $f_1, g_1 \in \mathcal{H}_1$ , and an inner product  $(f, g) = (f_0, g_0)_0 + (f_1, g_1)_1$ . It is easy to verify that such a bottom-up pasting is consistent with the aforementioned top-down decomposition; see Problem 2.4.

Consider the two-dimensional vector space. Equip the space  $\mathcal{H}_0 = \{f : f = (a, 0)^T, a \text{ real}\}$  with the inner product  $(f, g)_0 = a_f a_g$ , where  $f = (a_f, 0)^T$  and  $g = (a_g, 0)^T$ , and equip  $\mathcal{H}_1 = \{f : f = (0, b)^T, b \text{ real}\}$  with the inner product  $(f, g)_1 = b_f b_g$ , where  $f = (0, b_f)^T$  and  $g = (0, b_g)^T$ .  $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$  has elements of the form  $f = f_0 + f_1 = (a_f, 0)^T + (0, b_f)^T = (a_f, b_f)^T$  and  $g = (a_g, 0)^T + (0, b_g)^T = (a_g, b_g)^T$ , and an inner product  $(f, g) = (f_0, g_0)_0 + (f_1, g_1)_1 = a_f a_g + b_f b_g$ .

A non-negative definite bilinear form  $J(f, g)$  in a linear space  $\mathcal{H}$  defines a **semi-inner-product** in  $\mathcal{H}$  which induces a square **seminorm**  $J(f) = J(f, f)$ . Unless  $J(f, g)$  is positive definite, the **null space**  $\mathcal{N}_J = \{f : J(f, f) = 0, f \in \mathcal{H}\}$  is a linear subspace of  $\mathcal{H}$  containing more elements than just 0. With a nondegenerate  $\mathcal{N}_J$ , one typically can define another non-negative definite bilinear form  $\tilde{J}(f, g)$  in  $\mathcal{H}$  satisfying the following conditions: (i) it is positive definite when restricted to  $\mathcal{N}_J$ , so  $\tilde{J}(f) = \tilde{J}(f, f)$  defines a square full norm in  $\mathcal{N}_J$  and (ii) for every  $f \in \mathcal{H}$ , there exists  $g \in \mathcal{N}_J$  such that  $\tilde{J}(f - g) = 0$ . With such an  $\tilde{J}(f, g)$ , it is easy to verify that  $J(f, g)$  is positive definite in the linear subspace  $\mathcal{N}_{\tilde{J}} = \{f : \tilde{J}(f, f) = 0, f \in \mathcal{H}\}$  and that  $(J + \tilde{J})(f, g)$  is positive definite in  $\mathcal{H}$ . Hence, a semi-inner-product can be made a full inner product either via restriction to a subspace or via augmentation by an extra term, both through the definition of an inner product in its null space. If  $\mathcal{H}$  is complete under the norm induced by  $(J + \tilde{J})(f, g)$ , then it is easy to see that  $\mathcal{N}_J$  and  $\mathcal{N}_{\tilde{J}}$  form a tensor sum decomposition of  $\mathcal{H}$ .

In the two-dimensional vector space  $\mathcal{H}$  with elements  $f = (a_f, b_f)^T$  and  $g = (a_g, b_g)^T$ ,  $J(f, g) = b_f b_g$  defines a semi-inner-product with the null space  $\mathcal{N}_J = \{f : f = (a, 0)^T, a \text{ real}\}$ . Define  $\tilde{J}(f, g) = a_f a_g$ , which satisfies the two conditions specified above. It follows that  $\mathcal{N}_{\tilde{J}} = \{f : f = (0, b)^T, b \text{ real}\}$ , in which  $J(f, g) = b_f b_g$  is positive definite. Clearly,  $(J + \tilde{J})(f, g) = b_f b_g + a_f a_g$  is positive definite in  $\mathcal{H}$ .

**Example 2.1 ( $L_2$  space)** All square integrable functions on  $[0, 1]$  form a Hilbert space

$$\mathcal{L}_2[0, 1] = \{f : \int_0^1 f^2 dx < \infty\}$$

with an inner product  $(f, g) = \int_0^1 fg \, dx$ . The space

$$\mathcal{G} = \{f : f = gI_{[x \leq 0.5]}, g \in \mathcal{L}_2[0, 1]\}$$

is a closed linear subspace with an orthogonal complement

$$\mathcal{G}^c = \{f : f = gI_{[x \geq 0.5]}, g \in \mathcal{L}_2[0, 1]\}.$$

Note that elements in  $\mathcal{L}_2[0, 1]$  are defined not by individual functions but by equivalent classes.

The bilinear form  $J(f, g) = \int_0^{0.5} fg \, dx$  defines a semi-inner-product in  $\mathcal{L}_2[0, 1]$ , with a null space

$$\mathcal{N}_J = \mathcal{G}^c = \{f : f = gI_{[x \geq 0.5]}, g \in \mathcal{L}_2[0, 1]\}.$$

Define  $\tilde{J}(f, g) = C \int_{0.5}^1 fg \, dx$ , with  $C > 0$  a constant; one has an inner product  $(f, g) = (J + \tilde{J})(f, g) = \int_0^{0.5} fg \, dx + C \int_{0.5}^1 fg \, dx$  on  $\mathcal{L}_2[0, 1]$ . On  $\mathcal{G} = \mathcal{L}_2 \ominus \mathcal{N}_J$ ,  $J(f, g)$  is a full inner product.  $\square$

**Example 2.2 (Euclidean space)** Functions on  $\{1, \dots, K\}$  are vectors of length  $K$ . Consider the Euclidean  $K$ -space with an inner product

$$(f, g) = \sum_{x=1}^K f(x)g(x) = f^T g.$$

The space  $\mathcal{G} = \{f : f(1) = \dots = f(K)\}$  is a closed linear subspace with an orthogonal complement  $\mathcal{G}^c = \{f : \sum_{x=1}^K f(x) = 0\}$ .

Write  $\bar{f} = \sum_{x=1}^K f(x)/K$ . The bilinear form

$$J(f, g) = \sum_{x=1}^K (f(x) - \bar{f})(g(x) - \bar{g}) = f^T \left( I - \frac{1}{K} \mathbf{1}\mathbf{1}^T \right) g$$

defines a semi-inner-product in the vector space with a null space

$$\mathcal{N}_J = \mathcal{G} = \{f : f(1) = \dots = f(K)\}.$$

Define  $\tilde{J}(f, g) = C \bar{f} \bar{g} = C f^T (\mathbf{1}\mathbf{1}^T / K) g$ , with  $C > 0$  a constant; one has an inner product in the vector space,

$$(f, g) = (J + \tilde{J})(f, g) = f^T \left( I + \frac{C-1}{K} \mathbf{1}\mathbf{1}^T \right) g,$$

which reduces to the Euclidean inner product when  $C = 1$ . On  $\mathcal{G}^c = \{f : \sum_{x=1}^K f(x) = 0\}$ ,  $J(f, g)$  is a full inner product.  $\square$

### 2.1.2 Riesz Representation Theorem

For every  $g$  in a Hilbert space  $\mathcal{H}$ ,  $L_g f = (g, f)$  defines a continuous linear functional  $L_g$ . Conversely, every continuous linear functional  $L$  in  $\mathcal{H}$  has a representation  $Lf = (g_L, f)$  for some  $g_L \in \mathcal{H}$ , called the **representer** of  $L$ , as the following theorem asserts.

**Theorem 2.2 (Riesz representation)** *For every continuous linear functional  $L$  in a Hilbert space  $\mathcal{H}$ , there exists a unique  $g_L \in \mathcal{H}$  such that  $Lf = (g_L, f)$ ,  $\forall f \in \mathcal{H}$ .*

*Proof:* Let  $\mathcal{N}_L = \{f : Lf = 0\}$  be the null space of  $L$ . Since  $L$  is continuous,  $\mathcal{N}_L$  is a closed linear subspace. If  $\mathcal{N}_L = \mathcal{H}$ , take  $g_L = 0$ . When  $\mathcal{N}_L \subset \mathcal{H}$ , there exists a nonzero element  $g_0 \in \mathcal{H} \ominus \mathcal{N}_L$ . Since  $(Lf)g_0 - (Lg_0)f \in \mathcal{N}_L$ ,  $((Lf)g_0 - (Lg_0)f, g_0) = 0$ . Some algebra yields

$$Lf = \left( \frac{Lg_0}{(g_0, g_0)} g_0, f \right).$$

Hence, one can take  $g_L = (Lg_0)g_0/(g_0, g_0)$ . The uniqueness is trivial.  $\square$

The continuity of  $L$  is necessary for the theorem to hold, or otherwise  $\mathcal{N}_L$  is no longer closed and the proof breaks down.

All linear functionals in a finite-dimensional Hilbert space are continuous. Actually, there is an isomorphism between any  $K$ -dimensional Hilbert space and the Euclidean  $K$ -space. See Problems 2.5 and 2.6.

### 2.1.3 Reproducing Kernel and Non-Negative Definite Function

The likelihood part  $L(f)$  of the penalized likelihood functional  $L(f) + (\lambda/2)J(f)$  usually involves evaluations; thus, for it to be continuous in  $f$ , one needs the continuity of the **evaluation functional**  $[x]f = f(x)$ . Consider a Hilbert space  $\mathcal{H}$  of functions on domain  $\mathcal{X}$ . If the evaluation functional  $[x]f = f(x)$  is continuous in  $\mathcal{H}$ ,  $\forall x \in \mathcal{X}$ , then  $\mathcal{H}$  is called a **reproducing kernel Hilbert space**.

By the Riesz representation theorem, there exists  $R_x \in \mathcal{H}$ , the representer of the evaluation functional  $[x](\cdot)$ , such that  $(R_x, f) = f(x)$ ,  $\forall f \in \mathcal{H}$ . The symmetric bivariate function  $R(x, y) = R_x(y) = (R_x, R_y)$  has the reproducing property  $(R(x, \cdot), f(\cdot)) = f(x)$  and is called the **reproducing kernel** of the space  $\mathcal{H}$ . The reproducing kernel is unique when it exists (Problem 2.7).

The  $\mathcal{L}_2[0, 1]$  space of Example 2.1 is not a reproducing kernel Hilbert space. In fact, since the elements in  $\mathcal{L}_2[0, 1]$  are defined by equivalent classes but not individual functions, evaluation is not even well defined. A finite-dimensional Hilbert space is always a reproducing kernel Hilbert space since all linear functionals are continuous.

**Example 2.3 (Euclidean space)** Consider again the Euclidean  $K$ -space with  $(f, g) = f^T g$ , with vectors perceived as functions on  $\mathcal{X} = \{1, \dots, K\}$ . The evaluation functional  $[x]f = f(x)$  is simply coordinate extraction. Since  $f(x) = e_x^T f$ , where  $e_x$  is the  $x$ th unit vector, one has  $R_x(y) = I_{[x=y]}$ . A bivariate function on  $\{1, \dots, K\}$  can be written as a square matrix, and the reproducing kernel in the Euclidean space is simply the identity matrix.  $\square$

A bivariate function  $F(x, y)$  on  $\mathcal{X}$  is said to be a **non-negative definite function** if  $\sum_{i,j} \alpha_i \alpha_j F(x_i, x_j) \geq 0$ ,  $\forall x_i \in \mathcal{X}$ ,  $\forall \alpha_i$  real. For  $R(x, y) = R_x(y)$  a reproducing kernel, it is easy to verify that

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

so  $R(x, y)$  is non-negative definite. As a matter of fact, there exists a one-to-one correspondence between reproducing kernel Hilbert spaces and non-negative definite functions, as the following theorem asserts.

**Theorem 2.3** *For every reproducing kernel Hilbert space  $\mathcal{H}$  of functions on  $\mathcal{X}$ , there corresponds an unique reproducing kernel  $R(x, y)$ , which is non-negative definite. Conversely, for every non-negative definite function  $R(x, y)$  on  $\mathcal{X}$ , there corresponds a unique reproducing kernel Hilbert space  $\mathcal{H}$  that has  $R(x, y)$  as its reproducing kernel.*

By Theorem 2.3, one may construct a reproducing kernel Hilbert space simply by specifying its reproducing kernel. The following lemma is needed in the proof of the theorem.

**Lemma 2.4** *Let  $R(x, y)$  be any non-negative definite function on  $\mathcal{X}$ . If*

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j R(x_i, x_j) = 0,$$

*then  $\sum_{i=1}^n \alpha_i R(x_i, x) = 0$ ,  $\forall x \in \mathcal{X}$ .*

*Proof:* Augment the  $(x_i, \alpha_i)$  sequence by adding  $(x_0, \alpha_0)$ , where  $x_0 \in \mathcal{X}$  and  $\alpha_0$  real are arbitrary. Since

$$0 \leq \sum_{i=0}^n \sum_{j=0}^n \alpha_i \alpha_j R(x_i, x_j) = 2\alpha_0 \sum_{i=1}^n \alpha_i R(x_i, x_0) + \alpha_0^2 R(x_0, x_0)$$

and  $R(x_0, x_0) \geq 0$ , it is necessary that  $\sum_{i=1}^n \alpha_i R(x_i, x_0) = 0$ .  $\square$

*Proof of Theorem 2.3:* Only the converse needs a proof. Given  $R(x, y)$ , write  $R_x = R(x, \cdot)$ ; one starts with the linear space

$$\mathcal{H}^* = \{f : f = \sum_i \alpha_i R_{x_i}, x_i \in \mathcal{X}, \alpha_i \text{ real}\},$$



and defines in  $\mathcal{H}^*$  an inner product

$$\left( \sum_i \alpha_i R_{x_i}, \sum_j \beta_j R_{y_j} \right) = \sum_{i,j} \alpha_i \beta_j R(x_i, y_j).$$

It is trivial to verify the properties of inner product for such a  $(f, g)$ , except that  $(f, f) = 0$  holds only for  $f = 0$ , which is proved in Lemma 2.4. It is also easy to verify that  $(R_x, f) = f(x)$ ,  $\forall f \in \mathcal{H}^*$ .

By the Cauchy-Schwarz inequality,

$$|f(x)| = |(R_x, f)| \leq \sqrt{R(x, x)} \|f\|,$$

so convergence in norm implies pointwise convergence. For every Cauchy sequence  $\{f_n\}$  in  $\mathcal{H}^*$ ,  $\{f_n(x)\}$  is a Cauchy sequence on the real line converging to a limit. Note also that  $|\|f_n\| - \|f_m\|| \leq \|f_n - f_m\|$ , so  $\{\|f_n\|\}$  has a limit as well. The limit point of  $\{f_n\}$  can then be defined by  $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ ,  $\forall x \in \mathcal{X}$ , with  $\|f\| = \lim_{n \rightarrow \infty} \|f_n\|$ . It will be shown shortly that  $\|f\|$ , thus defined, is unique; that is, for two Cauchy sequences  $\{f_n\}$  and  $\{g_n\}$  satisfying  $\lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} g_n(x)$ ,  $\forall x \in \mathcal{X}$ , it is necessary that  $\lim_{n \rightarrow \infty} \|f_n\| = \lim_{n \rightarrow \infty} \|g_n\|$ . Adjoining all these limit points of Cauchy sequences to  $\mathcal{H}^*$ , one obtains a complete linear space  $\mathcal{H}$  with the norm  $\|f\|$ . It is easy to verify that  $(f, g) = (\|f + g\|^2 - \|f\|^2 - \|g\|^2)/2$  extends the inner product from  $\mathcal{H}^*$  to  $\mathcal{H}$  and that  $(R_x, f) = f(x)$  holds in  $\mathcal{H}$ , so  $\mathcal{H}$  is a reproducing kernel Hilbert space with  $R(x, y)$  as its reproducing kernel.

We now verify the uniqueness of the definition of  $\|f\|$  in the completed space, and it suffices to show that for every Cauchy sequence  $\{f_n\}$  in  $\mathcal{H}^*$  satisfying  $\lim_{n \rightarrow \infty} f_n(x) = 0$ ,  $\forall x \in \mathcal{X}$ , it necessarily holds that  $\lim_{n \rightarrow \infty} \|f_n\| = 0$ . We prove the assertion by negation. Suppose  $f_n(x) \rightarrow 0$ ,  $\forall x \in \mathcal{X}$ , but  $\|f_n\|^2 \rightarrow 3\delta > 0$ . Take  $\epsilon \in (0, \delta)$ . For  $n$  and  $m$  sufficiently large, one has  $\|f_n\|^2, \|f_m\|^2 > 2\delta$  and  $\|f_n - f_m\|^2 < \epsilon$ . Fix such an  $m$  and write  $f_m = \sum_i \alpha_i R_{x_i}$  a finite sum. Since  $f_n(x) \rightarrow 0$ ,  $\forall x \in \mathcal{X}$ , it follows that  $\sum_i \alpha_i f_n(x_i) \rightarrow 0$ . Hence, for  $n$  sufficiently large,

$$|(f_n, f_m)| = |(f_n, \sum_i \alpha_i R_{x_i})| = |\sum_i \alpha_i f_n(x_i)| < \epsilon.$$

Now,

$$\epsilon > \|f_n - f_m\|^2 = \|f_n\|^2 + \|f_m\|^2 - 2(f_n, f_m) > 4\delta - 2\epsilon > 2\delta,$$

a contradiction.

It remains to be shown that if a space  $\tilde{\mathcal{H}}$  has  $R(x, y)$  as its reproducing kernel, then  $\tilde{\mathcal{H}}$  must be identical to the space  $\mathcal{H}$  constructed above. Since  $R_x = R(x, \cdot) \in \tilde{\mathcal{H}}$ ,  $\forall x \in \mathcal{X}$ , so  $\mathcal{H} \subseteq \tilde{\mathcal{H}}$ . Now, for any  $h \in \tilde{\mathcal{H}} \ominus \mathcal{H}$ , by orthogonality,  $h(x) = (R_x, h) = 0$ ,  $\forall x \in \mathcal{X}$ , so  $\tilde{\mathcal{H}} = \mathcal{H}$ . The proof is now complete.  $\square$

From the construction in the proof, one can see that the space  $\mathcal{H}$  corresponding to  $R$  is generated from the “columns”  $R_x = R(\cdot, x)$  of  $R$ , very much like a vector space generated from the columns of a matrix.

In the sections to follow, we will be constantly decomposing reproducing kernel Hilbert spaces into tensor sums or pasting up larger spaces by taking tensor sums of smaller ones. The following theorem spells out some of the rules in such operations.

**Theorem 2.5** *If the reproducing kernel  $R$  of a space  $\mathcal{H}$  on domain  $\mathcal{X}$  can be decomposed into  $R = R_0 + R_1$ , where  $R_0$  and  $R_1$  are both non-negative definite,  $R_0(x, \cdot), R_1(x, \cdot) \in \mathcal{H}, \forall x \in \mathcal{X}$ , and  $(R_0(x, \cdot), R_1(y, \cdot)) = 0, \forall x, y \in \mathcal{X}$ , then the spaces  $\mathcal{H}_0$  and  $\mathcal{H}_1$  corresponding respectively to  $R_0$  and  $R_1$  form a tensor sum decomposition of  $\mathcal{H}$ . Conversely, if  $R_0$  and  $R_1$  are both non-negative definite and  $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$ , then  $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$  has a reproducing kernel  $R = R_0 + R_1$ .*

*Proof:* By the orthogonality between  $R_0(x, \cdot)$  and  $R_1(y, \cdot)$ ,

$$R_0(x, y) = (R_0(x, \cdot), R(y, \cdot)) = (R_0(x, \cdot), R_0(y, \cdot)),$$

so the inner product in  $\mathcal{H}_0$  is consistent with that in  $\mathcal{H}$ ; hence,  $\mathcal{H}_0$  is a closed linear subspace of  $\mathcal{H}$ . Now, for every  $f \in \mathcal{H}$ , let  $f_0$  be the projection of  $f$  in  $\mathcal{H}_0$  and write  $f = f_0 + f_0^c$ . Straightforward calculation yields

$$\begin{aligned} f(x) &= (R(x, \cdot), f) \\ &= (R_0(x, \cdot), f_0) + (R_0(x, \cdot), f_0^c) + (R_1(x, \cdot), f_0) + (R_1(x, \cdot), f_0^c) \\ &= f_0(x) + (R_1(x, \cdot), f_0^c), \end{aligned}$$

so  $(R_1(x, \cdot), f_0^c) = f(x) - f_0(x) = f_0^c(x)$ . This shows that  $R_1$  is the reproducing kernel of  $\mathcal{H} \ominus \mathcal{H}_0$ ; hence,  $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ .

For the converse, it is trivial to verify that

$$(R(x, \cdot), f) = (R_0(x, \cdot), f_0)_0 + (R_1(x, \cdot), f_1)_1 = f_0(x) + f_1(x) = f(x),$$

where  $f = f_0 + f_1 \in \mathcal{H}$  with  $f_0 \in \mathcal{H}_0$  and  $f_1 \in \mathcal{H}_1$ , and  $(\cdot, \cdot)_0$  and  $(\cdot, \cdot)_1$  are the inner products in  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , respectively.  $\square$

## 2.2 Smoothing Splines on $\{1, \dots, K\}$

As discussed in Example 2.3, a function on the discrete domain  $\mathcal{X} = \{1, \dots, K\}$  is a vector of length  $K$ , evaluation is coordinate extraction, and a reproducing kernel can be written as a non-negative definite matrix. A linear functional in a finite-dimensional space is always continuous, so a vector space equipped with an inner product is a reproducing kernel Hilbert space.

Let  $B$  be any  $K \times K$  non-negative definite matrix. Consider the column space of  $B$ ,  $\mathcal{H}_B = \{f : f = B\mathbf{c} = \sum_j c_j B(\cdot, j)\}$ , equipped with the inner product  $(f, g) = f^T B g$ . The standard eigenvalue decomposition gives

$$B = U D U^T = (U_1, U_2) \begin{pmatrix} D_1 & O \\ O & O \end{pmatrix} \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} = U_1 D_1 U_1^T,$$

where the diagonal of  $D_1$  contains the positive eigenvalues of  $B$  and the columns of  $U_1$  are the associated eigenvectors. The Moore-Penrose inverse of  $B$  has an expression  $B^+ = U_1 D_1^{-1} U_1^T$ . It is clear that  $\mathcal{H}_B = \mathcal{H}_{B^+} = \{f : f = U_1 \mathbf{c}\}$ . Now,  $B^+ B = U_1 U_1^T$  is the projection matrix onto  $\mathcal{H}_B$ , so  $B^+ B f = f$ ,  $\forall f \in \mathcal{H}_B$ . It then follows that

$$[x]f = f(x) = e_x^T f = e_x^T B^+ B f = (B^+ e_x)^T B f,$$

$\forall f \in \mathcal{H}_B$  (i.e., the representer of  $[x](\cdot)$  is the  $x$ th column of  $B^+$ ). Hence, the reproducing kernel is given by  $R(x, y) = B^+(x, y)$ , where  $B^+(x, y)$  is the  $(x, y)$ th entry of  $B^+$ . The result of Example 2.3 is a trivial special case with  $B = I$ .

The duality between  $(f, g) = f^T B g$  and  $R = B^+$  provides a useful insight into the relation between the inner product in a space and the corresponding reproducing kernel: *In a sense, the inner product and the reproducing kernel are inverses of each other.*

Now, consider a decomposition of the reproducing kernel in the Euclidean  $K$ -space,  $R(x, y) = I_{[x=y]} = 1/K + (I_{[x=y]} - 1/K)$ , or in matrix terms,  $I = (\mathbf{1}\mathbf{1}^T/K) + (I - \mathbf{1}\mathbf{1}^T/K)$ . Since  $(\mathbf{1}\mathbf{1}^T/K)(I - \mathbf{1}\mathbf{1}^T/K) = O$ ,  $(R_0(x, \cdot), R_1(y, \cdot)) = 0$ ,  $\forall x, y$ . By Theorem 2.5, the decomposition defines a tensor sum decomposition of the space  $R^K = \mathcal{H}_0 \oplus \mathcal{H}_1$ , where  $\mathcal{H}_0 = \{f : f(1) = \dots = f(K)\}$  and  $\mathcal{H}_1 = \{f : \sum_{x=1}^K f(x) = 0\}$ . The inner products in  $\mathcal{H}_0$  and  $\mathcal{H}_1$  have expressions  $(f, g)_0 = f^T g = f^T (\mathbf{1}\mathbf{1}^T/K) g$  and  $(f, g)_1 = f^T g = f^T (I - \mathbf{1}\mathbf{1}^T/K) g$ , respectively, where  $\mathbf{1}\mathbf{1}^T/K$  is the Moore-Penrose inverse of  $R_0 = \mathbf{1}\mathbf{1}^T/K$  and  $I - \mathbf{1}\mathbf{1}^T/K$  is the Moore-Penrose inverse of  $R_1 = I - \mathbf{1}\mathbf{1}^T/K$ . The decomposition defines a one-way ANOVA decomposition with an averaging operator  $Af = \sum_{x=1}^K f(x)/K$ . See Problem 2.8 for a construction yielding a one-way ANOVA decomposition with an averaging operator  $Af = f(1)$ .

Regression on  $\mathcal{X} = \{1, \dots, K\}$  yields the classical one-way ANOVA model. Consider a roughness penalty

$$J(f) = \sum_{x=1}^K (f(x) - \bar{f})^2 = f^T \left( I - \frac{\mathbf{1}\mathbf{1}^T}{K} \right) f,$$

where  $\bar{f} = \sum_{x=1}^K f(x)/K$ . The minimizer of

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda \sum_{x=1}^K (\eta(x) - \bar{\eta})^2 \quad (2.3)$$

defines a shrinkage estimate being shrunk toward a constant. Similarly, if one sets  $J(f) = f^T f$ , then the minimizer of

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda \sum_{x=1}^K \eta^2(x) \quad (2.4)$$

defines a shrinkage estimate being shrunk toward zero. Hence, smoothing splines on a discrete domain reduce to shrinkage estimates.

The roughness penalty  $\sum_{x=1}^K (f(x) - \bar{f})^2$  appears natural for  $x$  nominal. For  $x$  ordinal, however, one may consider alternatives such as

$$\sum_{x=2}^K (f(x) - f(x-1))^2,$$

which have the same null space but use different “scaling” in the penalized contrast space  $\mathcal{H}_1 = \{f : \sum_{x=1}^K f(x) = 0\}$ .

## 2.3 Polynomial Smoothing Splines on $[0, 1]$

The cubic smoothing spline of §1.1.1 is a special case of the polynomial smoothing splines, the minimizers of

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda \int_0^1 (\eta^{(m)})^2 dx, \quad (2.5)$$

in the space  $\mathcal{C}^{(m)}[0, 1] = \{f : f^{(m)} \in \mathcal{L}_2[0, 1]\}$ . Equipped with appropriate inner products, the space  $\mathcal{C}^{(m)}[0, 1]$  can be made a reproducing kernel Hilbert space.

We will present two such constructions and outline an approach to the computation of polynomial smoothing splines. The two constructions yield identical results for univariate smoothing, but provide building blocks satisfying different side conditions for multivariate smoothing with built-in ANOVA decompositions.

### 2.3.1 A Reproducing Kernel in $\mathcal{C}^{(m)}[0, 1]$

For  $f \in \mathcal{C}^{(m)}[0, 1]$ , the standard Taylor expansion gives

$$f(x) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!} f^{(\nu)}(0) + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} f^{(m)}(u) du, \quad (2.6)$$

where  $(\cdot)_+ = \max(0, \cdot)$ . With an inner product

$$(f, g) = \sum_{\nu=0}^{m-1} f^{(\nu)}(0) g^{(\nu)}(0) + \int_0^1 f^{(m)} g^{(m)} dx, \quad (2.7)$$

it can be shown that the representer of evaluation  $[x](\cdot)$  is

$$R_x(y) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!} \frac{y^\nu}{\nu!} + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} \frac{(y-u)_+^{m-1}}{(m-1)!} du. \quad (2.8)$$

To see this, note that  $R_x^{(\nu)}(0) = x^\nu/\nu!$ ,  $\nu = 0, \dots, m-1$ , and that  $R_x^{(m)}(y) = (x-y)_+^{m-1}/(m-1)!$ . Plugging these into (2.7) with  $g = R_x$ , one obtains the right-hand side of (2.6), so  $(R_x, f) = f(x)$ .

The two terms of the reproducing kernel  $R(x, y) = R_x(y)$ ,

$$R_0(x, y) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!} \frac{y^\nu}{\nu!}, \quad (2.9)$$

and

$$R_1(x, y) = \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} \frac{(y-u)_+^{m-1}}{(m-1)!} du, \quad (2.10)$$

are both non-negative definite themselves, and it is also easy to verify the other conditions of Theorem 2.5. To  $R_0$  there corresponds the space of polynomials  $\mathcal{H}_0 = \{f : f^{(m)} = 0\}$  with an inner product  $(f, g)_0 = \sum_{\nu=0}^{m-1} f^{(\nu)}(0)g^{(\nu)}(0)$ , and to  $R_1$  there corresponds the orthogonal complement of  $\mathcal{H}_0$ ,

$$\mathcal{H}_1 = \{f : f^{(\nu)}(0) = 0, \nu = 0, \dots, m-1, \int_0^1 (f^{(m)})^2 dx < \infty\}, \quad (2.11)$$

with an inner product  $(f, g)_1 = \int_0^1 f^{(m)}g^{(m)}dx$ . The space  $\mathcal{H}_0$  can be further decomposed into the tensor sum of  $m$  subspaces of monomials  $\{f : f \propto (\cdot)^\nu\}$  with inner products  $f^{(\nu)}(0)g^{(\nu)}(0)$  and reproducing kernels  $(x^\nu/\nu!)(y^\nu/\nu!)$ ,  $\nu = 0, \dots, m-1$ .

Setting  $m = 1$ , one has  $R_0(x, y) = 1$  and

$$R_1(x, y) = \int_0^1 I_{[u < x]} I_{[u < y]} du = x \wedge y, \quad (2.12)$$

where  $x \wedge y = \min(x, y)$ . This setting is useful for the computation of a linear smoothing spline, the minimizer of

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda \int_0^1 \dot{\eta}^2 dx. \quad (2.13)$$

Setting  $m = 2$ , one has  $R_0(x, y) = 1 + xy$  and

$$\begin{aligned} R_1(x, y) &= \int_0^1 (x-u)_+(y-u)_+ du \\ &= (x \wedge y)^2 (3(x \vee y) - (x \wedge y)) / 6, \end{aligned} \quad (2.14)$$

where  $x \vee y = \max(x, y)$ . The latter formula can be used in the computation of a cubic smoothing spline.

For  $m = 1$ , the tensor sum decomposition characterized by  $R = R_0 + R_1 = [1] + [x \wedge y]$  naturally defines a one-way ANOVA decomposition with an averaging operator  $Af = f(0)$ , where the corresponding  $\mathcal{H}_0$  spans the “mean” space and  $\mathcal{H}_1$  spans the “contrast” space; see §1.3.1 for discussions on ANOVA decomposition and averaging operator.

For  $m = 2$ , the same ANOVA decomposition is characterized by the kernel decomposition

$$R = R_{00} + [R_{01} + R_1] = [1] + [xy + \{(x \wedge y)^2(3(x \vee y) - (x \wedge y))/6\}],$$

where  $R_0 = 1 + xy$  is further decomposed into the sum of  $R_{00} = 1$  and  $R_{01} = xy$ . The kernel  $R_{00}$  generates the “mean” space and the kernels  $R_{01}$  and  $R_1$  together generate the “contrast” space, with  $R_{01}$  contributing to the “parametric contrast” and  $R_1$  to the “nonparametric contrast.”

### 2.3.2 Computation of Polynomial Smoothing Splines

Given the sampling points  $x_i$ ,  $i = 1, \dots, n$  in (2.5) and noting that the space  $\{f : f = \sum_{i=1}^n \alpha_i R_1(x_i, \cdot)\}$  is a closed linear subspace of  $\mathcal{H}_1$  given in (2.11), one may write  $\eta \in \mathcal{C}^{(m)}[0, 1]$  as

$$\eta(x) = \sum_{\nu=0}^{m-1} d_\nu \frac{x^\nu}{\nu!} + \sum_{i=1}^n c_i R_1(x_i, x) + \rho(x), \quad (2.15)$$

where  $c_i$  and  $d_\nu$  are real coefficients,  $R_1$  is given in (2.10), and

$$\rho \in \mathcal{H}_1 \ominus \{f : f = \sum_{i=1}^n c_i R_1(x_i, \cdot)\}.$$

By orthogonality,  $\rho(x_i) = (R_1(x_i, \cdot), \rho) = 0$ ,  $i = 1, \dots, n$ . Denoting by  $S$  the  $n \times m$  matrix with the  $(i, \nu)$ th entry  $x_i^\nu/\nu!$  and by  $Q$  the  $n \times n$  matrix with the  $(i, j)$ th entry  $R_1(x_i, x_j)$ , (2.5) can be written as

$$(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c} + n\lambda(\rho, \rho), \quad (2.16)$$

where the fact that  $\int_0^1 R_1^{(m)}(x_i, x) R_1^{(m)}(x_j, x) dx = R_1(x_i, x_j)$  is used. Note that  $\rho$  only appears in the third term in (2.16), which is minimized at  $\rho = 0$ . Hence, a polynomial smoothing spline resides in a space

$$\mathcal{H}_0 \oplus \{f : f = \sum_{i=1}^n c_i R_1(x_i, \cdot)\},$$

of finite dimension, and so can be computed via the minimization of the first two terms of (2.16) with respect to  $\mathbf{c}$  and  $\mathbf{d}$ .

In this approach to the computation of polynomial smoothing splines, one needs the reproducing kernel  $R_1$  that corresponds to a space  $\mathcal{H}_1$  in which the roughness penalty  $\int_0^1 (f^{(m)})^2 dx$  is a full square norm, plus a basis that spans the null space of the penalty.

### 2.3.3 Another Reproducing Kernel in $\mathcal{C}^{(m)}[0, 1]$

The bilinear form  $\int_0^1 f^{(m)} g^{(m)} dx$  is a semi-inner-product in  $\mathcal{C}^{(m)}[0, 1]$ , which can be augmented to a full inner product by the addition of an inner product in its null space, the space  $\{f : f^{(m)} = 0\}$  of polynomials up to order  $m - 1$ . In §2.3.1, we used  $\sum_{\nu=0}^{m-1} f^{(\nu)}(0)g^{(\nu)}(0)$  as the inner product in  $\{f : f^{(m)} = 0\}$ . In this section, we will use a different inner product,  $\sum_{\nu=0}^{m-1} (\int_0^1 f^{(\nu)} dx)(\int_0^1 g^{(\nu)} dx)$ , in  $\{f : f^{(m)} = 0\}$ , and derive the reproducing kernel associated with

$$(f, g) = \sum_{\nu=0}^{m-1} \left( \int_0^1 f^{(\nu)} dx \right) \left( \int_0^1 g^{(\nu)} dx \right) + \int_0^1 f^{(m)} g^{(m)} dx, \quad (2.17)$$

which defines an inner product different from that in (2.7).

The sought-after reproducing kernel can most conveniently be expressed in terms of the functions

$$k_r(x) = - \left( \sum_{\mu=-\infty}^{-1} + \sum_{\mu=1}^{\infty} \right) \frac{\exp(2\pi \mathbf{i} \mu x)}{(2\pi \mathbf{i} \mu)^r}, \quad r = 1, 2, \dots, \quad (2.18)$$

where  $\mathbf{i} = \sqrt{-1}$ . It is easy to verify that for  $r > 1$ ,  $k_r$  is well defined and continuous on the real line, and for  $r = 1$ , it is well defined and continuous at noninteger points; see Problem 2.9(a). It is also easy to verify that  $k_r(x)$  is real-valued and is periodic with period 1; see Problem 2.9(b). It can be seen that  $k_r^{(p)} = k_{r-p}$ ,  $p = 1, \dots, r-2$  and that  $k_r^{(r-1)}(x) = k_1(x)$  for  $x$  not an integer. It is known that  $k_1(x) = x - 0.5$  on  $(0, 1)$  (Problem 2.9(c)), and we define  $k_0 = 1$ . The  $k_r$  functions are actually scaled Bernoulli polynomials,  $k_r(x) = B_r(x)/r!$ ; see Abramowitz and Stegun (1964, Chap. 23) for a comprehensive list of results concerning the Bernoulli polynomials  $B_r(x)$ .

From the properties listed above, it is easy to verify that  $\int_0^1 k_\mu^{(\nu)} dx = \delta_{\mu, \nu}$ ,  $\mu, \nu = 0, \dots, m-1$ , where  $\delta_{\mu, \nu}$  is the Kronecker delta. It then follows that  $k_\nu$ ,  $\nu = 0, \dots, m-1$  form an orthonormal basis of  $\mathcal{H}_0 = \{f : f^{(m)} = 0\}$  under the inner product  $(f, g)_0 = \sum_{\nu=0}^{m-1} (\int_0^1 f^{(\nu)} dx)(\int_0^1 g^{(\nu)} dx)$  and that

$$R_0(x, y) = \sum_{\nu=0}^{m-1} k_\nu(x) k_\nu(y) \quad (2.19)$$

is the reproducing kernel in  $\mathcal{H}_0$ ; see Problem 2.5(c) for the definition of orthonormal basis. In fact,  $\mathcal{H}_0$  can be further decomposed into the tensor sum of  $m$  subspaces  $\{f : f \propto k_\nu\}$  with inner products  $(\int_0^1 f^{(\nu)} dx)(\int_0^1 g^{(\nu)} dx)$  and reproducing kernels  $k_\nu(x)k_\nu(y)$ ,  $\nu = 0, \dots, m-1$ , respectively.

We now show that in the space

$$\mathcal{H}_1 = \{f : \int_0^1 f^{(\nu)} dx = 0, \nu = 0, \dots, m-1, f^{(m)} \in \mathcal{L}_2[0, 1]\} \quad (2.20)$$

with a square norm  $(f, g)_1 = \int_0^1 f^{(m)} g^{(m)} dx$ , the function

$$R_x(y) = k_m(x)k_m(y) + (-1)^{m-1}k_{2m}(x-y) \quad (2.21)$$

is the representer of evaluation  $[x](\cdot)$ . From the properties of  $k_r$ , it is easy to verify that  $\int_0^1 R_x^{(\nu)}(y)dy = 0$ ,  $\nu = 0, \dots, m-1$ , and that  $R_x^{(m)}(y) = k_m(x) - k_m(x-y) \in \mathcal{L}_2[0, 1]$ , so  $R_x \in \mathcal{H}_1$  for  $\mathcal{H}_1$  given in (2.20). Integrating by parts, and using the periodicity of  $k_r$ ,  $r > 1$ , and the fact that  $\int_0^1 f^{(\nu)} dx = 0$ ,  $\nu = 0, \dots, m-1$ , one can show that, for  $m > 1$ ,

$$\begin{aligned} (R_x, f)_1 &= \int_0^1 (k_m(x) - k_m(x-y)) f^{(m)}(y) dy \\ &= - \int_0^1 k_{m-1}(x-y) \dot{f}^{(m-1)}(y) dy \\ &= \dots = - \int_0^1 k_1(x-y) \dot{f}(y) dy; \end{aligned} \quad (2.22)$$

see Problem 2.10. Now, since

$$k_1(x-y) = \begin{cases} x-y-0.5 = k_1(x) - y, & y \in (0, x), \\ (1+x-y)-0.5 = k_1(x) - y + 1, & y \in (x, 1), \end{cases}$$

straightforward calculation yields

$$\begin{aligned} & - \int_0^1 k_1(x-y) \dot{f}(y) dy \\ &= - \int_0^1 k_1(x) \dot{f}(y) dy + \int_0^1 y \dot{f}(y) dy - \int_x^1 \dot{f}(y) dy \\ &= 0 + f(1) - (f(1) - f(x)) = f(x). \end{aligned}$$

The result holds for  $m = 1$  via direct calculation. This proves that

$$R_1(x, y) = k_m(x)k_m(y) + (-1)^{m-1}k_{2m}(x-y) \quad (2.23)$$

is the reproducing kernel of  $\mathcal{H}_1$  given in (2.20).

Obviously,  $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$ , so by the converse of Theorem 2.5,  $\mathcal{C}^{(m)}[0, 1] = \mathcal{H}_0 \oplus \mathcal{H}_1$  has the reproducing kernel  $R = R_0 + R_1$ . The identity

$$f(x) = \sum_{\nu=0}^{m-1} k_\nu(x) \int_0^1 f^{(\nu)}(y) dy + \int_0^1 (k_m(x) - k_m(x-y)) f^{(m)}(y) dy, \quad (2.24)$$

$\forall f \in \mathcal{C}^{(m)}[0, 1]$ , may be called a generalized Taylor expansion, where the scaled Bernoulli polynomials  $k_\nu(x)$  play the role of the scaled monomials  $x^\nu/\nu!$  in the standard Taylor expansion of (2.6). The standard Taylor expansion is asymmetric with respect to the domain  $[0, 1]$ , in the sense that



a swapping of the two ends 0 and 1 would change its composition entirely, whereas the generalized Taylor expansion of (2.24) is symmetric with respect to the domain.

The computation of polynomial smoothing splines as outlined in §2.3.2 can also be performed by using the  $R_1$  of (2.23) instead of that of (2.10). Also, one may use any basis  $\{\phi_\nu\}_{\nu=0}^{m-1}$  of the subspace  $\mathcal{H}_0$  in the place of  $\{x^\nu/\nu!\}_{\nu=0}^{m-1}$  in the expression of  $\eta$  given in (2.15). The coefficients  $c_i$  and  $d_\nu$  will be different when different  $\phi_\nu$  and  $R_1$  are used, but the function estimate

$$\eta(x) = \sum_{\nu=0}^{m-1} d_\nu \phi_\nu(x) + \sum_{i=1}^n c_i R_1(x_i, x)$$

will remain the same regardless of the choices of  $\phi_\nu$  and  $R_1$ .

When  $m = 1$ ,  $R_0(x, y) = 1$  and

$$R_1(x, y) = k_1(x)k_1(y) + k_2(x - y). \quad (2.25)$$

When  $m = 2$ ,  $R_0(x, y) = 1 + k_1(x)k_1(y)$  and

$$R_1(x, y) = k_2(x)k_2(y) - k_4(x - y). \quad (2.26)$$

The  $R_1$  in (2.25) and (2.26) can be used in the computation of linear and cubic smoothing splines in lieu of those in (2.12) and (2.14). To calculate  $R_1$  in (2.25) and (2.26), one has, on  $x \in [0, 1]$ ,

$$\begin{aligned} k_2(x) &= \frac{1}{2} \left( k_1^2(x) - \frac{1}{12} \right), \\ k_4(x) &= \frac{1}{24} \left( k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240} \right), \end{aligned} \quad (2.27)$$

where  $k_1(x) = x - 0.5$ ; see Problem 2.11. Note that  $k_2$  and  $k_4$  are symmetric with respect to 0.5 on  $[0, 1]$ , so for  $x \in [-1, 0]$ ,

$$k_2(x) = k_2(x + 1) = k_2(0.5 + (x + 0.5)) = k_2(0.5 - (x + 0.5)) = k_2(-x),$$

and likewise,  $k_4(x) = k_4(-x)$ . It then follows that  $k_2(x - y) = k_2(|x - y|)$  and  $k_4(x - y) = k_4(|x - y|)$ , for  $x, y \in [0, 1]$ .

For  $m = 1$ , the tensor sum decomposition characterized by  $R = R_0 + R_1 = [1] + [k_1(x)k_1(y) + k_2(x - y)]$  defines a one-way ANOVA decomposition with an averaging operator  $Af = \int_0^1 f dx$ , where the corresponding  $\mathcal{H}_0$  spans the “mean” space and  $\mathcal{H}_1$  spans the “contrast” space.

For  $m = 2$ , the same ANOVA decomposition is characterized by the kernel decomposition

$$R = R_{00} + [R_{01} + R_1] = [1] + [k_1(x)k_1(y) + \{k_2(x)k_2(y) - k_4(x - y)\}],$$

where  $R_0 = 1 + k_1(x)k_1(y)$  is further decomposed into the sum of  $R_{00} = 1$  and  $R_{01} = k_1(x)k_1(y)$ . The kernel  $R_{00}$  generates the “mean” space and

the kernels  $R_{01}$  and  $R_1$  together generate the “contrast” space, with  $R_{01}$  contributing to the “parametric contrast” and  $R_1$  to the “nonparametric contrast.”

## 2.4 Smoothing Splines on Product Domains

To incorporate the ANOVA decomposition introduced in §1.3.2 for the estimation of a multivariate function, one may construct a tensor product reproducing kernel Hilbert space. Given Theorem 2.3, the construction of the space can be done through the construction of the reproducing kernel, for which one uses reproducing kernels on the marginal domains. One-way ANOVA decompositions on the marginal domains naturally induce an ANOVA decomposition on the product domain.

We begin with some general discussion of tensor product reproducing kernel Hilbert spaces, where it is shown that the products of reproducing kernels on the marginal domains form reproducing kernels on the product domain. The construction is then illustrated with marginal domains  $\{1, \dots, K\}$  and  $[0, 1]$ , using the (marginal) reproducing kernels introduced in §§2.2 and 2.3.

### 2.4.1 Tensor Product Reproducing Kernel Hilbert Spaces

A convenient approach to the construction of reproducing kernel Hilbert spaces on a product domain  $\prod_{\gamma=1}^r \mathcal{X}_\gamma$  is by taking the tensor product of spaces constructed on the marginal domains  $\mathcal{X}_\gamma$ . The construction builds on the following theorem.

**Theorem 2.6** *For  $R_{(1)}(x_{(1)}, y_{(1)})$  non-negative definite on  $\mathcal{X}_1$  and  $R_{(2)}(x_{(2)}, y_{(2)})$  non-negative definite on  $\mathcal{X}_2$ ,  $R(x, y) = R_{(1)}(x_{(1)}, y_{(1)})R_{(2)}(x_{(2)}, y_{(2)})$  is non-negative definite on  $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ .*

*Proof:* It suffices to show that, for two non-negative definite matrices  $A$  and  $B$  of the same size, their entrywise product,  $A \circ B$ , is necessarily non-negative definite. By elementary matrix theory,  $A$  and  $B$  are non-negative definite if and only if there exist vectors  $a_i$  and  $b_j$  such that  $A = \sum_i a_i a_i^T$  and  $B = \sum_j b_j b_j^T$ . Now,

$$\begin{aligned} A \circ B &= \left( \sum_i a_i a_i^T \right) \circ \left( \sum_j b_j b_j^T \right) \\ &= \sum_{i,j} (a_i a_i^T) \circ (b_j b_j^T) = \sum_{i,j} (a_i \circ b_j)(a_i \circ b_j)^T, \end{aligned}$$

so  $A \circ B$  is non-negative definite.  $\square$

By Theorem 2.3, every non-negative definite function  $R$  on domain  $\mathcal{X}$  corresponds to a reproducing kernel Hilbert space with  $R$  as its reproducing

kernel. Given  $\mathcal{H}_{(1)}$  on  $\mathcal{X}_1$  with reproducing kernel  $R_{(1)}$  and  $\mathcal{H}_{(2)}$  on  $\mathcal{X}_2$  with reproducing kernel  $R_{(2)}$ ,  $R = R_{(1)}R_{(2)}$  is non-negative definite on  $\mathcal{X}_1 \times \mathcal{X}_2$  by Theorem 2.6. The reproducing kernel Hilbert space corresponding to such an  $R$  is called the **tensor product space** of  $\mathcal{H}_{(1)}$  and  $\mathcal{H}_{(2)}$ , and is denoted by  $\mathcal{H}_{(1)} \otimes \mathcal{H}_{(2)}$ . The operation extends to multiple-term products recursively.

Suppose one has reproducing kernel Hilbert spaces  $\mathcal{H}_{(\gamma)}$  on domains  $\mathcal{X}_\gamma$ ,  $\gamma = 1, \dots, \Gamma$ , respectively. Further, assume that the spaces have one-way ANOVA decompositions built in via the tensor sum decompositions  $\mathcal{H}_{(\gamma)} = \mathcal{H}_{0(\gamma)} \oplus \mathcal{H}_{1(\gamma)}$ , where  $\mathcal{H}_{0(\gamma)} = \{f : f \propto 1\}$  has a reproducing kernel  $R_{0(\gamma)} \propto 1$  and  $\mathcal{H}_{1(\gamma)}$  has a reproducing kernel  $R_{1(\gamma)}$  satisfying side conditions  $A_\gamma R_{1(\gamma)}(x_{(\gamma)}, \cdot) = 0$ ,  $\forall x_{(\gamma)} \in \mathcal{X}_\gamma$ , where  $A_\gamma$  are the averaging operators defining the one-way ANOVA decompositions on  $\mathcal{X}_\gamma$ . The tensor product space  $\mathcal{H} = \otimes_{\gamma=1}^\Gamma \mathcal{H}_{(\gamma)}$  has a tensor sum decomposition

$$\mathcal{H} = \bigotimes_{\gamma=1}^\Gamma (\mathcal{H}_{0(\gamma)} \oplus \mathcal{H}_{1(\gamma)}) = \bigoplus_{\mathcal{S}} \left\{ \left( \bigotimes_{\gamma \in \mathcal{S}} \mathcal{H}_{1(\gamma)} \right) \otimes \left( \bigotimes_{\gamma \notin \mathcal{S}} \mathcal{H}_{0(\gamma)} \right) \right\} = \bigoplus_{\mathcal{S}} \mathcal{H}_{\mathcal{S}}, \quad (2.28)$$

which parallels (1.7) on page 7, where the summation is over all subsets  $\mathcal{S} \subseteq \{1, \dots, \Gamma\}$ . The term  $\mathcal{H}_{\mathcal{S}}$  has a reproducing kernel  $R_{\mathcal{S}} \propto \prod_{\gamma \in \mathcal{S}} R_{1(\gamma)}$ , and the projection of  $f \in \mathcal{H}$  in  $\mathcal{H}_{\mathcal{S}}$  is the  $f_{\mathcal{S}}$  appearing in (1.7). The minimizer of  $L(f) + (\lambda/2)J(f)$  in a tensor product reproducing kernel Hilbert space is called a **tensor product smoothing spline**. Examples of the construction follow.

### 2.4.2 Reproducing Kernel Hilbert Spaces on $\{1, \dots, K\}^2$

Set  $A_\gamma f = \sum_{x_{(\gamma)}=1}^{K_\gamma} f(x)/K_\gamma$  on discrete domains  $\mathcal{X}_\gamma = \{1, \dots, K_\gamma\}$ ,  $\gamma = 1, 2$ . The marginal reproducing kernels that define the one-way ANOVA decomposition on  $\mathcal{X}_\gamma$  can be taken as  $R_{0(\gamma)}(x_{(\gamma)}, y_{(\gamma)}) = 1/K_\gamma$  and

$$R_{1(\gamma)}(x_{(\gamma)}, y_{(\gamma)}) = I_{[x_{(\gamma)}=y_{(\gamma)}]} - 1/K_\gamma,$$

$\gamma = 1, 2$ , as given in §2.2.

A function on  $\{1, \dots, K_1\} \times \{1, \dots, K_2\}$  can be written as a vector of length  $K_1 K_2$ ,

$$f = (f(1, 1), \dots, f(1, K_2), \dots, f(K_1, 1), \dots, f(K_1, K_2))^T,$$

and a reproducing kernel as a  $(K_1 K_2) \times (K_1 K_2)$  matrix. Using matrix notation, the products of the marginal reproducing kernels  $R_{0(\gamma)}$  and  $R_{1(\gamma)}$  given above and the subspaces they correspond to are listed in Table 2.1, where  $\mathbf{1}_K$  is of length  $K$ ,  $I_K$  is of size  $K \times K$ , and, as a matrix operator,  $\otimes$  denotes the Kronecker product of matrices. The corresponding inner products are defined by the Moore-Penrose inverses of these matrices, which

TABLE 2.1. Product reproducing kernels on  $\{1, \dots, K_1\} \times \{1, \dots, K_2\}$ .

Subspace	Reproducing kernel
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{0(2)}$	$(\mathbf{1}_{K_1} \mathbf{1}_{K_1}^T / K_1) \otimes (\mathbf{1}_{K_2} \mathbf{1}_{K_2}^T / K_2)$
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{1(2)}$	$(\mathbf{1}_{K_1} \mathbf{1}_{K_1}^T / K_1) \otimes (I_{K_2} - \mathbf{1}_{K_2} \mathbf{1}_{K_2}^T / K_2)$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{0(2)}$	$(I_{K_1} - \mathbf{1}_{K_1} \mathbf{1}_{K_1}^T / K_1) \otimes (\mathbf{1}_{K_2} \mathbf{1}_{K_2}^T / K_2)$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{1(2)}$	$(I_{K_1} - \mathbf{1}_{K_1} \mathbf{1}_{K_1}^T / K_1) \otimes (I_{K_2} - \mathbf{1}_{K_2} \mathbf{1}_{K_2}^T / K_2)$

are themselves because they are idempotent. The decomposition of (2.28) is seen to be

$$\begin{aligned}
\mathcal{H} &= (\mathcal{H}_{0(1)} \oplus \mathcal{H}_{1(1)}) \otimes (\mathcal{H}_{0(2)} \oplus \mathcal{H}_{1(2)}) \\
&= (\mathcal{H}_{0(1)} \otimes \mathcal{H}_{0(2)}) \oplus (\mathcal{H}_{1(1)} \otimes \mathcal{H}_{0(2)}) \\
&\quad \oplus (\mathcal{H}_{0(1)} \otimes \mathcal{H}_{1(2)}) \oplus (\mathcal{H}_{1(1)} \otimes \mathcal{H}_{1(2)}) \\
&= \mathcal{H}_{\{\}} \oplus \mathcal{H}_{\{1\}} \oplus \mathcal{H}_{\{2\}} \oplus \mathcal{H}_{\{1,2\}},
\end{aligned} \tag{2.29}$$

where  $\mathcal{H}_{\{\}}$  spans the constant,  $\mathcal{H}_{\{1\}}$  spans the  $x_{(1)}$  main effect,  $\mathcal{H}_{\{2\}}$  spans the  $x_{(2)}$  main effect, and  $\mathcal{H}_{\{1,2\}}$  spans the interaction.

If one would like to use the averaging operator  $Af = f(1)$  on a marginal domain  $\{1, \dots, K\}$ , the  $K$ -dimensional vector space may be decomposed alternatively as

$$\mathcal{H}_0 \oplus \mathcal{H}_1 = \{f : f(1) = \dots = f(K)\} \oplus \{f : f(1) = 0\},$$

with the reproducing kernels given by  $R_0 = 1$  and  $R_1(x, y) = I_{[x=y \neq 1]}$ ; see Problem 2.8.

### 2.4.3 Reproducing Kernel Hilbert Spaces on $[0, 1]^2$

Set  $Af = \int_0^1 f dx$  on  $[0, 1]$ . The tensor product reproducing kernel Hilbert spaces on  $[0, 1]^2$  can be constructed using the reproducing kernels (2.19) and (2.23) derived in §2.3.3.

**Example 2.4 (Tensor product linear spline)** Setting  $m = 1$  in §2.3.3, one has

$$\begin{aligned}
\{f : \dot{f} \in \mathcal{L}_2[0, 1]\} &= \{f : f \propto 1\} \oplus \{f : \int_0^1 f dx = 0, \dot{f} \in \mathcal{L}_2[0, 1]\} \\
&= \mathcal{H}_0 \oplus \mathcal{H}_1,
\end{aligned}$$

with reproducing kernels  $R_0(x, y) = 1$  and  $R_1(x, y) = k_1(x)k_1(y) + k_2(x - y)$ . This marginal space can be used on both axes to construct a tensor product reproducing kernel Hilbert space with the structure of (2.28), with averaging operators  $A_\gamma f = \int_0^1 f dx_{(\gamma)}$ ,  $\gamma = 1, 2$ . The reproducing kernels and the corresponding inner products in the subspaces are listed in Table 2.2.  $\square$

TABLE 2.2. Reproducing Kernels and Inner Products in Example 2.4.

Subspace	Reproducing Kernel	Inner Product
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{0(2)}$	1	$(\int_0^1 \int_0^1 f)(\int_0^1 \int_0^1 g)$
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{1(2)}$	$k_1(x_{(2)})k_1(y_{(2)}) + k_2(x_{(2)} - y_{(2)})$	$\int_0^1 (\int_0^1 \dot{f}_{(2)} dx_{(1)}) (\int_0^1 \dot{g}_{(2)} dx_{(1)}) dx_{(2)}$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{0(2)}$	$k_1(x_{(1)})k_1(y_{(1)}) + k_2(x_{(1)} - y_{(1)})$	$\int_0^1 (\int_0^1 \dot{f}_{(1)} dx_{(2)}) (\int_0^1 \dot{g}_{(1)} dx_{(2)}) dx_{(1)}$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{1(2)}$	$[k_1(x_{(1)})k_1(y_{(1)}) + k_2(x_{(1)} - y_{(1)})][k_1(x_{(2)})k_1(y_{(2)}) + k_2(x_{(2)} - y_{(2)})]$	$\int_0^1 \int_0^1 \ddot{f}_{(12)} \ddot{g}_{(12)}$

TABLE 2.3. Reproducing Kernels and Inner Products in Example 2.5.

Subspace	Reproducing Kernel	Inner Product
$\mathcal{H}_{00(1)} \otimes \mathcal{H}_{00(2)}$	1	$(\int_0^1 \int_0^1 f)(\int_0^1 \int_0^1 g)$
$\mathcal{H}_{01(1)} \otimes \mathcal{H}_{00(2)}$	$k_1(x_{(1)})k_1(y_{(1)})$	$(\int_0^1 \int_0^1 \dot{f}_{(1)}) (\int_0^1 \int_0^1 \dot{g}_{(1)})$
$\mathcal{H}_{01(1)} \otimes \mathcal{H}_{01(2)}$	$k_1(x_{(1)})k_1(y_{(1)})k_1(x_{(2)})k_1(y_{(2)})$	$(\int_0^1 \int_0^1 \ddot{f}_{(12)}) (\int_0^1 \int_0^1 \ddot{g}_{(12)})$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{00(2)}$	$k_2(x_{(1)})k_2(y_{(1)}) - k_4(x_{(1)} - y_{(1)})$	$\int_0^1 (\int_0^1 \dot{f}_{(11)} dx_{(2)}) (\int_0^1 \ddot{g}_{(11)} dx_{(2)}) dx_{(1)}$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{01(2)}$	$[k_2(x_{(1)})k_2(y_{(1)}) - k_4(x_{(1)} - y_{(1)})]k_1(x_{(2)})k_1(y_{(2)})$	$\int_0^1 (\int_0^1 \dot{f}_{(112)} dx_{(2)}) (\int_0^1 \ddot{g}_{(112)}^{(3)} dx_{(2)}) dx_{(1)}$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{1(2)}$	$[k_2(x_{(1)})k_2(y_{(1)}) - k_4(x_{(1)} - y_{(1)})][k_2(x_{(2)})k_2(y_{(2)}) - k_4(x_{(2)} - y_{(2)})]$	$\int_0^1 \int_0^1 \ddot{f}_{(1122)}^{(4)} \ddot{g}_{(1122)}^{(4)}$

**Example 2.5 (Tensor product cubic spline)** Setting  $m = 2$  in §2.3.3, one has

$$\begin{aligned} \{f : \ddot{f} \in \mathcal{L}_2[0, 1]\} &= \{f : f \propto 1\} \oplus \{f : f \propto k_1\} \\ &\quad \oplus \{f : \int_0^1 f dx = \int_0^1 \dot{f} dx = 0, \ddot{f} \in \mathcal{L}_2[0, 1]\} \\ &= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1, \end{aligned}$$

where  $\mathcal{H}_{01} \oplus \mathcal{H}_1$  forms the contrast in a one-way ANOVA decomposition with an averaging operator  $Af = \int_0^1 f dx$ . The corresponding reproducing kernels are  $R_{00}(x, y) = 1$ ,  $R_{01}(x, y) = k_1(x)k_1(y)$ , and  $R_1(x, y) = k_2(x)k_2(y) - k_4(x - y)$ . Note that  $\int_0^1 R_{01}(x, y) dy = \int_0^1 R_1(x, y) dy = 0$ ,  $\forall x \in [0, 1]$ . Using this space on both marginal domains, one can construct a tensor product space with nine tensor sum terms. The subspace  $\mathcal{H}_{00(1)} \otimes \mathcal{H}_{00(2)}$  spans the constant term in (1.7) on page 7, the subspaces  $\mathcal{H}_{00(1)} \otimes (\mathcal{H}_{01(2)} \oplus \mathcal{H}_{1(2)})$  and  $(\mathcal{H}_{01(1)} \oplus \mathcal{H}_{1(1)}) \otimes \mathcal{H}_{00(2)}$  span the main effects, and the subspace  $(\mathcal{H}_{01(1)} \oplus \mathcal{H}_{1(1)}) \otimes (\mathcal{H}_{01(2)} \oplus \mathcal{H}_{1(2)})$  spans the interaction. The reproducing kernels and the corresponding inner products in some of the subspaces are listed in Table 2.3. The separation of  $\mathcal{H}_{01}$  and  $\mathcal{H}_1$  is intended to facilitate adequate numerical treatment of the different components; it is not needed for the characterization of the ANOVA decomposition in (2.28).  $\square$

For the averaging operator  $Af = f(0)$ , similar tensor product reproducing kernel Hilbert spaces can be constructed using the marginal spaces described in §2.3.1; details are to be worked out in Problem 2.13. Note that it is not necessary to use the same marginal space on both axes. Actually, the choice of the order  $m$  and that of the averaging operator  $Af$  on different axes are unrelated to each other. Although the reproducing kernels of §§2.3.1 and 2.3.3 lead to identical polynomial smoothing splines for univariate smoothing on  $[0, 1]$ , they do yield different tensor product smoothing splines on  $[0, 1]^2$ , as their respective roughness penalties are different.

#### 2.4.4 Reproducing Kernel Hilbert Spaces on $\{1, \dots, K\} \times [0, 1]$

Setting  $A_1 f = \sum_{x_{(1)}=1}^K f(x)/K$  on  $\mathcal{X}_1 = \{1, \dots, K\}$  and  $A_2 f = \int_0^1 f dx_{(2)}$  on  $\mathcal{X}_2 = [0, 1]$ , tensor product spaces with the structure of (2.28) built in can be constructed using the marginal spaces used in §§2.4.2 and 2.4.3.

**Example 2.6** One construction of a tensor product space is by using  $R_{0(1)}(x_{(1)}, y_{(1)}) = 1/K$  and  $R_{1(1)}(x_{(1)}, y_{(1)}) = I_{[x_{(1)}=y_{(1)}]} - 1/K$  on  $\mathcal{X}_1$  and  $R_{0(2)}(x_{(2)}, y_{(2)}) = 1$  and  $R_{1(2)}(x_{(2)}, y_{(2)}) = k_1(x_{(2)})k_1(y_{(2)}) + k_2(x_{(2)} - y_{(2)})$  on  $\mathcal{X}_2$ . The reproducing kernels and the corresponding inner products in the subspaces are listed in Table 2.4.  $\square$

**Example 2.7** Using  $R_{0(1)} = 1/K$  and  $R_{1(1)} = I_{[x_{(1)}=y_{(1)}]} - 1/K$  on  $\mathcal{X}_1$  and  $R_{00(2)} = 1$ ,  $R_{01(2)} = k_1(x_{(2)})k_1(y_{(2)})$ , and  $R_{1(2)} = k_2(x_{(2)})k_2(y_{(2)}) - k_4(x_{(2)} - y_{(2)})$  on  $\mathcal{X}_2$ , one can construct a tensor product space with six tensor sum terms. The subspace  $\mathcal{H}_{0(1)} \otimes \mathcal{H}_{00(2)}$  spans the constant,  $\mathcal{H}_{0(1)} \otimes (\mathcal{H}_{01(2)} \oplus \mathcal{H}_{1(2)})$  and  $\mathcal{H}_{1(1)} \otimes \mathcal{H}_{00(2)}$  span the main effects, and  $\mathcal{H}_{1(1)} \otimes (\mathcal{H}_{01(2)} \oplus \mathcal{H}_{1(2)})$  spans the interaction. The reproducing kernels and the corresponding inner products in the subspaces are listed in Table 2.5.  $\square$

#### 2.4.5 Multiple-Term Reproducing Kernel Hilbert Spaces: General Form

The examples of tensor product reproducing kernel Hilbert spaces on product domains presented above all contain multiple tensor sum terms. In general, a multiple-term reproducing kernel Hilbert space can be written as  $\mathcal{H} = \oplus_{\beta} \mathcal{H}_{\beta}$ , where  $\beta$  is a generic index, with subspaces  $\mathcal{H}_{\beta}$  having inner products  $(f_{\beta}, g_{\beta})_{\beta}$  and reproducing kernels  $R_{\beta}$ , where  $f_{\beta}$  is the projection of  $f$  in  $\mathcal{H}_{\beta}$ . It is often convenient to write  $(f, g)_{\beta}$  for  $(f_{\beta}, g_{\beta})_{\beta}$ , which can be formally defined as a semi-inner-product in  $\mathcal{H}$  satisfying  $(f - f_{\beta}, f - f_{\beta})_{\beta} = 0$ .

The subspaces  $\mathcal{H}_{\beta}$  are independent modules, and the within-module metrics implied by the inner products  $(f_{\beta}, g_{\beta})_{\beta}$  are not necessarily comparable between the modules. Allowing for intermodule rescaling of the metrics, an inner product in  $\mathcal{H}$  can be specified via

$$J(f, g) = \sum_{\beta} \theta_{\beta}^{-1} (f_{\beta}, g_{\beta})_{\beta}, \quad (2.30)$$

where  $\theta_{\beta} \in (0, \infty)$  are tunable parameters. The reproducing kernel associated with (2.30) is  $R_J = \sum_{\beta} \theta_{\beta} R_{\beta}$ , as

$$J(R_J(x, \cdot), f) = \sum_{\beta} \theta_{\beta}^{-1} (\theta_{\beta} R_{\beta}(x, \cdot), f_{\beta})_{\beta} = \sum_{\beta} f_{\beta}(x) = f(x).$$

When some of the  $\theta_{\beta}$  are set to  $\infty$  in (2.30),  $J(f, g)$  defines a semi-inner-product in  $\mathcal{H} = \oplus_{\beta} \mathcal{H}_{\beta}$ . Such a semi-inner-product may be used to specify  $J(f) = J(f, f)$  for use in  $L(f) + (\lambda/2)J(f)$ . Subspaces not contributing to  $J(f)$  form the null space of  $J(f)$ ,  $\mathcal{N}_J = \{f : J(f) = 0\}$ . Subspaces contributing to  $J(f)$  form the space  $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$ , in which  $J(f, g)$  is a full inner product.

Observing  $Y_i = \eta(x_i) + \epsilon_i$ , where  $x_i \in \mathcal{X}$  is a product domain and  $\epsilon_i \sim N(0, \sigma^2)$ , one may estimate  $\eta$  via the minimization of

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda J(\eta), \quad (2.31)$$

TABLE 2.4. Reproducing kernels and inner products in Example 2.6.

Subspace	Reproducing kernel	Inner product
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{0(2)}$	$1/K$	$(\sum_{x_{(1)}=1}^K \int_0^1 f)(\sum_{x_{(1)}=1}^K \int_0^1 g)/K$
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{1(2)}$	$[k_1(x_{(2)})k_1(y_{(2)}) + k_2(x_{(2)} - y_{(2)})]/K$	$\int_0^1 (\sum_{x_{(1)}=1}^K \dot{f}_{(2)})(\sum_{x_{(1)}=1}^K \dot{g}_{(2)})/K$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{0(2)}$	$I_{[x_{(1)}=y_{(1)}]} - 1/K$	$\sum_{x_{(1)}=1}^K (\int_0^1 (I - A_1)f)(\int_0^1 (I - A_1)g)$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{1(2)}$	$(I_{[x_{(1)}=y_{(1)}]} - 1/K)[k_1(x_{(2)})k_1(y_{(2)}) + k_2(x_{(2)} - y_{(2)})]$	$\int_0^1 \sum_{x_{(1)}=1}^K (I - A_1)\dot{f}_{(2)}(I - A_1)\dot{g}_{(2)}$

TABLE 2.5. Reproducing kernels and inner products in Example 2.7.

Subspace	Reproducing kernel	Inner product
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{00(2)}$	$1/K$	$(\sum_{x_{(1)}=1}^K \int_0^1 f)(\sum_{x_{(1)}=1}^K \int_0^1 g)/K$
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{01(2)}$	$k_1(x_{(2)})k_1(y_{(2)})/K$	$(\sum_{x_{(1)}=1}^K \int_0^1 \dot{f}_{(2)})(\sum_{x_{(1)}=1}^K \int_0^1 \dot{g}_{(2)})/K$
$\mathcal{H}_{0(1)} \otimes \mathcal{H}_{1(2)}$	$[k_2(x_{(2)})k_2(y_{(2)}) - k_4(x_{(2)} - y_{(2)})]/K$	$\int_0^1 (\sum_{x_{(1)}=1}^K \ddot{f}_{(2)})(\sum_{x_{(1)}=1}^K \ddot{g}_{(2)})/K$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{00(2)}$	$I_{[x_{(1)}=y_{(1)}]} - 1/K$	$\sum_{x_{(1)}=1}^K (\int_0^1 (I - A_1)f)(\int_0^1 (I - A_1)g)$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{01(2)}$	$(I_{[x_{(1)}=y_{(1)}]} - 1/K)k_1(x_{(2)})k_1(y_{(2)})$	$\sum_{x_{(1)}=1}^K (\int_0^1 (I - A_1)\dot{f}_{(2)})(\int_0^1 (I - A_1)\dot{g}_{(2)})$
$\mathcal{H}_{1(1)} \otimes \mathcal{H}_{1(2)}$	$(I_{[x_{(1)}=y_{(1)}]} - 1/K)[k_2(x_{(2)})k_2(y_{(2)}) + k_4(x_{(2)} - y_{(2)})]$	$\int_0^1 \sum_{x_{(1)}=1}^K (I - A_1)\ddot{f}_{(2)}(I - A_1)\ddot{g}_{(2)}$



where  $J(f) = J(f, f)$  is as given above. The minimizer of (2.31) defines a smoothing spline on  $\mathcal{X}$ . The computation strategy outlined in §2.3.2 readily applies here, with the subspaces  $\mathcal{H}_0$  and  $\mathcal{H}_1$  in §2.3.2 replaced by  $\mathcal{N}_J$  and  $\mathcal{H}_J$ , respectively.

When some of the  $\theta_\beta$  are set to 0 in  $J(f) = J(f, f)$ , the corresponding  $f_\beta$  are not allowed in the estimate. One simply eliminates the corresponding  $\mathcal{H}_\beta$  from the tensor sum.

Note that for the computation of a smoothing spline, all that one needs are a basis of  $\mathcal{N}_J$  and the reproducing kernel  $R_J$  associated with  $J(f)$  in  $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$ . In particular, the explicit form of  $J(f)$  is *not* needed.

**Example 2.8** Consider the construction of Example 2.5 on  $\mathcal{X} = [0, 1]^2$ . Denote  $\mathcal{H}_{\nu,\mu} = \mathcal{H}_{\nu(1)} \otimes \mathcal{H}_{\mu(2)}$ ,  $\nu, \mu = 00, 01, 1$ , with inner products  $(f, g)_{\nu,\mu}$  and reproducing kernels  $R_{\nu,\mu} = R_{\nu(1)}R_{\mu(2)}$ . One may set

$$\begin{aligned} J(f, g) = & \theta_{1,00}^{-1}(f, g)_{1,00} + \theta_{1,01}^{-1}(f, g)_{1,01} \\ & + \theta_{00,1}^{-1}(f, g)_{00,1} + \theta_{01,1}^{-1}(f, g)_{01,1} + \theta_{1,1}^{-1}(f, g)_{1,1} \end{aligned}$$

and minimize (2.31) in  $\mathcal{H} = \oplus_{\nu,\mu} \mathcal{H}_{\nu,\mu}$ . The null space of  $J(f) = J(f, f)$  is

$$\begin{aligned} \mathcal{N}_J = & \mathcal{H}_{00,00} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{00,01} \oplus \mathcal{H}_{01,01} \\ = & \text{span}\{\phi_{00,00}, \phi_{01,00}, \phi_{00,01}, \phi_{01,01}\} \\ = & \text{span}\{1, k_1(x_{(1)}), k_1(x_{(2)}), k_1(x_{(1)})k_1(x_{(2)})\}, \end{aligned}$$

where the basis functions  $\phi_{\nu,\mu}$  are explicitly specified. The minimizer of (2.31) in  $\mathcal{H} = \oplus_{\nu,\mu} \mathcal{H}_{\nu,\mu}$  has an expression

$$\eta(x) = \sum_{\nu,\mu=00,01} d_{\nu,\mu} \phi_{\nu,\mu}(x) + \sum_{i=1}^n c_i R_J(x_i, x),$$

where

$$R_J = \theta_{1,00} R_{1,00} + \theta_{1,01} R_{1,01} + \theta_{00,1} R_{00,1} + \theta_{01,1} R_{01,1} + \theta_{1,1} R_{1,1}.$$

The projections of  $\eta$  in  $\mathcal{H}_{\nu,\mu}$  are readily available from the expression. For example,  $\eta_{01,00} = d_{01,00} \phi_{01,00}(x)$  and  $\eta_{01,1} = \sum_{i=1}^n c_i \theta_{01,1} R_{01,1}(x_i, x)$ .

To fit an additive model, one may set

$$J(f, g) = \theta_{1,00}^{-1}(f, g)_{1,00} + \theta_{00,1}^{-1}(f, g)_{00,1}$$

and minimize (2.31) in  $\mathcal{H}_a = \mathcal{H}_{00,00} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{1,00} \oplus \mathcal{H}_{00,01} \oplus \mathcal{H}_{00,1}$ . The null space is now

$$\mathcal{N}_J = \mathcal{H}_{00,00} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{00,01} = \text{span}\{\phi_{00,00}, \phi_{01,00}, \phi_{00,01}\},$$

and  $\mathcal{H}_J = \mathcal{H}_{1,00} \oplus \mathcal{H}_{00,1}$  with a reproducing kernel

$$R_J = \theta_{1,00}R_{1,00} + \theta_{00,1}R_{00,1}.$$

The spaces  $\mathcal{H}_{01,01}$ ,  $\mathcal{H}_{1,01}$ ,  $\mathcal{H}_{01,1}$ , and  $\mathcal{H}_{1,1}$  are eliminated from  $\mathcal{H}_a$ .  $\square$

## 2.5 Bayes Model

Penalized likelihood estimation in a reproducing kernel Hilbert space  $\mathcal{H}$  with the penalty  $J(f)$  a square (semi) norm is equivalent to a certain empirical Bayes model with a Gaussian prior. The prior has a diffuse component in the null space  $\mathcal{N}_J$  of  $J(f)$  and a proper component in  $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$  with mean zero and a covariance function proportional to the reproducing kernel  $R_J$  in  $\mathcal{H}_J$ . The Bayes model may also be perceived as a mixed-effect model, with the fixed effects residing in  $\mathcal{N}_J$  and the random effects residing in  $\mathcal{H}_J$ .

We start the discussion with the familiar shrinkage estimates on discrete domains, followed by the polynomial smoothing splines on  $[0, 1]$ . The calculus is seen to depend only on the null space  $\mathcal{N}_J$  of  $J(f)$  and the reproducing kernel  $R_J$  in its orthogonal complement  $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$ , hence applies to smoothing splines in general. The general results are noted concerning the general multiple-term smoothing splines of §2.4.5.

### 2.5.1 Shrinkage Estimates as Bayes Estimates

Consider the classical one-way ANOVA model with independent observations  $Y_i \sim N(\eta(x_i), \sigma^2)$ ,  $i = 1, \dots, n$ , where  $x_i \in \{1, \dots, K\}$ . With a prior  $\eta \sim N(0, bI)$ , it is easy to see that the posterior mean of  $\eta$  is given by the minimizer of

$$\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \frac{1}{b} \sum_{x=1}^K \eta^2(x). \quad (2.32)$$

Setting  $b = \sigma^2/n\lambda$ , (2.32) is equivalent to (2.4) of §2.2.

Now, consider  $\eta = \alpha \mathbf{1} + \eta_1$ , with independent priors  $\alpha \sim N(0, \tau^2)$  for the mean and  $\eta_1 \sim N(0, b(I - \mathbf{1}\mathbf{1}^T/K))$  for the contrast. Note that  $\eta_1^T \mathbf{1} = 0$  almost surely and that  $\bar{\eta} = \sum_{x=1}^K \eta(x)/K = \alpha$ . The posterior mean of  $\eta$  is given by the minimizer of

$$\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \frac{1}{\tau^2} \bar{\eta}^2 + \frac{1}{b} \sum_{x=1}^K (\eta(x) - \bar{\eta})^2. \quad (2.33)$$

Letting  $\tau^2 \rightarrow \infty$  and setting  $b = \sigma^2/n\lambda$ , (2.33) reduces to (2.3) of §2.2. In the limit,  $\alpha$  is said to have a diffuse prior. This setting may also be considered as a mixed-effect model, with  $\alpha\mathbf{1}$  being the fixed effect and  $\eta_1$  being the random effect.

Next we look at a two-way ANOVA model on  $\{1, \dots, K_1\} \times \{1, \dots, K_2\}$  using the notation of §2.4.2. Assume that  $\eta = \eta_\emptyset + \eta_1 + \eta_2 + \eta_{1,2}$  has four independent components, with priors  $\eta_\emptyset \sim N(0, b\theta_\emptyset R_\emptyset)$ ,  $\eta_1 \sim N(0, b\theta_1 R_1)$ ,  $\eta_2 \sim N(0, b\theta_2 R_2)$ , and  $\eta_{1,2} \sim N(0, b\theta_{1,2} R_{1,2})$ , where  $R_\emptyset = R_{0(1)}R_{0(2)}$ ,  $R_1 = R_{1(1)}R_{0(2)}$ ,  $R_2 = R_{0(1)}R_{1(2)}$ , and  $R_{1,2} = R_{1(1)}R_{1(2)}$ , as given in Table 2.1. Note that  $R_\beta$ 's are orthogonal to each other and that an  $\eta_\beta$  resides in the column space of  $R_\beta$  almost surely. The posterior mean of  $\eta$  is given by the minimizer of

$$\frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \frac{1}{b} \sum_{\beta} \theta_\beta^{-1} \eta^T R_\beta^+ \eta. \quad (2.34)$$

Setting  $b = \sigma^2/n\lambda$  and  $J(f) = \sum_{\beta} \theta_\beta^{-1} f^T R_\beta^+ f$ , (2.34) reduces to (2.31) of §2.4.5, which defines a bivariate smoothing spline on a discrete product domain. A  $\theta_\beta = \infty$  in  $J(f)$  puts  $\eta_\beta$  in  $\mathcal{N}_J$ , which is equivalent to a diffuse prior, or a fixed effect in a mixed-effect model. To obtain the additive model, one simply eliminates  $\eta_{1,2}$  by setting  $\theta_{1,2} = 0$ .

## 2.5.2 Polynomial Smoothing Splines as Bayes Estimates

Consider  $\eta = \eta_0 + \eta_1$  on  $[0, 1]$ , with  $\eta_0$  and  $\eta_1$  having independent Gaussian priors with mean zero and covariance functions,

$$E[\eta_0(x)\eta_0(y)] = \tau^2 R_0(x, y) = \tau^2 \sum_{\nu=0}^{m-1} \frac{x^\nu y^\nu}{\nu! \nu!},$$

$$E[\eta_1(x)\eta_1(y)] = b R_1(x, y) = b \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} \frac{(y-u)_+^{m-1}}{(m-1)!} du,$$

where  $R_0$  and  $R_1$  are taken from (2.9) and (2.10) of §2.3.1. Observing  $Y_i \sim N(\eta(x_i), \sigma^2)$ , the joint distribution of  $\mathbf{Y}$  and  $\eta(x)$  is normal with mean zero and a covariance matrix

$$\begin{pmatrix} bQ + \tau^2 S S^T + \sigma^2 I & b\xi + \tau^2 S\phi \\ b\xi^T + \tau^2 \phi^T S^T & bR_1(x, x) + \tau^2 \phi^T \phi \end{pmatrix}, \quad (2.35)$$

where  $Q$  is  $n \times n$  with the  $(i, j)$ th entry  $R_1(x_i, x_j)$ ,  $S$  is  $n \times m$  with the  $(i, \nu)$ th entry  $x_i^{\nu-1}/(\nu-1)!$ ,  $\xi$  is  $n \times 1$  with the  $i$ th entry  $R_1(x_i, x)$ , and  $\phi$  is  $m \times 1$  with the  $\nu$ th entry  $x^{\nu-1}/(\nu-1)!$ . Using a standard result on multivariate normal distribution (see, e.g., Johnson and Wichern (1992, Result 4.6)), the posterior mean of  $\eta(x)$  is seen to be

$$\begin{aligned}
E[\eta(x)|\mathbf{Y}] &= (b\boldsymbol{\xi}^T + \tau^2\boldsymbol{\phi}^T S^T)(bQ + \tau^2 SS^T + \sigma^2 I)^{-1}\mathbf{Y} \\
&= \boldsymbol{\xi}^T(Q + \rho SS^T + n\lambda I)^{-1}\mathbf{Y} \\
&\quad + \boldsymbol{\phi}^T \rho S^T(Q + \rho SS^T + n\lambda I)^{-1}\mathbf{Y},
\end{aligned} \tag{2.36}$$

where  $\rho = \tau^2/b$  and  $n\lambda = \sigma^2/b$ .

**Lemma 2.7** *Suppose  $M$  is symmetric and nonsingular and  $S$  is of full column rank.*

$$\lim_{\rho \rightarrow \infty} (\rho SS^T + M)^{-1} = M^{-1} - M^{-1}S(S^T M^{-1}S)^{-1}S^T M^{-1}, \tag{2.37}$$

$$\lim_{\rho \rightarrow \infty} \rho S^T (\rho SS^T + M)^{-1} = (S^T M^{-1}S)^{-1}S^T M^{-1}. \tag{2.38}$$

*Proof:* It can be verified that (Problem 2.17)

$$\begin{aligned}
(\rho SS^T + M)^{-1} &= \\
M^{-1} - M^{-1}S(S^T M^{-1}S)^{-1}(I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1}S^T M^{-1}.
\end{aligned} \tag{2.39}$$

Equation (2.37) follows trivially from (2.39). Substituting (2.39) into the left-hand side of (2.38), some algebra leads to

$$\begin{aligned}
\rho S^T (\rho SS^T + M)^{-1} &= \rho(I - (I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1})S^T M^{-1} \\
&= (S^T M^{-1}S)^{-1}(I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1}S^T M^{-1}.
\end{aligned}$$

Letting  $\rho \rightarrow \infty$  yields (2.38).  $\square$

Setting  $\rho \rightarrow \infty$  in (2.36) and applying Lemma 2.7, the posterior mean  $E[\eta(x)|\mathbf{Y}]$  is of the form  $\boldsymbol{\xi}^T \mathbf{c} + \boldsymbol{\phi}^T \mathbf{d}$ , with the coefficients given by

$$\begin{aligned}
\mathbf{c} &= (M^{-1} - M^{-1}S(S^T M^{-1}S)^{-1}S^T M^{-1})\mathbf{Y}, \\
\mathbf{d} &= (S^T M^{-1}S)^{-1}S^T M^{-1}\mathbf{Y},
\end{aligned} \tag{2.40}$$

where  $M = Q + n\lambda I$ .

**Theorem 2.8** *The polynomial smoothing spline of (2.5) is the posterior mean of  $\eta = \eta_0 + \eta_1$ , where  $\eta_0$  diffuses in  $\text{span}\{x^{\nu-1}, \nu = 1, \dots, m\}$  and  $\eta_1$  has a Gaussian process prior with mean zero and a covariance function*

$$bR_1(x, y) = b \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} \frac{(y-u)_+^{m-1}}{(m-1)!} du,$$

for  $b = \sigma^2/n\lambda$ .

*Proof:* The only thing that remains to be verified is that  $\mathbf{c}$  and  $\mathbf{d}$  in (2.40) minimize (2.16) on page 36. Differentiating (2.16) with respect to  $\mathbf{c}$  and  $\mathbf{d}$  and setting the derivatives to 0, one gets

$$\begin{aligned} Q\{(Q + n\lambda I)\mathbf{c} + S\mathbf{d} - \mathbf{Y}\} &= 0, \\ S^T\{Q\mathbf{c} + S\mathbf{d} - \mathbf{Y}\} &= 0. \end{aligned} \quad (2.41)$$

It is easy to verify that  $\mathbf{c}$  and  $\mathbf{d}$  given in (2.40) satisfy (2.41).  $\square$

### 2.5.3 Smoothing Splines as Bayes Estimates: General Form

Besides the choices of covariance functions  $R_0$  and  $R_1$ , nothing is specific to polynomial smoothing splines in the derivation of §2.5.2. In general, consider a reproducing kernel Hilbert space  $\mathcal{H} = \oplus_{\beta=0}^p \mathcal{H}_\beta$  on a domain  $\mathcal{X}$  with an inner product

$$(f, g) = \sum_{\beta=0}^p \theta_\beta^{-1}(f, g)_\beta = \sum_{\beta=0}^p \theta_\beta^{-1}(f_\beta, g_\beta)_\beta$$

and a reproducing kernel

$$R(x, y) = \sum_{\beta=0}^p \theta_\beta R_\beta(x, y),$$

where  $(f, g)_\beta$  is an inner product in  $\mathcal{H}_\beta$  with a reproducing kernel  $R_\beta$ ,  $f_\beta$  is the projection of  $f$  in  $\mathcal{H}_\beta$ , and  $\mathcal{H}_0$  is finite dimensional. Observing  $Y_i \sim N(\eta(x_i), \sigma^2)$ , a smoothing spline on  $\mathcal{X}$  can be defined as the minimizer of the functional

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda \sum_{\beta=1}^p \theta_\beta^{-1}(\eta, \eta)_\beta \quad (2.42)$$

in  $\mathcal{H}$ ; see also (2.31) of §2.4.5. A smoothing spline thus defined is a Bayes estimate of  $\eta = \sum_{\beta=0}^p \eta_\beta$ , where  $\eta_0$  has a diffuse prior in  $\mathcal{H}_0$  and  $\eta_\beta$ ,  $\beta = 1, \dots, p$ , have mean zero Gaussian process priors on  $\mathcal{X}$  with covariance functions  $E[\eta_\beta(x)\eta_\beta(y)] = b\theta_\beta R_\beta(x, y)$ , independent of each other, where  $b = \sigma^2/n\lambda$ . Treated as a mixed-effect model,  $\eta_0$  contains the fixed effects and  $\eta_\beta$ ,  $\beta = 1, \dots, p$ , are the random effects.

## 2.6 Minimization of Penalized Functional

As an optimization object, analytical properties of the penalized likelihood functional  $L(f) + (\lambda/2)J(f)$  can be studied under general functional analytical conditions such as the continuity, convexity, and differentiability of  $L(f)$  and  $J(f)$ . Among such properties are the existence of the minimizer and the equivalence of penalized optimization and constrained optimization.

We first show that the penalized likelihood estimate exists as long as the maximum likelihood estimate uniquely exists in the null space  $\mathcal{N}_J$  of  $J(f)$ .

We then prove that the minimization of  $L(f) + (\lambda/2)J(f)$  is equivalent to the minimization of  $L(f)$  subject to a constraint of the form  $J(f) \leq \rho$  for some  $\rho \geq 0$ , and quantify the relation between  $\rho$  and  $\lambda$ .

### 2.6.1 Existence of Minimizer

A functional  $A(f)$  in a linear space  $\mathcal{L}$  is said to be **convex** if for  $f, g \in \mathcal{L}$ ,  $A(\alpha f + (1-\alpha)g) \leq \alpha A(f) + (1-\alpha)A(g)$ ,  $\forall \alpha \in (0, 1)$ ; the convexity is strict if the equality holds only for  $f = g$ .

**Theorem 2.9 (Existence)** *Suppose  $L(f)$  is a continuous and convex functional in a Hilbert space  $\mathcal{H}$  and  $J(f)$  is a square (semi) norm in  $\mathcal{H}$  with a null space  $\mathcal{N}_J$ , of finite dimension. If  $L(f)$  has a unique minimizer in  $\mathcal{N}_J$ , then  $L(f) + (\lambda/2)J(f)$  has a minimizer in  $\mathcal{H}$ .*

The minus log likelihood  $L(f|\text{data})$  in (1.3) is usually convex in  $f$ , as will be verified on a case-by-case basis in later chapters. The quadratic functional  $J(f)$  is convex; see Problem 2.18. A minimizer of  $L(f)$  is unique in  $\mathcal{N}_J$  if the convexity is strict in it, which is often the case.

Without loss of generality, one may set  $\lambda = 2$  in the theorem. The proof of the theorem builds on the following two lemmas, with  $L(f)$  and  $J(f)$  in the lemmas being the same as those in Theorem 2.9.

**Lemma 2.10** *If a continuous and convex functional  $A(f)$  has a unique minimizer in  $\mathcal{N}_J$ , then it has a minimizer in the cylinder area  $C_\rho = \{f : f \in \mathcal{H}, J(f) \leq \rho\}$ ,  $\forall \rho \in (0, \infty)$ .*

**Lemma 2.11** *If  $L(f) + J(f)$  has a minimizer in  $C_\rho = \{f : f \in \mathcal{H}, J(f) \leq \rho\}$ ,  $\forall \rho \in (0, \infty)$ , then it has a minimizer in  $\mathcal{H}$ .*

The rest of the section are the proofs.

*Proof of Lemma 2.10:* Let  $\|\cdot\|_0$  be the norm in  $\mathcal{N}_J$ , and  $f_0$  be the unique minimizer of  $A(f)$  in  $\mathcal{N}_J$ . By Theorem 4 of Tapia and Thompson (1978, p. 162),  $A(f)$  has a minimizer in a “rectangle”

$$R_{\rho, \gamma} = \{f : f \in \mathcal{H}, J(f) \leq \rho, \|f - f_0\|_0 \leq \gamma\}.$$

Now, if the lemma is not true (i.e., that  $A(f)$  has no minimizer in  $C_\rho$  for some  $\rho$ ), then a minimizer  $f_\gamma$  of  $A(f)$  in  $R_{\rho, \gamma}$  must satisfy  $\|f_\gamma - f_0\|_0 = \gamma$ . By the convexity of  $A(f)$  and the fact that  $A(f_\gamma) \leq A(f_0)$ ,

$$A(\alpha f_\gamma + (1-\alpha)f_0) \leq \alpha A(f_\gamma) + (1-\alpha)A(f_0) \leq A(f_0), \quad (2.43)$$

for  $\alpha \in (0, 1)$ . Now, take a sequence  $\gamma_i \rightarrow \infty$  and set  $\alpha_i = \gamma_i^{-1}$ , and write  $\alpha_i f_{\gamma_i} + (1-\alpha_i)f_0 = f_i^o + f_i^*$ , where  $f_i^o \in \mathcal{N}_J$  and  $f_i^* \in \mathcal{H} \ominus \mathcal{N}_J$ . It is

easy to check that  $\|f_i^o - f_0\|_0 = 1$  and that  $J(f_i^*) \leq \alpha_i^2 \rho$ . Since  $\mathcal{N}_J$  is finite dimensional,  $\{f_i^o\}$  has a convergent subsequence converging to, say,  $f_1 \in \mathcal{N}_J$ , and  $\|f_1 - f_0\|_0 = 1$ . It is apparent that  $f_i^* \rightarrow 0$ . By the continuity of  $A(f)$  and (2.43),  $A(f_1) \leq A(f_0)$ , which contradicts the fact that  $f_0$  uniquely minimizes  $A(f)$  in  $\mathcal{N}_J$ . Hence,  $\|f_\gamma - f_0\|_0 = \gamma$  cannot hold for all  $\gamma \in (0, \infty)$ . This completes the proof.  $\square$

*Proof of Lemma 2.11:* Without loss of generality we assume  $L(0) = 0$ . If the lemma is not true, then a minimizer  $f_\rho$  of  $L(f) + J(f)$  in  $C_\rho$  must fall on the boundary of  $C_\rho$  for every  $\rho$  (i.e.,  $J(f_\rho) = \rho$ ,  $\forall \rho \in (0, \infty)$ ). By the convexity of  $L(f)$ ,

$$L(\alpha f_\rho) \leq \alpha L(f_\rho), \quad (2.44)$$

for  $\alpha \in (0, 1)$ . By the definition of  $f_\rho$ ,

$$L(f_\rho) + J(f_\rho) \leq L(\alpha f_\rho) + J(\alpha f_\rho). \quad (2.45)$$

Combining (2.44) and (2.45) and substituting  $J(f_\rho) = \rho$ , one obtains

$$L(\alpha f_\rho)/\alpha + \rho \leq L(\alpha f_\rho) + \alpha^2 \rho,$$

which, after some algebra, yields

$$L(\alpha f_\rho) \leq -\alpha(1 + \alpha)\rho. \quad (2.46)$$

Now, choose  $\alpha = \rho^{-1/2}$ . Since  $J(\alpha f_\rho) = 1$ , (2.46) leads to

$$L(f_1) \leq -(\rho^{1/2} + 1),$$

which is impossible for large enough  $\rho$ . This proves the lemma.  $\square$

*Proof of Theorem 2.9:* Applying Lemma 2.10 on  $A(f) = L(f) + J(f)$  leads to the condition of Lemma 2.11, and the lemma, in turn, yields the theorem.  $\square$

## 2.6.2 Penalized and Constrained Optimization

For a functional  $A(f)$  in a linear space  $\mathcal{L}$ , define  $A_{f,g}(\alpha) = A(f + \alpha g)$  as functions of  $\alpha$  real indexed by  $f, g \in \mathcal{L}$ . If  $\dot{A}_{f,g}(0)$  exists and is linear in  $g$ ,  $\forall f, g \in \mathcal{L}$ ,  $A(f)$  is said to be **Fréchet differentiable** in  $\mathcal{L}$ , and  $\dot{A}_{f,g}(0)$  is the **Fréchet derivative** of  $A$  at  $f$  in the direction of  $g$ .

**Theorem 2.12** Suppose  $L(f)$  is continuous, convex, and Fréchet differentiable in a Hilbert space  $\mathcal{H}$ , and  $J(f)$  is a square (semi) norm in  $\mathcal{H}$ . If  $f^*$  minimizes  $L(f)$  in  $C_\rho = \{f : f \in \mathcal{H}, J(f) \leq \rho\}$ , then  $f^*$  minimizes  $L(f) + (\lambda/2)J(f)$  in  $\mathcal{H}$ , where the Lagrange multiplier relates to  $\rho$  via  $\lambda = -\rho^{-1} \dot{L}_{f^*, f_1^*}(0) \geq 0$ , with  $f_1^*$  being the projection of  $f^*$  in  $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$ . Conversely, if  $f^o$  minimizes  $L(f) + (\lambda/2)J(f)$  in  $\mathcal{H}$ , where  $\lambda > 0$ , then  $f^o$  minimizes  $L(f)$  in  $\{f : f \in \mathcal{H}, J(f) \leq J(f^o)\}$ .

The minus log likelihood  $L(f|\text{data})$  in (1.3) is usually Fréchet differentiable, as will be verified on a case-by-case basis in later chapters.

*Proof of Theorem 2.12:* If  $J(f^*) < \rho$ , then by the convexity of  $L(f)$ ,  $f^*$  is a global minimizer of  $L(f)$ , so the result holds with  $\lambda = \dot{L}_{f^*, f_1^*}(0) = 0$ .

In general,  $J(f^*) = \rho$ ; thus,  $f^*$  minimizes  $L(f)$  on the boundary contour  $C_\rho^o = \{f : f \in \mathcal{H}, J(f) = \rho\}$ . It is easy to verify that  $\dot{J}_{f,g}(0) = 2J(f, g)$ , where  $J(f, g)$  is the (semi) inner product associated with  $J(f)$ . The space tangent to the contour  $C_\rho^o$  at  $f^*$  is thus  $\mathcal{G} = \{g : J(f^*, g) = J(f_1^*, g) = 0\}$ .

Pick an arbitrary  $g \in \mathcal{G}$ . When  $J(g) = 0$ ,  $f^* + \alpha g \in C_\rho^o$ . Since

$$0 \leq L(f^* + \alpha g) - L(f^*) = \alpha \dot{L}_{f^*, g}(0) + o(\alpha),$$

one has  $\dot{L}_{f^*, g}(0) = 0$ . When  $J(g) \neq 0$ , without loss of generality one may scale  $g$  so that  $J(g) = \rho$ ; then,  $\sqrt{1 - \alpha^2}f^* + \alpha g \in C_\rho^o$ . Now, write  $\gamma = (\sqrt{1 - \alpha^2} - 1)/\alpha$ . By the linearity of  $\dot{L}_{f,g}(0)$  in  $g$ , one has

$$\begin{aligned} 0 &\leq L(\sqrt{1 - \alpha^2}f^* + \alpha g) - L(f^*) \\ &= L(f^* + \alpha(\gamma f^* + g)) - L(f^*) \\ &= \alpha\gamma \dot{L}_{f^*, f^*}(0) + \alpha \dot{L}_{f^*, g}(0) + o(\alpha) \\ &= \alpha \dot{L}_{f^*, g}(0) + o(\alpha), \end{aligned}$$

where  $\alpha\gamma = \sqrt{1 - \alpha^2} - 1 = O(\alpha^2) = o(\alpha)$ ; so, again,  $\dot{L}_{f^*, g}(0) = 0$ .

It is easy to see that  $J(f_1^*) = \rho$  and that  $\mathcal{G}^c = \text{span}\{f_1^*\}$ . Now, every  $f \in \mathcal{H}$  has a unique decomposition  $f = \beta f_1^* + g$ , with  $\beta$  real and  $g \in \mathcal{G}$ ; hence,

$$\begin{aligned} \dot{L}_{f^*, f}(0) + \frac{\lambda}{2} \dot{J}_{f^*, f}(0) &= \dot{L}_{f^*, \beta f_1^*}(0) + \dot{L}_{f^*, g}(0) + \lambda J(f^*, \beta f_1^* + g) \\ &= \beta \dot{L}_{f^*, f_1^*}(0) + \beta \lambda \rho. \end{aligned} \tag{2.47}$$

With  $\lambda = -\rho^{-1} \dot{L}_{f^*, f_1^*}(0)$ , (2.47) is annihilated for all  $f \in \mathcal{H}$ ; thus,  $f^*$  minimizes  $L(f) + (\lambda/2)J(f)$ . Finally, note that  $L(f^* - \alpha f_1^*) \geq L(f^*)$  for  $\alpha \in (0, 1)$ , so  $\dot{L}_{f^*, f_1^*}(0) \leq 0$ . The converse is straightforward and is left as an exercise (Problem 2.21).  $\square$

## 2.7 Bibliographic Notes

### Section 2.1

The theory of Hilbert space is at the core of many advanced analysis courses. The elementary materials presented in §2.1.1 provide a minimal exposition for our need. An excellent treatment of vector spaces can be found in Rao (1973, Chap. 1). Proofs of the Riesz representation theorem



can be found in many references, of different levels of abstraction; the one given in §2.1.2 was taken from Akhiezer and Glazman (1961). The theory of reproducing kernel Hilbert space was developed by Aronszajn (1950), which remains the primary reference on the subject. The exposition in §2.1.3 is minimally sufficient to serve our need.

## Section 2.2

Shrinkage estimates are among basic techniques in classical decision theory and Bayesian statistics; see, e.g., Lehmann and Casella (1998, §5.5). The interpretation of shrinkage estimates as smoothing splines on discrete domains has not appeared elsewhere. Vector spaces are much more familiar to statisticians than reproducing kernel Hilbert spaces, and this section is intended to help the reader to gain further insights into entities in a reproducing kernel Hilbert space.

## Section 2.3

The space  $\mathcal{C}^{(m)}[0, 1]$  with the inner product (2.7) and the representer of evaluation (2.8) derived from the standard Taylor expansion are standard results found in numerical analysis literature; see, e.g., Schumaker (1981, Chap. 8). The reproducing kernel (2.21) of  $\mathcal{C}^{(m)}[0, 1]$  associated with the inner product (2.17) was derived by Craven and Wahba (1979), and was used more often than (2.8) as marginal kernels in tensor product smoothing splines. Results concerning Bernoulli polynomials can be found in Abramowitz and Stegun (1964, Chap. 23).

The computational strategy outlined in §2.3.2 was derived by Kimeldorf and Wahba (1971) in the setting of Chebyshev splines, of which the polynomial smoothing splines of (2.5) are special cases; see §4.5.2 for Chebyshev splines. For many years, however, the device was not used much in actual numerical computation. The reasons were multifold. First, algorithms based on (2.16) are of order  $O(n^3)$ , whereas  $O(n)$  algorithms exist for polynomial smoothing splines; see §§3.4 and 3.10. Second, portable numerical linear algebra software and powerful desktop computing were not available until much later. Since the late 1980s, generic algorithms and software have been developed based on (2.16) for the computation of smoothing splines, univariate and multivariate alike; see §3.4 for details.

## Section 2.4

A comprehensive treatment of tensor product reproducing kernel Hilbert spaces can be found in Aronszajn (1950), where Theorem 2.6 was quoted as a classical result of I. Schur. The proof given here was suggested by Liqing Yan.

The idea of tensor product smoothing splines was conceived by Barry (1986) and Wahba (1986). Dozens of references appeared in the literature since then, among which Chen (1991), Gu and Wahba (1991b, 1993a, 1993b), Gu (1992b, 1995a, 1996, 2004), Wahba, Wang, Gu, Klein, and Klein (1995) and Gu and Ma (2011) registered notable innovations in the theory and practice of the tensor product spline technique. The materials of §§2.4.3–2.4.5 are scattered in these references. The materials of §2.4.2, however, had not appeared in the smoothing literature prior to the first edition of this book.

## Section 2.5

The Bayes model of polynomial smoothing splines was first observed by Kimeldorf and Wahba (1970a, 1970b). The materials of §§2.5.2 and 2.5.3 are mainly taken from Wahba (1978, 1983). The elementary materials of §2.5.1 in the familiar discrete setting provide insights into the general results. In Bayesian statistics, such models are more specifically referred to as empirical Bayes models; see, e.g., Berger (1985, §4.5).

## Section 2.6

The existence of penalized likelihood estimates has been discussed by many authors in various settings; see, e.g., Tapia and Thompson (1978, Chap. 4) and Silverman (1982). The general result of Theorem 2.9 and the elementary proof are taken from Gu and Qiu (1993).

The relation between penalized optimization and constrained optimization in the context of natural polynomial splines was noted by Schoenberg (1964), where  $L(f)$  was a least squares functional. The general result of Theorem 2.12 was adapted from the discussion of Gill, Murray, and Wright (1981, §3.4) on constrained nonlinear optimization.

# 2.8 Problems

## Section 2.1

**2.1** Prove the Cauchy-Schwarz inequality of (2.1).

**2.2** Prove the triangle inequality of (2.2).

**2.3** Let  $\mathcal{H}$  be a Hilbert space and  $\mathcal{G} \subset \mathcal{H}$  a closed linear subspace. For every  $f \in \mathcal{H}$ , prove that the projection of  $f$  in  $\mathcal{G}$ ,  $f_{\mathcal{G}} \in \mathcal{G}$ , that satisfies

$$\|f - f_{\mathcal{G}}\| = \inf_{g \in \mathcal{G}} \|f - g\|$$

uniquely exists.

- (a) Show that there exists a sequence  $\{g_n\} \subset \mathcal{G}$  such that

$$\lim_{n \rightarrow \infty} \|f - g_n\| = \delta = \inf_{g \in \mathcal{G}} \|f - g\|.$$

- (b) Show that

$$\|g_m - g_n\|^2 = 2\|f - g_m\|^2 + 2\|f - g_n\|^2 - 4\|f - \frac{g_m + g_n}{2}\|^2.$$

Since  $\lim_{m,n \rightarrow \infty} \|f - \frac{g_m + g_n}{2}\| = \delta$ ,  $\{g_n\}$  is a Cauchy sequence.

- (c) Show the uniqueness of  $f_{\mathcal{G}}$  using the triangle inequality.

**2.4** Given Hilbert spaces  $\mathcal{H}_0$  and  $\mathcal{H}_1$  satisfying  $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$ , prove that the space  $\mathcal{H} = \{f : f = f_0 + f_1, f_0 \in \mathcal{H}_0, f_1 \in \mathcal{H}_1\}$  with an inner product  $(f, g) = (f_0, g_0)_0 + (f_1, g_1)_1$  is a Hilbert space, where  $f = f_0 + f_1$ ,  $g = g_0 + g_1$ ,  $f_0, g_0 \in \mathcal{H}_0$ ,  $f_1, g_1 \in \mathcal{H}_1$ , and  $(\cdot, \cdot)_0$  and  $(\cdot, \cdot)_1$  are the inner products in  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , respectively. Prove that  $\mathcal{H}_0$  and  $\mathcal{H}_1$  are the orthogonal complements of each other as closed linear subspaces of  $\mathcal{H}$ .

**2.5** The isomorphism between a  $K$ -dimensional Hilbert space  $\mathcal{H}$  and the Euclidean  $K$ -space is outlined in the following steps:

- (a) Take any  $\phi \in \mathcal{H}^0 = \mathcal{H}$  nonzero, denote  $\phi_1 = \phi/\|\phi\|$ , and obtain

$$\mathcal{H}^1 = \mathcal{H}^0 \ominus \{f : f = \alpha\phi_1, \alpha \text{ real}\}.$$

Prove that  $\mathcal{H}^1$  contains nonzero elements if  $K > 1$ .

- (b) Repeat step (a) for  $\mathcal{H}^{i-1}$ ,  $i = 2, \dots, K$ , to obtain  $\phi_i$  and

$$\mathcal{H}^i = \mathcal{H}^{i-1} \ominus \{f : f = \alpha\phi_i, \alpha \text{ real}\}.$$

Prove that  $\mathcal{H}^{K-1} = \{f : f = \alpha\phi_K, \alpha \text{ real}\}$ , so  $\mathcal{H}^K = \{0\}$ .

- (c) Verify that  $(\phi_i, \phi_j) = \delta_{i,j}$ , where  $\delta_{i,j}$  is the Kronecker delta. The elements  $\phi_i$ ,  $i = 1, \dots, K$ , are said to form an orthonormal basis of  $\mathcal{H}$ . For every  $f \in \mathcal{H}$ , there is a unique representation  $f = \sum_{i=1}^K \alpha_i \phi_i$ , where  $\alpha_i$  are real coefficients.
- (d) Prove that the mapping  $f \leftrightarrow \alpha$ , where  $\alpha$  are the coefficients of  $f$ , defines an isomorphism between  $\mathcal{H}$  and the Euclidean space.

**2.6** Prove that in an Euclidean space, every linear functional is continuous.

**2.7** Prove that the reproducing kernel of a Hilbert space, when it exists, is unique.

## Section 2.2

**2.8** On  $\mathcal{X} = \{1, \dots, K\}$ , the constructions of reproducing kernel Hilbert spaces outlined below yield a one-way ANOVA decomposition with an averaging operator  $Af = f(1)$ .

- (a) Verify that the reproducing kernel  $R_0 = 1 = \mathbf{1}\mathbf{1}^T$  generates the space  $\mathcal{H}_0 = \{f : f(1) = \dots = f(K)\}$  with an inner product  $(f, g)_0 = f^T(\mathbf{1}\mathbf{1}^T/K^2)g$ .
- (b) Verify that the reproducing kernel  $R_1 = I_{[x=y \neq 1]} = (I - e_1 e_1^T)$  generates the space  $\mathcal{H}_1 = \{f : f(1) = 0\}$  with an inner product  $(f, g)_1 = f^T(I - e_1 e_1^T)g$ , where  $e_1$  is the first unit vector.
- (c) Note that  $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$ , so  $\mathcal{H}_0 \oplus \mathcal{H}_1$  is well defined and has the reproducing kernel  $R_0 + R_1$ . With the expressions given in (a) and (b), however, one in general has  $(f_1, f_1)_0 \neq 0$  for  $f_1 \in \mathcal{H}_1$  and  $(f_0, f_0)_1 \neq 0$  for  $f_0 \in \mathcal{H}_0$ . Nevertheless,  $f = \mathbf{1}e_1^T f$  for  $f \in \mathcal{H}_0$ , so one may write  $(f, g)_0 = f^T(e_1 e_1^T)g$ . Similarly, as  $f = (I - \mathbf{1}e_1^T)f$  for  $f \in \mathcal{H}_1$ , one may write  $(f, g)_1 = f^T(I - e_1 \mathbf{1}^T)(I - \mathbf{1}e_1^T)g$ . Verify the new expressions of  $(f, g)_0$  and  $(f, g)_1$ . Check that with the new expressions,  $(f_1, f_1)_0 = 0$ ,  $\forall f_1 \in \mathcal{H}_1$ , and that  $(f_0, f_0)_1 = 0$ ,  $\forall f_0 \in \mathcal{H}_0$ , so the inner product in  $\mathcal{H}_0 \oplus \mathcal{H}_1$  can be written as  $(f, g) = (f, g)_0 + (f, g)_1$  with the new expressions.
- (d) Verify that  $(\mathbf{1}\mathbf{1}^T + I - e_1 e_1^T)^{-1} = e_1 e_1^T + (I - e_1 \mathbf{1}^T)(I - \mathbf{1}e_1^T)$  (i.e., the reproducing kernel  $R_0 + R_1$  and the inner product  $(f, g)_0 + (f, g)_1$  are inverses of each other).

## Section 2.3

**2.9** Consider the function  $k_r(x)$  of (2.18).

- (a) Prove that the infinite series converges for  $r > 1$  on the real line and for  $r = 1$  at noninteger points.
- (b) Prove that  $k_r(x)$  is real-valued.
- (c) Prove that  $k_1(x) = x - 0.5$  on  $x \in (0, 1)$ .

**2.10** Prove (2.22) through integration by parts, for  $m > 1$ . Note that  $k_r$ ,  $r > 1$ , are periodic with period 1 and that  $\int_0^1 f^{(\nu)} dx = 0$ ,  $\nu = 0, \dots, m-1$ .

**2.11** Derive the expressions of  $k_2(x)$  and  $k_4(x)$  on  $[0, 1]$  as given in (2.27) by successive integration from  $k_1(x) = x - .5$ . Note that for  $r > 1$ ,  $dk_r/dx = k_{r-1}$  and  $k_r(0) = k_r(1)$ .

## Section 2.4

**2.12** On  $\mathcal{X} = \{1, \dots, K_1\} \times \{1, \dots, K_2\}$ , construct tensor product reproducing kernel Hilbert spaces with the structure of (2.28).

- (a) With  $A_1 f = f(1, x_{(2)})$  and  $A_2 f = f(x_{(1)}, 1)$ .
- (b) With  $A_1 f = f(1, x_{(2)})$  and  $A_2 f = \sum_{x_{(2)}=1}^{K_2} f(x)/K_2$ .

**2.13** On  $\mathcal{X} = [0, 1]^2$ , construct tensor product reproducing kernel Hilbert spaces with the structure of (2.28).

- (a) With  $A_1 f = f(0, x_{(2)})$  and  $A_2 f = f(x_{(1)}, 0)$ , using (2.9) and (2.10) with  $m = 1, 2$ .
- (b) With  $A_1 f = f(0, x_{(2)})$  and  $A_2 f = \int_0^1 f dx_{(2)}$ , using (2.9), (2.10), (2.19) and (2.23), with  $m = 1, 2$ .

**2.14** On  $\mathcal{X} = \{1, \dots, K\} \times [0, 1]$ , construct tensor product reproducing kernel Hilbert spaces with the structure of (2.28).

- (a) With  $A_1 f = f(1, x_{(2)})$  and  $A_2 f = f(x_{(1)}, 0)$ .
- (b) With  $A_1 f = f(1, x_{(2)})$  and  $A_2 f = \int_0^1 f dx_{(2)}$ .
- (c) With  $A_1 f = \sum_{x_{(1)}=1}^K f(x)/K$  and  $A_2 f = f(x_{(1)}, 0)$ .

**2.15** To compute the tensor product smoothing splines of Example 2.8, one may use the strategy outlined in §2.3.2.

- (a) Specify the matrices  $S$  and  $Q$  in (2.16), for both the full model and the additive model.
- (b) Decompose the expression of  $\eta(x)$  into those of the constant, the main effects, and the interaction.

**2.16** In parallel to Example 2.8 and Problem 2.15, work out the corresponding details for the computation of tensor product smoothing splines on  $\{1, \dots, K\} \times [0, 1]$ , using the construction of Example 2.7.

## Section 2.5

**2.17** Verify (2.39).

## Section 2.6

**2.18** Prove that a quadratic functional  $J(f)$  is convex.

**2.19** Let  $A(f)$  be a strictly convex functional in a Hilbert space  $\mathcal{H}$ . Prove that if the minimizer of  $A(f)$  exists in  $\mathcal{H}$ , then it is also unique.

**2.20** Consider a strictly convex continuous function  $f(x)$  on  $(-\infty, \infty)^2$ . Prove that if  $f_1(x_{\langle 1 \rangle}) = f(x_{\langle 1 \rangle}, 0)$  has a minimizer, then  $f(x) + x_{\langle 2 \rangle}^2$  has a unique minimizer.

**2.21** Prove that if  $f^o$  minimizes  $L(f) + \lambda J(f)$ , where  $\lambda > 0$ , then  $f^o$  minimizes  $L(f)$  subject to  $J(f) \leq J(f^o)$ .



<http://www.springer.com/978-1-4614-5368-0>

Smoothing Spline ANOVA Models

Gu, C.

2013, XVIII, 433 p., Hardcover

ISBN: 978-1-4614-5368-0