

# To-do list

1. From Wikipedia. But don't really get it, although it might explain the Fisher information between linear functionals. . . . . 10
2. Insert figure squiggly line and smooth line. . . . . 14
3. Update graphics. . . . . 21
4. This is the same for any RKHS? . . . . . 26

# Contents

<b>2</b>	<b>Vector space of functions</b>	<b>2</b>
2.1	Some functional analysis . . . . .	3
2.2	Reproducing kernel Hilbert space theory . . . . .	12
2.3	Reproducing kernel Kreĭn space theory . . . . .	19
2.4	RKHS building blocks . . . . .	21
2.4.1	The RKHS of constant functions . . . . .	21
2.4.2	The canonical (linear) RKHS . . . . .	22
2.4.3	The fractional Brownian motion RKHS . . . . .	24
2.4.4	The squared exponential RKHS . . . . .	27
2.4.5	The Pearson RKHS . . . . .	29
2.5	Constructing RKKS from existing RKHS . . . . .	32
2.5.1	Sums, products and scaling of RKHS . . . . .	32
2.5.2	The polynomial RKKS . . . . .	34
2.5.3	The ANOVA RKKS . . . . .	35
2.6	Summary . . . . .	42
2.7	Miscellanea . . . . .	43
2.7.1	A vector space... of 'functions'? . . . . .	43
	<b>Bibliography</b>	<b>44</b>

Haziq Jamil

*Department of Statistics*

*London School of Economics and Political Science*

PhD thesis: ‘Regression modelling using Fisher information covariance kernels (I-priors)’

## Chapter 2

# Vector space of functions

One of the main assumptions for regression modelling with I-priors is that the regression functions lie in some vector space of functions. The purpose of this chapter is to provide a concise review of functional analysis leading up to the theory of reproducing kernel Hilbert and Kreĭn spaces (RKHS/RKKS). The interest with these RKHS and RKKS is that these spaces have well-established mathematical structure and offer desirable topologies. In particular, it allows the possibility of deriving the Fisher information for regression functions—this will be covered in Chapter 3. As we shall see, RKHS are also extremely convenient in that they may be specified completely via their reproducing kernels. Several of these function spaces are of interest to us, for example, spaces of linear functions, smoothing functions, and functions whose inputs are nominal values and even functions themselves. RKHS are widely studied in the applied statistical and machine learning literature, but perhaps RKKS are less so. To provide an early insight, RKKS are simply a generalisation of RKHS, and are defined as the difference between two RKHSs. The flexibility provided by RKKS will prove both useful and necessary, especially when considering the sums and products of scaled function spaces, as is done in I-prior modelling.

It is emphasised that a deep knowledge of functional analysis, including RKHS and RKKS theory, is not at all necessary for I-prior modelling, so perhaps the advanced reader may wish to skip Sections 2.1–2.3. Section 2.4 describes the fundamental RKHS of interest for I-prior regression, which we refer to as the “building block” RKHS/RKKS. The reason for this is that it is possible to construct new RKKS from existing ones, and this is described in Section 2.5.

A remark on notation: Sets and vector spaces are denoted by calligraphic letters, and as much as possible, we shall stick to the convention that  $\mathcal{F}$  denotes function spaces, and  $\mathcal{X}$  denotes set of covariates or function inputs. Occasionally, we will describe a generic Hilbert space denoted by  $\mathcal{H}$ . Elements of the vector space of real functions over a set  $\mathcal{X}$  are denoted  $f(\cdot)$ , or simply  $f$ . This distinguishes them from the actual evaluation of the function at an input point  $x \in \mathcal{X}$ , denoted  $f(x) \in \mathbb{R}$ . For a much cleaner read, we dispense with boldface notation for vectors and matrices when talking about them, without ambiguity, in the abstract sense.

## 2.1 Some functional analysis

The core study of functional analysis revolves around the treatment of functions as objects in vector spaces over a field<sup>1</sup>. Vector spaces, or linear spaces as they are sometimes known, may be endowed with some kind of structure so as to allow ideas such as closeness and limits to be conceived. Of particular interest to us is the structure brought about by *inner products*, which allow the rigorous mathematical study of various geometrical concepts such as lengths, directions, and orthogonality, among other things. We begin with the definition of an inner product.

**Definition 2.1** (Inner products). Let  $\mathcal{F}$  be a vector space over  $\mathbb{R}$ . A function  $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$  is said to be an inner product on  $\mathcal{F}$  if all of the following are satisfied:

- **Symmetry:**  $\langle f, g \rangle_{\mathcal{F}} = \langle g, f \rangle_{\mathcal{F}}, \forall f, g \in \mathcal{F}$ .
- **Linearity:**  $\langle af_1 + bf_2, g \rangle_{\mathcal{F}} = a\langle f_1, g \rangle_{\mathcal{F}} + b\langle f_2, g \rangle_{\mathcal{F}}, \forall f_1, f_2, g \in \mathcal{F}$  and  $\forall a, b \in \mathbb{R}$ .
- **Non-degeneracy:**  $\langle f, f \rangle_{\mathcal{F}} = 0 \Leftrightarrow f = 0$ .

Additionally, an inner product is said to be *positive definite* if  $\langle f, f \rangle_{\mathcal{F}} \geq 0, \forall f \in \mathcal{F}$ . Inner products need not necessarily be positive definite, and we shall revisit this fact later when we cover Krein spaces. However, for the purposes of the discussion moving forward, the inner products that are referenced are the positive definite kind, unless otherwise stated.

We can always define a *norm* on  $\mathcal{F}$  using the inner product as

$$\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}}. \quad (2.1)$$

---

<sup>1</sup>In this thesis, this will be  $\mathbb{R}$  exclusively.

{eq:normip}

Norms are another form of structure that specifically captures the notion of length. This is defined below.

**Definition 2.2** (Norms). Let  $\mathcal{F}$  be a vector space over  $\mathbb{R}$ . A non-negative function  $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \rightarrow [0, \infty)$  is said to be a norm on  $\mathcal{F}$  if all of the following are satisfied:

- **Absolute homogeneity:**  $\|\lambda f\|_{\mathcal{F}} = |\lambda| \cdot \|f\|_{\mathcal{F}}, \forall \lambda \in \mathbb{R}, \forall f \in \mathcal{F}$
- **Subadditivity:**  $\|f + g\|_{\mathcal{F}} \leq \|f\|_{\mathcal{F}} + \|g\|_{\mathcal{F}}, \forall f, g \in \mathcal{F}$
- **Point separating:**  $\|f\|_{\mathcal{F}} = 0 \Leftrightarrow f = 0$

The subadditivity property is also known as the *triangle inequality*. Also note that since  $\|-f\|_{\mathcal{F}} = \|f\|_{\mathcal{F}}$ , and by the triangle inequality and point separating property, we have that  $\|f\|_{\mathcal{F}} = \frac{1}{2}\|f\|_{\mathcal{F}} + \frac{1}{2}\|-f\|_{\mathcal{F}} \geq \frac{1}{2}\|f - f\|_{\mathcal{F}} = 0$ , thus implying non-negativity of norms. Several important relationships between norms and inner products hold in linear spaces, namely, the *Cauchy-Schwarz inequality*

$$|\langle f, g \rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \cdot \|g\|_{\mathcal{F}};$$

the *parallelogram law*

$$\|f + g\|_{\mathcal{F}}^2 - \|f - g\|_{\mathcal{F}}^2 = 2\|f\|_{\mathcal{F}}^2 + 2\|g\|_{\mathcal{F}}^2;$$

and the *polarisation identity*

$$\|f + g\|_{\mathcal{F}}^2 + \|f - g\|_{\mathcal{F}}^2 = 4\langle f, g \rangle_{\mathcal{F}},$$

for some  $f, g \in \mathcal{F}$ .

A vector space endowed with an inner product (c.f. norm) is called an inner product space (c.f. normed vector space). As a remark, inner product spaces can always be equipped with a norm using (2.1), but not always the other way around. A norm needs to satisfy the parallelogram law for an inner product to be properly defined.

The norm  $\|\cdot\|_{\mathcal{F}}$ , in turn, induces a metric (a notion of distance) on  $\mathcal{F}$ :  $D(f, g) = \|f - g\|_{\mathcal{F}}$ , for  $f, g \in \mathcal{F}$ . With these notions of distances, one may talk about sequences of functions in  $\mathcal{F}$  which are *convergent*, and sequences whose elements become arbitrarily close to one another as the sequence progresses (*Cauchy*).

**Definition 2.3** (Convergent sequence). A sequence  $\{f_n\}_{n=1}^{\infty}$  of elements of a normed vector space  $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$  is said to *converge* to some  $f \in \mathcal{F}$ , if for every  $\epsilon > 0$ ,  $\exists N = N(\epsilon) \in \mathbb{N}$ , such that  $\forall n > N$ ,  $\|f_n - f\|_{\mathcal{F}} < \epsilon$ .

**Definition 2.4** (Cauchy sequence). A sequence  $\{f_n\}_{n=1}^{\infty}$  of elements of a normed vector space  $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$  is said to be a Cauchy sequence if for every  $\epsilon > 0$ ,  $\exists N = N(\epsilon) \in \mathbb{N}$ , such that  $\forall n, m > N$ ,  $\|f_n - f_m\|_{\mathcal{F}} < \epsilon$ .

Every convergent sequence is Cauchy (from the triangle inequality), but the converse is not true. If the limit of the Cauchy sequence exists within the vector space, then the sequence converges to it. If the vector space contains the limits of all Cauchy sequences (or in other words, if every Cauchy sequence converges), then it is said to be *complete*.

There are special names given to complete vector spaces. A complete inner product space is known as a *Hilbert space*, while a complete normed space is called a *Banach space*. Out of interest, an inner product space that is not complete is sometimes known as a *pre-Hilbert space*, since its completion with respect to the norm induced by the inner product is a Hilbert space.

A subset  $\mathcal{G} \subseteq \mathcal{F}$  is a *closed subspace* of  $\mathcal{F}$  if it is closed under addition and multiplication by a scalar. That is, for any  $g, g' \in \mathcal{G}$ ,  $\lambda_1 g + \lambda_2 g'$  is also in  $\mathcal{G}$ . For Hilbert spaces, each closed subspace is also complete, and thus a Hilbert space in its own right. Although, as a remark, not every Hilbert subspace need be closed, and therefore complete.

Being vectors in a vector space, we can discuss mapping the vectors onto a different space, or in essence, having a function acted upon them. To establish terminology, we define linear functionals, bilinear form, and linear operators.

**Definition 2.5** (Linear functional). Let  $\mathcal{F}$  be a Hilbert space. A *functional*  $L$  is a map from  $\mathcal{F}$  to  $\mathbb{R}$ , and we denote its action on a function  $f$  as  $L(f)$ . A functional is called *linear* if it satisfies  $L(f + g) = L(f) + L(g)$  and  $L(\lambda f) = \lambda L(f)$ , for all  $f, g \in \mathcal{F}$  and  $\lambda \in \mathbb{R}$ .

**Definition 2.6** (Bilinear form). Let  $\mathcal{F}$  be a Hilbert space. A *bilinear form*  $B$  takes inputs  $f, g \in \mathcal{F}$  and returns a real value. It is linear in each argument separately, i.e.

- $B(\lambda_1 f + \lambda_2 g, h) = \lambda_1 B(f, h) + \lambda_2 B(g, h)$ ; and
- $B(f, \lambda_1 g + \lambda_2 h) = \lambda_1 B(f, g) + \lambda_2 B(f, h)$ ,

for all  $f, g, h \in \mathcal{F}$  and  $\lambda_1, \lambda_2 \in \mathbb{R}$ .

**Definition 2.7** (Linear operator). Let  $\mathcal{F}$  and  $\mathcal{G}$  be two Hilbert spaces over  $\mathbb{R}$ . An operator  $A$  is a map from  $\mathcal{F}$  to  $\mathcal{G}$ , and we denote its action on a function  $f \in \mathcal{F}$  as  $Af \in \mathcal{G}$ . A *linear operator* satisfies  $A(f + g) = A(f) + A(g)$  and  $A(\lambda f) = \lambda A(f)$ , for all  $f, g \in \mathcal{F}$  and  $\lambda \in \mathbb{R}$ .

The term ‘functional’ is classically used in calculus of variations to denote ‘a function of a function’, i.e. a function having another function as its input, and outputs a real number. Really, from a function space perspective, it is simply a mapping of functions onto another vector space (the reals in this case). More generally, if the output space is another Hilbert space, then it is an operator. An interesting property of these operators to look at, besides linearity, is whether or not they are *continuous*.

def:continuity

**Definition 2.8** (Continuity). Let  $\mathcal{F}$  and  $\mathcal{G}$  be two Hilbert spaces. A function  $A : \mathcal{F} \rightarrow \mathcal{G}$  is said to be *continuous at*  $g \in \mathcal{F}$ , if for every  $\epsilon > 0$ ,  $\exists \delta = \delta(\epsilon, g) > 0$  such that

$$\|f - g\|_{\mathcal{F}} < \delta \Rightarrow \|Af - Ag\|_{\mathcal{G}} < \epsilon.$$

$A$  is *continuous* on  $\mathcal{F}$ , if it is continuous at every point  $g \in \mathcal{F}$ . If, in addition,  $\delta$  depends on  $\epsilon$  only,  $A$  is said to be *uniformly continuous*.

Continuity in the sense of linear operators here means that a convergent sequence in  $\mathcal{F}$  can be mapped to a convergent sequence in  $\mathcal{G}$ . For a special case of linear operator, the evaluation functional, this means that a function in  $\mathcal{F}$  is continuous if the evaluation functional is continuous—more on this later in [Section 2.2](#). There is an even stronger notion of continuity called the *Lipschitz continuity*.

**Definition 2.9** (Lipschitz continuity). Let  $\mathcal{F}$  and  $\mathcal{G}$  be two Hilbert spaces. A function  $A : \mathcal{F} \rightarrow \mathcal{G}$  is *Lipschitz continuous* if  $\exists M > 0$  such that  $\forall f, f' \in \mathcal{F}$ ,

$$\|Af - Af'\|_{\mathcal{G}} \leq M\|f - f'\|_{\mathcal{F}}.$$

Clearly, Lipschitz continuity implies uniform continuity: choose  $\delta = \delta(\epsilon) := \epsilon/M$  and replace this in [Definition 2.8](#). A continuous, linear operator is also one that is bounded:

def:bounded op

**Definition 2.10** (Bounded operator). The linear operator  $A : \mathcal{F} \rightarrow \mathcal{G}$  between two Hilbert spaces  $\mathcal{F}$  and  $\mathcal{G}$  is said to be *bounded* if there exists some  $M > 0$  such that

$$\|Af\|_{\mathcal{G}} \leq M\|f\|_{\mathcal{F}}.$$

thm:bounded  
nt

The smallest such  $M$  is defined to be the *operator norm*, denoted  $\|A\| := \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$ .

**Lemma 2.1** (Equivalence of boundedness and continuity). *Let  $\mathcal{F}$  and  $\mathcal{G}$  be two Hilbert spaces, and  $A : \mathcal{F} \rightarrow \mathcal{G}$  a linear operator.  $A$  is a bounded if and only if it is continuous.*

*Proof.* Suppose that  $L$  is bounded. Then,  $\forall f, f' \in \mathcal{F}$ , there exists some  $M > 0$  such that  $\|A(f - g)\|_{\mathcal{G}} \leq M\|f - g\|_{\mathcal{G}}$ . Conversely, let  $A$  be a continuous linear operator, especially at the zero vector. In other words,  $\exists \delta > 0$  such that  $\|A(f)\|_{\mathcal{G}} = \|A(f + 0 - 0)\|_{\mathcal{G}} = \|A(f) - A(0)\|_{\mathcal{G}} \leq 1$ ,  $\forall f \in \mathcal{F}$  whenever  $\|f\|_{\mathcal{F}} \leq \delta$ . Thus, for all non-zero  $f \in \mathcal{F}$ ,

$$\begin{aligned} \|A(f)\|_{\mathcal{G}} &= \left\| \frac{\|f\|_{\mathcal{F}}}{\delta} A\left(\frac{\delta}{\|f\|_{\mathcal{F}}} f\right) \right\|_{\mathcal{G}} \\ &= \left| \frac{\|f\|_{\mathcal{F}}}{\delta} \right| \cdot \left\| A\left(\frac{\delta}{\|f\|_{\mathcal{F}}} f\right) \right\|_{\mathcal{G}} \\ &\leq \frac{\|f\|_{\mathcal{F}}}{\delta} \cdot 1, \end{aligned}$$

and thus  $A$  is bounded. □

So important is the concept of linearity and continuity, that there are specially named spaces which contain linear and continuous functionals.

**Definition 2.11** (Dual spaces). Let  $\mathcal{F}$  be a Hilbert space. The space  $\mathcal{F}^*$  of *linear functionals* is called the *algebraic dual space* of  $\mathcal{F}$ . The space  $\mathcal{F}'$  of *continuous linear functionals* is called the *continuous dual space* or alternatively, the *topological dual space*, of  $\mathcal{F}$ .

As it turns out, the algebraic dual space and continuous dual space coincide in finite-dimensional Hilbert spaces: take any  $L \in \mathcal{F}'$ ; since  $L$  is finite-dimensional, it is bounded, and therefore continuous (see [Lemma 2.1](#)) so  $L \in \mathcal{F}'$  and  $\mathcal{F}^* \subseteq \mathcal{F}'$ ; but  $\mathcal{F}' \subseteq \mathcal{F}^*$  trivially, so  $\mathcal{F}^* \equiv \mathcal{F}'$ . For infinite-dimensional Hilbert spaces, this is not so, but in any case, we will only be considering the continuous dual space in this thesis. The following result is an important one, which states that (continuous) linear functionals of an inner product space are nothing more than just inner products.

**Theorem 2.2** (Riesz representation). *Let  $\mathcal{F}$  be a Hilbert space. Every element  $L$  of the continuous dual space  $\mathcal{F}'$ , i.e. all continuous linear functionals  $L : \mathcal{F} \rightarrow \mathbb{R}$ , can be uniquely written in the form  $L = \langle \cdot, g \rangle_{\mathcal{F}}$ , for some  $g \in \mathcal{F}$ .*

*Proof.* Omitted—see [rudin1987real](#) for a proof. □

The notion of isometry (transformation that preserves distance) is usually associated with metric spaces—two metric spaces being isometric means that they are identical in as far as their metric properties are concerned. For Hilbert spaces (or normed spaces in general), there is an analogous concept as well in *isometric isomorphism* (a bijective isometry), such that two Hilbert spaces being isometrically isomorphic imply that they have exactly the same geometric structure, but may very well contain fundamentally different objects.

**Definition 2.12** (Isometric isomorphism). Two Hilbert spaces  $\mathcal{F}$  and  $\mathcal{G}$  are said to be *isometrically isomorphic* if there is a linear bijective map  $A : \mathcal{F} \rightarrow \mathcal{G}$  which preserves the inner product, i.e.

$$\langle f, f' \rangle_{\mathcal{F}} = \langle Af, Af' \rangle_{\mathcal{G}}.$$

A consequence of the Riesz representation theorem is that it gives us a canonical isometric isomorphism  $A : f \mapsto \langle \cdot, f \rangle_{\mathcal{F}}$  between  $\mathcal{F}$  and its continuous dual  $\mathcal{F}'$ , whereby  $\|Af\|_{\mathcal{F}'} = \|f\|_{\mathcal{F}}$ . Implicitly, this means that  $\mathcal{F}'$  is a Hilbert space as well.

Another important type of mapping is the mapping  $P$  of an element in  $\mathcal{F}$  onto a closed subspace  $\mathcal{G} \subset \mathcal{F}$ , such that  $Pf \in \mathcal{G}$  is closest to  $f$ . This mapping is called the *orthogonal projection*, due to the fact that such projections yield perpendicularity in the sense that  $\langle f - Pf, g \rangle_{\mathcal{G}} = 0$  for any  $g \in \mathcal{G}$ . The remainder  $f - Pf$  belongs to the *orthogonal complement* of  $\mathcal{G}$ .

**Definition 2.13** (Orthogonal complement). Let  $\mathcal{F}$  be a Hilbert space and  $\mathcal{G} \subset \mathcal{F}$  be a closed subspace. The linear subspace  $\mathcal{G}^{\perp} = \{f \mid \langle f, g \rangle_{\mathcal{G}} = 0, \forall g \in \mathcal{G}\}$  is called the orthogonal complement of  $\mathcal{G}$ .

**Theorem 2.3** (Orthogonal decomposition). *Let  $\mathcal{F}$  be a Hilbert space and  $\mathcal{G} \subset \mathcal{F}$  be a closed subspace. For every  $f \in \mathcal{F}$ , we can write  $f = g + g^c$ , where  $g \in \mathcal{G}$  and  $g^c \in \mathcal{G}^{\perp}$ , and this decomposition is unique.*

*Proof.* Omitted—see [rudin1987real](#) for a proof. □

We can write  $\mathcal{F} = \mathcal{G} \oplus \mathcal{G}^{\perp}$ , where the  $\oplus$  symbol denotes the *direct sum*, and such a decomposition is called a *tensor sum decomposition*. In infinite-dimensional Hilbert spaces, some subspaces are not closed, but all orthogonal complements are closed. In such spaces, the orthogonal complement of the orthogonal complement of  $\mathcal{G}$  is the closure of  $\mathcal{G}$ , i.e.  $(\mathcal{G}^{\perp})^{\perp} =: \overline{\mathcal{G}}$ , and we say that  $\mathcal{G}$  is dense in  $\overline{\mathcal{G}}$ . Another interesting fact regarding



thm:orthdec  
omp2

the orthogonal complement is that  $\mathcal{G} \cap \mathcal{G}^\perp = \{0\}$ , since any  $g \in \mathcal{G} \cap \mathcal{G}^\perp$  must be orthogonal to itself, i.e.  $\langle g, g \rangle_{\mathcal{G}} = 0$  implying that  $g = 0$ .

**Corollary 2.3.1.** *Let  $\mathcal{G}$  be a subspace of a Hilbert space  $\mathcal{F}$ . Then,  $\mathcal{G}^\perp = \{0\}$  if and only if  $\mathcal{G}$  is dense in  $\mathcal{F}$ .*

*Proof.* If  $\mathcal{G}^\perp = \{0\}$  then  $(\mathcal{G}^\perp)^\perp = \overline{\mathcal{G}} = \mathcal{F}$ . Conversely, since  $\mathcal{G}$  is dense in  $\mathcal{F}$ , we have  $\mathcal{G}^\perp = \overline{\mathcal{G}}^\perp = \mathcal{F}^\perp = \{0\}$ .  $\square$

Besides tensor sums, of importance is the concept of *tensor products*, which can be thought of as a generalisation of the outer product in Euclidean space.

**Definition 2.14** (Tensor products). Let  $x_1 \in \mathcal{H}_1$  and  $x_2 \in \mathcal{H}_2$  be two elements of two real Hilbert spaces. Then, the tensor product  $x_1 \otimes x_2 : \mathcal{H}_1 \times \mathcal{H}_2 \rightarrow \mathbb{R}$ , is a bilinear form defined as

$$(x_1 \otimes x_2)(y_1, y_2) = \langle x_1, y_1 \rangle_{\mathcal{H}_1} \langle x_2, y_2 \rangle_{\mathcal{H}_2}$$

for any  $(y_1, y_2) \in \mathcal{H}_1 \times \mathcal{H}_2$ .

Correspondingly, we may also define the *tensor product space*.

**Definition 2.15** (Tensor product space). The tensor product space  $\mathcal{H}_1 \otimes \mathcal{H}_2$  is the completion of the space

$$\mathcal{A} = \left\{ \sum_{j=1}^J x_{1j} \otimes x_{2j} \mid x_{1j} \in \mathcal{H}_1, x_{2j} \in \mathcal{H}_2, J \in \mathbb{N} \right\}.$$

with respect to the norm induced by the inner product

$$\left\langle \sum_{j=1}^J x_{1j} \otimes x_{2j}, \sum_{k=1}^K y_{1k} \otimes y_{2k} \right\rangle_{\mathcal{A}} = \sum_{j=1}^J \sum_{k=1}^K \langle x_{1j}, y_{1k} \rangle_{\mathcal{H}_1} \langle x_{2j}, y_{2k} \rangle_{\mathcal{H}_2}.$$

Interestingly, the tensor product can be viewed as an operator between two Hilbert spaces. That is, for each pair of elements  $(x_1, x_2) \in \mathcal{H}_1 \times \mathcal{H}_2$ , we define the operator  $A : \mathcal{H}_1 \rightarrow \mathcal{H}_2$  in the following way:

$$\begin{aligned} A_{x_1, x_2} : \mathcal{H}_1 &\rightarrow \mathcal{H}_2 \\ y_1 &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2 \end{aligned}$$

For some  $y_1 \in \mathcal{H}_1$  and  $y_2 \in \mathcal{H}_2$ , we have that

$$\begin{aligned}\langle A_{x_1, x_2}(y_1), y_2 \rangle_{\mathcal{H}_2} &= \langle \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2, y_2 \rangle_{\mathcal{H}_2} \\ &= \langle x_1, y_1 \rangle_{\mathcal{H}_1} \langle x_2, y_2 \rangle_{\mathcal{H}_2} \\ &= (x_1 \otimes x_2)(y_1, y_2).\end{aligned}$$

It is seen that the tensor product  $x_1 \otimes x_2$  is associated with the rank one operator  $B : \mathcal{H}'_1 \rightarrow \mathcal{H}_2$  defined by  $z \mapsto z(x_1)x_2$  with  $z = \langle x_1, \cdot \rangle_{\mathcal{H}_1}$ . We write  $B = x_1 \otimes x_2$ .

Therefore, this extends a linear identification between  $\mathcal{H}_1 \otimes \mathcal{H}_2$  and the space of finite-rank operators from  $\mathcal{H}'_1$  to  $\mathcal{H}_2$ . We now have three distinct interpretations of the tensor product:

- **Bilinear form** (as defined in Definition 3.5).

$$\begin{aligned}x_1 \otimes x_2 : \mathcal{H}_1 \times \mathcal{H}_2 &\rightarrow \mathbb{R} \\ (y_1, y_2) &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} \langle x_2, y_2 \rangle_{\mathcal{H}_2}\end{aligned}$$

for  $x_1, y_1 \in \mathcal{H}_1$  and  $x_2, y_2 \in \mathcal{H}_2$ .

- **Operator**.

$$\begin{aligned}x_1 \otimes x_2 : \mathcal{H}_1 &\rightarrow \mathcal{H}_2 \\ y_1 &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2\end{aligned}$$

- **General form** (as an element in the tensor space).

$$x_1 \otimes x_2 \in \mathcal{H}_1 \otimes \mathcal{H}_2.$$

For the last part of this introductory section on functional analysis, we discuss measures on Hilbert spaces, and in particular, a probability measure. Let  $\mathcal{H}$  be a real Hilbert space. As discussed earlier, we can define a metric on  $\mathcal{H}$  using  $D(x, x') = \|x - x'\|_{\mathcal{H}}$ , where the norm on  $\mathcal{H}$  is the norm induced by the inner product. A collection  $\Sigma$  of subsets of  $\mathcal{H}$  is called a  $\sigma$ -algebra if  $\emptyset \in \Sigma$ ,  $S \in \Sigma$  implies its complement  $S^c \in \Sigma$ , and  $S_j \in \Sigma$ ,  $j \geq 1$  implies  $\bigcup_{j=1}^{\infty} S_j \in \Sigma$ . The smallest  $\sigma$ -algebra containing all open subsets of  $\mathcal{H}$  is called the *Borel  $\sigma$ -algebra*, and its members the Borel sets. Denote by  $\mathcal{B}(\mathcal{H})$  the Borel  $\sigma$ -algebra of  $\mathcal{H}$ .

1. From Wikipedia. But don't really get it, although it might explain the Fisher information between linear functionals.

Recall that a function  $\nu : \Sigma \rightarrow [0, \infty]$  is called a *measure* if it satisfies

- **Non-negativity:**  $\nu(S) \geq 0$  for all  $S$  in  $\Sigma$ ;
- **Null empty set:**  $\nu(\emptyset) = 0$ ; and
- **$\sigma$ -additivity:** for all countable, mutually disjoint sets  $\{S_i\}_{i=1}^{\infty}$ ,

$$\nu\left(\bigcup_{i=1}^{\infty} S_i\right) = \sum_{i=1}^{\infty} \nu(S_i).$$

A measure  $\nu$  on  $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$  is called a *Borel measure* on  $\mathcal{H}$ . We shall only concern ourselves with finite Borel measures. In addition, if  $\nu(\mathcal{H}) = 1$  then  $\nu$  is a (*Borel*) *probability measure* and the measure space  $(\mathcal{H}, \mathcal{B}(\mathcal{H}), \nu)$  is a (*Borel*) *probability space*.

Let  $(\Omega, \mathcal{E}, P)$  be a probability space. We say that a mapping  $X : \Omega \rightarrow \mathcal{H}$  is a *random element* in  $\mathcal{H}$  if  $X^{-1}(B) \in \mathcal{E}$  for every Borel set, i.e.,  $X$  is a function such that for every  $B \in \mathcal{B}(\mathcal{H})$ , its preimage  $X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$  lies in  $\Sigma$ . This is simply a generalisation of the definition of random variables in regular Euclidean space. From this definition, we can also properly define random functions  $f$  in a Hilbert space of functions  $\mathcal{F}$ . In any case, every random element  $X$  induces a probability measure on  $\mathcal{H}$  defined by

$$\nu(B) = P(X^{-1}(B)) = P(\omega \in \Omega \mid X(\omega) \in B) = P(X \in B).$$

The measure  $\nu$  is called the *distribution* of  $X$ . The *density*  $p$  of  $X$  is a measurable function with the property that

$$P(X \in B) = \int_{X^{-1}(B)} \omega \, dP(\omega) = \int_B p(x) \, d\nu(x).$$

**Definition 2.16** (Mean vector). Let  $\nu$  be a Borel probability measure on a real Hilbert space  $\mathcal{H}$ . Supposing that a random element  $X$  of  $\mathcal{H}$  is *integrable*, that is to say

$$E\|X\|_{\mathcal{H}} = \int_{\mathcal{H}} \|x\|_{\mathcal{H}} \, d\nu(x) < \infty,$$

then the unique element  $\mu \in \mathcal{H}$  satisfying

$$\langle \mu, x' \rangle = \int_{\mathcal{H}} \langle x, x' \rangle_{\mathcal{H}} \, d\nu(x) = E\langle X, x' \rangle_{\mathcal{H}}$$

for all  $x' \in \mathcal{H}$  is called the *mean vector*.

**Definition 2.17** (Covariance operator). Let  $\nu$  be a Borel probability measure on a real Hilbert space  $\mathcal{H}$ . Suppose that a random element  $X$  of  $\mathcal{H}$  is *square integrable*, i.e.,  $\mathbb{E}\|X\|_{\mathcal{H}}^2 < \infty$ , and let  $\mu$  be the mean vector of  $X$ . Then the *covariance operator*  $C$  is defined by the mapping

$$\begin{aligned} C : \mathcal{H} &\rightarrow \mathcal{H} \\ x &\mapsto \mathbb{E} [\langle X - \mu, x \rangle_{\mathcal{H}} (X - \mu)]. \end{aligned}$$

The covariance operator  $C$  is also an element of  $\mathcal{H} \otimes \mathcal{H}$  that satisfies

$$\begin{aligned} \langle C, x \otimes x' \rangle_{\mathcal{H} \otimes \mathcal{H}} &= \int_{\mathcal{H}} \langle z - \mu, x \rangle_{\mathcal{H}} \langle z - \mu, x' \rangle_{\mathcal{H}} d\nu(z) \\ &= \mathbb{E} [\langle X - \mu, x \rangle_{\mathcal{H}} \langle X - \mu, x' \rangle_{\mathcal{H}}] \end{aligned}$$

for all  $x, x' \in \mathcal{H}$ .

From the definition of the covariance operator, we see that it induces a symmetric, bilinear form, which we shall denote by  $\text{Cov} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ , through

$$\begin{aligned} \langle Cx, x' \rangle_{\mathcal{H}} &= \langle \mathbb{E} [\langle X - \mu, x \rangle_{\mathcal{H}} (X - \mu)], x' \rangle_{\mathcal{H}} \\ &= \mathbb{E} [\langle X - \mu, x \rangle_{\mathcal{H}} \langle X - \mu, x' \rangle_{\mathcal{H}}] \\ &=: \text{Cov}(x, x'). \end{aligned}$$

**Definition 2.18** (Gaussian vectors). A random element  $X$  is called *Gaussian* if  $\langle X, x \rangle_{\mathcal{H}}$  has a normal distribution for all fixed  $x \in \mathcal{H}$ . A Gaussian vector  $X$  is characterised by its mean element  $\mu \in \mathcal{H}$  and its covariance  $C \in \mathcal{H} \otimes \mathcal{H}$ .

## 2.2 Reproducing kernel Hilbert space theory

The introductory section sets us up nicely to discuss the coveted reproducing kernel Hilbert space. This is a subset of Hilbert spaces for which its evaluation functionals are continuous (by definition, in fact). The majority of this section, apart from defining RKHS, is to convince ourselves that each and every RKHS of functions can be specified solely through its reproducing kernel. To begin, we consider a fundamental linear functional on a Hilbert space of functions  $\mathcal{F}$ , that assigns a value to  $f \in \mathcal{F}$  for each  $x \in \mathcal{X}$ .

sec:rkhs  
theory

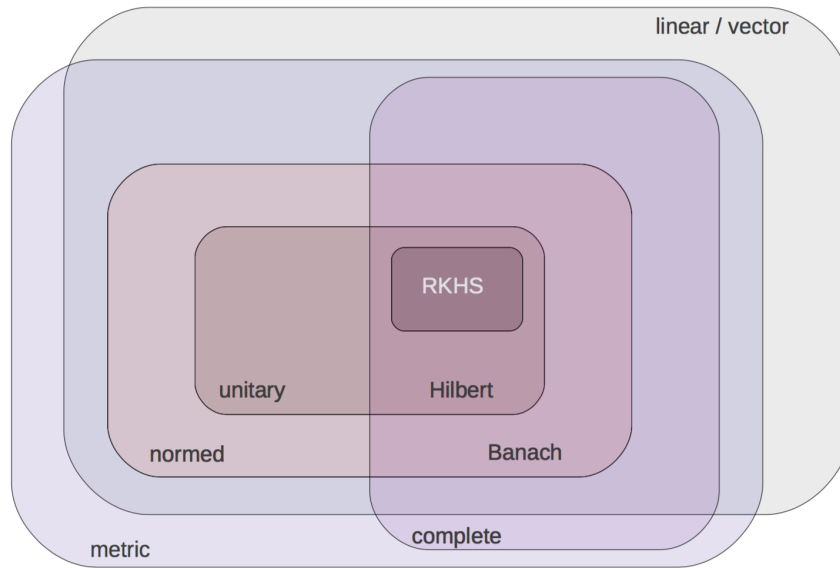


Figure 2.1: A hierarchy of vector spaces<sup>2</sup>.

**Definition 2.19** (Evaluation functional). Let  $\mathcal{F}$  be a vector space of functions  $f : \mathcal{X} \rightarrow \mathbb{R}$ , defined on a non-empty set  $\mathcal{X}$ . For a fixed  $x \in \mathcal{X}$ , the functional  $\delta_x : \mathcal{F} \rightarrow \mathbb{R}$  as defined by  $\delta_x(f) = f(x)$  is called the (Dirac) evaluation functional at  $x$ .

It is easy to see that evaluation functionals are always linear:  $\delta_x(\lambda f + g) = (\lambda f + g)(x) = \lambda f(x) + g(x) = \lambda \delta_x(f) + \delta_x(g)$ . This is in fact the linearity that was implied earlier on at the beginning of Chapter 2 when introducing the notion of functions behaving like vectors. As a remark, the calculation of the (penalised) likelihood functional involves evaluations. It is therefore important for the evaluation functional to be continuous. It turns out, this is exactly what RKHS provide.

**Definition 2.20** (Reproducing kernel Hilbert space). A Hilbert space  $\mathcal{F}$  of real-valued functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  on a non-empty set  $\mathcal{X}$  is called a *reproducing kernel Hilbert space* if the evaluation functional  $\delta_x : f \mapsto f(x)$  is continuous (equivalently, bounded) on  $\mathcal{F}$ ,  $\forall x \in \mathcal{X}$ .

Continuity of evaluation functionals in an RKHS means that functions that are close in RKHS norm imply that they are also close pointwise, but the converse is not neces-

<sup>2</sup>Reproduced from the lecture slides of Dino Sejdinovic and Arthur Gretton entitled ‘Foundations of Reproducing Kernel Hilbert Spaces: Advanced Topics in Machine Learning’, 2014. URL: [http://www.stats.ox.ac.uk/~sejdinovic/teaching/atml14/Theory\\_slides2\\_2014.pdf](http://www.stats.ox.ac.uk/~sejdinovic/teaching/atml14/Theory_slides2_2014.pdf).

thm:normpoi  
ntconv

sarily true. This gives some reassurance when trying to estimate  $f$  from  $\mathcal{F}$  using the norm of  $\mathcal{F}$  as a criterion for selection. More formally,

**Corollary 2.3.2** (Norm convergence implies pointwise convergence in RKHS). *Let  $\mathcal{F}$  be an RKHS of real functions over  $\mathcal{X}$ , and let  $f_n$  be a sequence of points in  $\mathcal{F}$ . Then, for some  $f \in \mathcal{F}$ ,*

$$\lim_{n \rightarrow \infty} \|f_n - f\|_{\mathcal{F}} = 0 \quad \Rightarrow \quad \lim_{n \rightarrow \infty} |f_n(x) - f(x)| = 0.$$

*Proof.* Suppose  $\mathcal{F}$  is an RKHS with reproducing kernel  $h$ . Then,

$$\begin{aligned} |\delta_x(f) - \delta_x(g)| &= |\delta_x(f - g)| \\ &= |(f - g)(x)| \\ &= |\langle f - g, h(\cdot, x) \rangle_{\mathcal{F}}| \quad (\text{reproducing property}) \\ &\leq \|h(\cdot, x)\|_{\mathcal{F}} \cdot \|f - g\|_{\mathcal{F}} \quad (\text{by Cauchy-Schwarz}) \\ &= \sqrt{h(x, x)} \cdot \|f - g\|_{\mathcal{F}}. \end{aligned}$$

□

Insert figure squiggly line and smooth line.

While the continuity condition by definition is what makes an RKHS, it is neither easy to check this condition in practice, nor is it intuitive as to the meaning of its name. In fact, there isn't even any mention of what a reproducing kernel actually is. In order to benefit from the desirable continuity property of RKHS, we should look at this from another, more intuitive, perspective. By invoking the Riesz representation theorem, we see that for all  $x \in \mathcal{X}$ , there exists a unique element  $h_x \in \mathcal{F}$  such that

$$f(x) = \delta_x(f) = \langle f, h_x \rangle_{\mathcal{F}}, \forall f \in \mathcal{F}$$

holds. Since  $h_x$  itself is a function in  $\mathcal{F}$ , it holds that for every  $x' \in \mathcal{X}$  there exists a  $h_{x'} \in \mathcal{F}$  such that

$$h_x(x') = \delta_{x'}(h_x) = \langle h_x, h_{x'} \rangle_{\mathcal{F}}.$$

This leads us to the definition of a *reproducing kernel* of an RKHS—the very notion that inspired its name.

def:repkern

**Definition 2.21** (Reproducing kernels). Let  $\mathcal{F}$  be a Hilbert space of functions over a non-empty set  $\mathcal{X}$ . A symmetric, bivariate function  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is called a *kernel*, and it is a *reproducing kernel* of  $\mathcal{F}$  if  $h$  satisfies

- $\forall x \in \mathcal{X}, h(\cdot, x) \in \mathcal{F}$ ; and
- $\forall x \in \mathcal{X}, f \in \mathcal{F}, \langle f, h(\cdot, x) \rangle_{\mathcal{F}} = f(x)$  (the reproducing property).

In particular, for any  $x, x' \in \mathcal{X}$ ,

$$h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}}.$$

An important property for reproducing kernels of a RKHS is that they are positive-definite functions. That is,  $\forall \lambda_1, \dots, \lambda_n \in \mathbb{R}$  and  $\forall x_1, \dots, x_n \in \mathcal{X}$ ,

$$\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j h(x_i, x_j) \geq 0.$$

thm:posdef

**Claim 2.4** (Reproducing kernels of RKHS are positive-definite). *Let  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be a reproducing kernel for a Hilbert space  $\mathcal{F}$ . Then  $h$  is a symmetric and positive definite function.*

*Proof.*

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j h(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \langle h(\cdot, x_i), h(\cdot, x_j) \rangle_{\mathcal{F}} \\ &= \left\langle \sum_{i=1}^n \lambda_i h(\cdot, x_i), \sum_{j=1}^n \lambda_j h(\cdot, x_j) \right\rangle_{\mathcal{F}} \\ &= \left\| \sum_{i=1}^n \lambda_i h(\cdot, x_i) \right\|_{\mathcal{F}}^2 \\ &\geq 0 \end{aligned}$$

□

*Remark 2.1.* In the kernel method literature, a *kernel*  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is usually defined as the inner product between inputs in feature space. That is, take  $\phi : \mathcal{X} \rightarrow \mathcal{V}$ ,  $x \mapsto \phi(x)$ , where  $\mathcal{V}$  is a Hilbert space. Then the kernel is defined as  $h(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}$ , for any  $x, x' \in \mathcal{X}$ .  $\mathcal{V}$  is known as the *feature space* and  $\phi$  the *feature map*. In many

mathematical models involving feature space mappings, elucidation of the feature map and feature space is not necessary, and computation is made simpler by the use of kernels (known as the *kernel trick*). Note that kernels defined in this manner are positive definite, while in this thesis, we opt for a more general definition allowing for non-positive kernels.

Introducing the following definition of the *kernel matrix* (also known as the *Gram matrix*) is useful at this point.

**Definition 2.22** (Kernel matrix). Let  $\{x_1, \dots, x_n\}$  be a sample of points, where each  $x_i \in \mathcal{X}$ , and  $h$  a kernel over  $\mathcal{X}$ . Define the *kernel matrix*  $\mathbf{H}$  for  $h$  as the  $n \times n$  matrix with  $(i, j)$  entries equal to  $h(x_i, x_j)$ .

Now, one might ask what the relationship between a reproducing kernel and a RKHS is. We assert the following:

- **RKHS  $\Leftrightarrow$  reproducing kernel.** For every RKHS  $\mathcal{F}$  of functions over a set  $\mathcal{X}$ , there corresponds a unique, positive-definite reproducing kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , and vice-versa.
- **P.d. function  $\Rightarrow$  RKHS.** For every positive-definite function  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , there corresponds a unique RKHS  $\mathcal{F}$  that has  $h$  as its reproducing kernel.

In essence, there is a bijection between the set of positive-definite kernels and the set of reproducing kernel Hilbert spaces. The rest of this subsection is a discussion of this assertion, which is proven by the two theorems that follow.

**Theorem 2.5** (RKHS uniqueness). *Let  $\mathcal{F}$  be a Hilbert space of functions over  $\mathcal{X}$ .  $\mathcal{F}$  is a RKHS if and only if  $\mathcal{F}$  has a reproducing kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , and that  $h$  is unique to  $\mathcal{F}$ .*

*Proof.* First we tackle existence, i.e., we prove that  $\mathcal{F}$  is a RKHS if and only if  $\mathcal{F}$  has a reproducing kernel. Suppose  $\mathcal{F}$  is a Hilbert space of functions, and  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a reproducing kernel for  $\mathcal{F}$ . Then, choosing  $\delta = \epsilon / \|h(\cdot, x)\|_{\mathcal{F}}$ , for any  $f \in \mathcal{F}$  such that  $\|f - g\|_{\mathcal{F}} < \delta$ , we have

$$\begin{aligned}
 |\delta_x(f) - \delta_x(g)| &= |(f - g)(x)| \\
 &= |\langle f - g, h(\cdot, x) \rangle_{\mathcal{F}}| \quad (\text{reproducing property}) \\
 &\leq \|h(\cdot, x)\|_{\mathcal{F}} \cdot \|f - g\|_{\mathcal{F}} \quad (\text{by Cauchy-Schwarz}) \\
 &= \epsilon.
 \end{aligned}$$

thm:rkhsuni  
que



Thus, the evaluation functional is (uniformly) continuous on  $\mathcal{F}$ , and by definition,  $\mathcal{F}$  is a RKHS. Now suppose that  $\mathcal{F}$  is a RKHS, and  $h$  is a kernel function over  $\mathcal{X} \times \mathcal{X}$ . The reproducing property of  $h$  is had by following the argument preceding [Definition 2.21](#).

As for uniqueness, assume that the RKHS  $\mathcal{F}$  has two reproducing kernels  $h_1$  and  $h_2$ . Then,  $\forall f \in \mathcal{F}$  and  $\forall x \in \mathcal{X}$ ,

$$\langle f, h_1(\cdot, x) - h_2(\cdot, x) \rangle_{\mathcal{F}} = f(x) - f(x) = 0.$$

In particular, if we take  $f = h_1(\cdot, x) - h_2(\cdot, x)$ , we obtain  $\|h_1(\cdot, x) - h_2(\cdot, x)\|_{\mathcal{F}}^2 = 0$ . Thus,  $h_1 = h_2$ .  $\square$

**Theorem 2.6** (Moore-Aronszajn). *If  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is a positive-definite function then there exists a unique RKHS whose reproducing kernel is  $h$ .*

*Sketch proof.* Most of the details here have been omitted, except for the parts which we feel are revealing as to the properties of an RKHS. For a complete proof, see [berlinet2011reproducing](#). Start with the linear space

$$\mathcal{F}_0 = \left\{ f_n : \mathcal{X} \rightarrow \mathbb{R} \mid f_n = \sum_{i=1}^n w_i h(\cdot, x_i), x_i \in \mathcal{X}, w_i \in \mathbb{R}, n \in \mathbb{N} \right\}$$

and endow this linear space with the following inner product:

$$\left\langle \sum_{i=1}^n w_i h(\cdot, x_i), \sum_{j=1}^m w'_j h(\cdot, x'_j) \right\rangle_{\mathcal{F}_0} = \sum_{i=1}^n \sum_{j=1}^m w_i w'_j h(x_i, x'_j).$$

It may be shown that this indeed a valid inner-product satisfying the conditions laid in [Definition 2.1](#). At this point, the reproducing property is already had:

$$\begin{aligned} \langle f_n, h(\cdot, x) \rangle_{\mathcal{F}_0} &= \left\langle \sum_{i=1}^n w_i h(\cdot, x_i), h(\cdot, x) \right\rangle_{\mathcal{F}_0} \\ &= \sum_{i=1}^n w_i h(x_i, x) \\ &= f_n(x), \end{aligned}$$

for any  $f_n \in \mathcal{F}_0$ .

thm:moorea

Let  $\mathcal{F}$  be the completion of  $\mathcal{F}_0$  with respect to this inner product. In other words, define  $\mathcal{F}$  to be the set of functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  for which there exists a Cauchy sequence  $\{f_n\}_{n=1}^\infty$  in  $\mathcal{F}_0$  converging pointwise to  $f \in \mathcal{F}$ . The inner product for  $\mathcal{F}$  is defined to be

$$\langle f, f' \rangle_{\mathcal{F}} = \lim_{n \rightarrow \infty} \langle f_n, f'_n \rangle_{\mathcal{F}_0}.$$

The sequence  $\{\langle f_n, f'_n \rangle_{\mathcal{F}_0}\}_{n=1}^\infty$  is convergent and does not depend on the sequence chosen, but only on the limits  $f$  and  $f'$  ([berlinet2011reproducing](#)). We may check that this indeeds defines a valid inner product. The reproducing property carries over to the completion:

$$\begin{aligned} \langle f, h(\cdot, x) \rangle_{\mathcal{F}} &= \lim_{n \rightarrow \infty} \langle f_n, h(\cdot, x) \rangle_{\mathcal{F}_0} \\ &= \lim_{n \rightarrow \infty} f_n(x) \\ &= f(x). \end{aligned}$$

To prove uniqueness, let  $\mathcal{G}$  be another RKHS with reproducing kernel  $h$ .  $\mathcal{F}$  has to be a closed subspace of  $\mathcal{G}$ , since  $h(\cdot, x) \in \mathcal{G}$  for all  $x \in \mathcal{X}$ , and because  $\mathcal{G}$  is complete and contains  $\mathcal{F}_0$  and hence its completion. Using the orthogonal decomposition theorem, we have  $\mathcal{G} = \mathcal{F} \oplus \mathcal{F}^\perp$ , i.e. any  $g \in \mathcal{G}$  can be decomposed as  $g = f + f^c$ ,  $f \in \mathcal{F}$  and  $f^c \in \mathcal{F}^\perp$ . For each element  $g \in \mathcal{G}$  we have that, for all  $x \in \mathcal{X}$ ,

$$\begin{aligned} g(x) &= \langle g, h(\cdot, x) \rangle_{\mathcal{G}} \\ &= \langle f + f^c, h(\cdot, x) \rangle_{\mathcal{G}} \\ &= \langle f, h(\cdot, x) \rangle_{\mathcal{G}} + \underbrace{\langle f^c, h(\cdot, x) \rangle_{\mathcal{G}}}_{\rightarrow 0} \\ &= f(x) \end{aligned}$$

so therefore  $g \in \mathcal{F}$  too. It must be that  $\mathcal{F} \equiv \mathcal{G}$ . □

A consequence of the above proof is that we can show that any function  $f$  in a RKHS  $\mathcal{F}$  with kernel  $h$  can be written in the form  $f(x) = \sum_{i=1}^n h(x, x_i) w_i$ , with some  $(w_1, \dots, w_n) \in \mathbb{R}^n$ ,  $n \in \mathbb{N}$ . More precisely,  $\mathcal{F}$  is the completion of the space  $\mathcal{G} = \text{span}\{h(\cdot, x) \mid x \in \mathcal{X}\}$  endowed with the inner product as stated in [Section 2.2](#).

## 2.3 Reproducing kernel Kreĭn space theory

In this section, we shall review basic Kreĭn and reproducing kernel Kreĭn space theory, and comment on the similarity and differences between it and RKHS. Kreĭn spaces are spaces endowed with a Hilbertian topology, characterised by an inner product which is non-positive.

**Definition 2.23** (Negative and indefinite inner products). Let  $\langle \cdot, \cdot \rangle_{\mathcal{F}}$  be an inner product of a vector space  $\mathcal{F}$ , as per [Definition 2.1](#). An inner product is said to be *negative-definite* if for all  $f \in \mathcal{F}$ ,  $\langle f, f \rangle_{\mathcal{F}} \leq 0$ . It is *indefinite* if it is neither positive- nor negative-definite.

**Definition 2.24** (Kreĭn space). An inner product space  $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$  is a *Kreĭn space* if there exists two Hilbert spaces  $(\mathcal{F}_+, \langle \cdot, \cdot \rangle_{\mathcal{F}_+})$  and  $(\mathcal{F}_-, \langle \cdot, \cdot \rangle_{\mathcal{F}_-})$  spanning  $\mathcal{F}$  such that

- All  $f \in \mathcal{F}$  can be decomposed into  $f = f_+ + f_-$ , where  $f_+ \in \mathcal{F}_+$  and  $f_- \in \mathcal{F}_-$ .
- This decomposition is orthogonal, i.e.  $\mathcal{F}_+ \cup \mathcal{F}_- = \{0\}$ , and  $\langle f_+, f_- \rangle_{\mathcal{F}} = 0$  for all  $f_+ \in \mathcal{F}_+$  and  $f_- \in \mathcal{F}_-$ , with the inner product on  $\mathcal{F}$  defined below.
- $\forall f, f' \in \mathcal{F}$ ,  $\langle f, f' \rangle_{\mathcal{F}} = \langle f_+, f'_+ \rangle_{\mathcal{F}_+} - \langle f_-, f'_- \rangle_{\mathcal{F}_-}$ .

Let  $P$  be the projection of the Kreĭn space  $\mathcal{F}$  onto  $\mathcal{F}_+$ , and  $Q = I - P$  the projection onto  $\mathcal{F}_-$ . These are called the *fundamental projections* of  $\mathcal{F}$ . We shall refer to  $\mathcal{F}_+$  as the *positive subspace*, and  $\mathcal{F}_-$  as the *negative subspace*. These monikers stem from the fact that for all  $f, f' \in \mathcal{F}$ ,  $\langle Pf, Pf' \rangle_{\mathcal{F}_+} \geq 0$  while  $\langle Qf, Qf' \rangle_{\mathcal{F}_-} \leq 0$ . We introduce the notation  $\ominus$  to refer to the Kreĭn space decomposition:  $\mathcal{F} = \mathcal{F}_+ \ominus \mathcal{F}_-$ . There is then a notion of an *associated Hilbert space*.

**Definition 2.25** (Associated Hilbert space). Let  $\mathcal{F}$  be a Kreĭn space with decomposition into Hilbert spaces  $\mathcal{F}_+$  and  $\mathcal{F}_-$ . Denote by  $\mathcal{F}_{\mathcal{H}}$  the associated Hilbert space defined by  $\mathcal{F}_{\mathcal{H}} = \mathcal{F}_+ \oplus \mathcal{F}_-$ , with inner product

$$\langle f, f' \rangle_{\mathcal{F}_{\mathcal{H}}} = \langle f_+, f'_+ \rangle_{\mathcal{F}_+} + \langle f_-, f'_- \rangle_{\mathcal{F}_-},$$

for all  $f, f' \in \mathcal{F}$ .

The associated Hilbert space can be found via the linear operator  $J = P - Q$  called the *fundamental symmetry*. That is, a Kreĭn space  $\mathcal{F}$  can be turned into its associated Hilbert space by using the positive-definite inner product of the associated Hilbert space as  $\langle f, f' \rangle_{\mathcal{F}_{\mathcal{H}}} = \langle f, Jf' \rangle_{\mathcal{F}}$ , for all  $f, f' \in \mathcal{F}$ . The converse is true too: Starting from a

Hilbert space  $\mathcal{F}_{\mathcal{H}}$  and an operator  $J$ , the vector space endowed with the inner product  $\langle f, f' \rangle_{\mathcal{F}} = \langle f, Jf' \rangle_{\mathcal{F}_{\mathcal{H}}}$ , for all  $f, f' \in \mathcal{F}$ , is a Kreĭn space.

We realise that for a Kreĭn space  $\mathcal{F}$ ,  $|\langle f, f \rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}_{\mathcal{H}}}^2$  for all  $f \in \mathcal{F}$ , and we say that  $\mathcal{F}_{\mathcal{H}}$  majorises the  $\mathcal{F}$ , and in fact it is the smallest Hilbert space to do so. The strong topology on  $\mathcal{F}$  is defined to be the topology arising from the norm of  $\mathcal{F}_{\mathcal{H}}$ , and this does not depend on the decomposition chosen ([ong2004learning](#)).

**Definition 2.26** (Reproducing kernel Krein space). A Krein space  $\mathcal{F}$  of real-valued functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  on a non-empty set  $\mathcal{X}$  is called a *reproducing kernel Krein space* if the evaluation functional  $\delta_x : f \mapsto f(x)$  is continuous on  $\mathcal{F}$ ,  $\forall x \in \mathcal{X}$ , endowed with its strong topology (i.e. the topology of its associated Hilbert space  $\mathcal{F}_{\mathcal{H}}$ ).

One might wonder whether the uniqueness theorem ([Theorem 2.5](#)) holds for RKKS. Indeed, for every RKKS  $\mathcal{F}$  of functions over a set  $\mathcal{X}$ , there corresponds a unique reproducing kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ .

**Lemma 2.7** (Uniqueness of kernel for RKKS). *Let  $\mathcal{F}$  be a RKKS of functions over a set  $\mathcal{X}$ , with  $\mathcal{F} = \mathcal{F}_+ \oplus \mathcal{F}_-$ . Then,  $\mathcal{F}_+$  and  $\mathcal{F}_-$  are both RKHS with kernel  $h_+$  and  $h_-$ , and the kernel  $h = h_+ - h_-$  is a unique, symmetric, reproducing kernel for  $\mathcal{F}$ .*

*Proof.* Since  $\mathcal{F}$  is a RKKS, evaluation functionals are continuous on  $\mathcal{F}$  with respect to topology of the associated Hilbert space  $\mathcal{F}_{\mathcal{H}} = \mathcal{F}_+ \oplus \mathcal{F}_-$ . Therefore,  $\mathcal{F}_{\mathcal{H}}$  is a RKHS, and so too are  $\mathcal{F}_+$  and  $\mathcal{F}_-$  with respective kernels  $h_+$  and  $h_-$ .

Furthermore,  $h(\cdot, x) \in \mathcal{F}$  since  $h_+(\cdot, x) \in \mathcal{F}_+$  and  $h_-(\cdot, x) \in \mathcal{F}_-$  for some  $x \in \mathcal{X}$ . Then, for any  $f \in \mathcal{F}$ ,

$$\begin{aligned}
 \langle f, h(\cdot, x) \rangle_{\mathcal{F}} &= \langle f, h_+(\cdot, x) \rangle_{\mathcal{F}} - \langle f, h_-(\cdot, x) \rangle_{\mathcal{F}} \\
 &= \langle f_+, h_+(\cdot, x) \rangle_{\mathcal{F}_+} - \langle f_-, h_+(\cdot, x) \rangle_{\mathcal{F}_-} \\
 &\quad - \langle f_+, h_-(\cdot, x) \rangle_{\mathcal{F}_+} + \langle f_-, h_-(\cdot, x) \rangle_{\mathcal{F}_-} \\
 &= f_+(x) + f_-(x) \\
 &= f(x)
 \end{aligned}$$

The last two lines are achieved by linearity of evaluation functionals ( $\delta_x(f_+) + \delta_x(f_-) = \delta_x(f_+ + f_-)$ ), and the fact that  $f = f_+ + f_-$  (by the Kreĭn space decomposition). We have

that  $h = h_+ - h_-$  is a reproducing kernel for  $\mathcal{F}$ . Uniqueness follows as a consequence of the non-degeneracy condition of the respective inner products for  $\mathcal{F}_+$  and  $\mathcal{F}_-$ .  $\square$

*Remark 2.2.* Unlike reproducing kernels of RKHSs, reproducing kernels of RKKSs may not be positive-definite.

The analogue of the Moore-Aronszajn theorem holds partially for RKKS, up to uniqueness. That is, there is *at least* one associated RKKS with kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  if and only if  $h$  can be decomposed as the difference between two positive kernels  $h_+$  and  $h_-$  over  $\mathcal{X}$ , i.e.,  $h = h_+ - h_-$ . The proof of this statement is rather involved, so is omitted in the interest of maintaining coherence to the discussion at hand. This subject has been studied by various authors, one may refer to works by [alpay1991some](#), and [mary2003hilbertian](#).

The take-away message as we close this section is that there is no bijection, but a surjection, between the set of RKKS and the set of bivariate, symmetric functions over  $\mathcal{X} \times \mathcal{X}$ . In any case, Hilbertian topology applies to Kreĭn spaces via the associated Hilbert space, and in particular, RKKS provide a functional space for which evaluation functionals are continuous. The motivation for the use of Kreĭn spaces will become clear when constructing function spaces out of (scaled) building block RKHS later in [Section 2.5](#).

## 2.4 RKHS building blocks

This section describes what we refer to as the “building block” RKHS of functions. In the context of regression modelling, we may assume that the regression function lies in any one of these single RKHS, although it may be more appropriate to consider function spaces built upon these RKHS for more complex models. Construction of new function spaces from these building block RKHS will be discussed in the next section.

3. Update graphics.

### 2.4.1 The RKHS of constant functions

The vector space of constant functions  $\mathcal{F}$  over a set  $\mathcal{X}$  contains the functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  such that  $f(x) = c_f \in \mathbb{R}, \forall x \in \mathcal{X}$ . These functions would be useful to model an overall average, i.e. an “intercept effect”. The space  $\mathcal{F}$  can be equipped with a norm to form an RKHS, as shown in the following lemma.

**Proposition 2.8** (RKHS of constant functions). *The space  $\mathcal{F}$  as described above endowed with the norm  $\|f\|_{\mathcal{F}} = |c_f|$  forms an RKHS with the reproducing kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  as defined, rather simply by,*

$$h(x, x') = 1,$$

*known as the constant kernel.*

*Proof.* If  $\mathcal{F}$  is an RKHS with kernel  $h$  as described, then  $\mathcal{F}$  is spanned by the functions  $h(\cdot, x) = 1$ , so it is clear that  $\mathcal{F}$  consists of constant functions over  $\mathcal{X}$ . On the other hand, if the space  $\mathcal{F}$  is equipped with the inner product  $\langle f, f' \rangle_{\mathcal{F}} = c_f c_{f'}$ , then the reproducing property follows, since  $\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = c_f = f(x)$ . Hence,  $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}} = |c_f|$ .  $\square$



Figure 2.2: Sample paths from the RKHS of constant functions.

#### 2.4.2 The canonical (linear) RKHS

Consider a function space  $\mathcal{F}$  over  $\mathcal{X}$  which consists of functions of the form  $f_{\beta} : \mathcal{X} \rightarrow \mathbb{R}$ ,  $f_{\beta} : x \mapsto \langle x, \beta \rangle_{\mathcal{X}}$  for some  $\beta \in \mathbb{R}$ . Suppose that  $\mathcal{X} \equiv \mathbb{R}^p$ , then  $\mathcal{F}$  consists of the linear functions  $f_{\beta}(x) = x^{\top} \beta$ . More generally, if  $\mathcal{X}$  is a Hilbert space, then its continuous dual consists of elements of the form  $f_{\beta} = \langle \cdot, \beta \rangle_{\mathcal{X}}$  by the Riesz representation theorem. We can show that the continuous dual space of  $\mathcal{X}$  is a RKHS which consists of these linear functions.

**Proposition 2.9** (The canonical RKHS). *The continuous dual space a Hilbert space  $\mathcal{X}$ , denoted by  $\mathcal{X}'$ , is a RKHS of linear functions over  $\mathcal{X}$  of the form  $\langle \cdot, \beta \rangle_{\mathcal{X}}$ ,  $\beta \in \mathcal{X}$ . Its*

reproducing kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is defined by

$$h(x, x') = \langle x, x' \rangle_{\mathcal{X}}.$$

*Proof.* Define  $f_\beta := \langle \cdot, \beta \rangle_{\mathcal{X}}$  for some  $\beta \in \mathcal{X}$ . Clearly this is linear and continuous, so  $f_\beta \in \mathcal{X}'$ , and so  $\mathcal{X}'$  is a Hilbert space containing functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  of the form  $f_\beta(x) = \langle x, \beta \rangle_{\mathcal{X}}$ . By the Riesz representation theorem, every element of  $\mathcal{X}'$  has the form  $f_\beta$ . It also gives us a natural isometric isomorphism such that the following is true:

$$\langle \beta, \beta' \rangle_{\mathcal{X}} = \langle f_\beta, f_{\beta'} \rangle_{\mathcal{X}'}$$

Hence, for any  $f_\beta \in \mathcal{X}'$ ,

$$\begin{aligned} f_\beta(x) &= \langle x, \beta \rangle_{\mathcal{X}} \\ &= \langle f_x, f_\beta \rangle_{\mathcal{X}'} \\ &= \langle \langle \cdot, x \rangle_{\mathcal{X}}, f_\beta \rangle_{\mathcal{X}'}. \end{aligned}$$

Thus,  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  as defined by  $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$  is the reproducing kernel of  $\mathcal{X}'$ .  $\square$

In many other literature, the kernel  $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$  is also known as the *linear kernel*. The use of the term ‘canonical’ is fitting not just due to the relation between a Hilbert space and its continuous dual space. Let  $\phi : \mathcal{X} \rightarrow \mathcal{V}$  be the feature map from the space of covariates (inputs) to some feature space  $\mathcal{V}$ . Suppose both  $\mathcal{X}$  and  $\mathcal{V}$  are Hilbert spaces, then a kernel is defined as

$$h(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}.$$

Taking the feature map to be  $\phi(x) = \langle \cdot, x \rangle_{\mathcal{X}}$ , we can prove the reproducing property to obtain  $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ , which implies  $\phi(x) = h(\cdot, x)$ , and thus  $\phi$  is the *canonical feature map* (steinwart2008support).

The origin of a Hilbert space may be arbitrary, in which case a centring may be appropriate. We define the centred canonical RKHS as follows.

**Definition 2.27** (Centred canonical RKHS). Let  $\mathcal{X}$  be a Hilbert space,  $P$  be a probability measure over  $\mathcal{X}$ , and  $\mu \in \mathcal{X}$  be the mean of a random element  $X \in \mathcal{X}$ . Define  $(\mathcal{X} - \mu)'$ , the continuous dual space of  $\mathcal{X} - \mu$ , to be the *centred canonical RKHS*.  $(\mathcal{X} - \mu)'$  consists

of the centred linear functions  $f_\beta(x) = \langle x - \mu, \beta \rangle_{\mathcal{X}}$ , for  $\beta \in \mathcal{X}$ , such that  $\mathbb{E} f_\beta(X) = 0$ . The reproducing kernel of  $(\mathcal{X} - \mu)'$  is

$$h(x, x') = \langle x - \mu, x' - \mu \rangle_{\mathcal{X}}.$$

*Proof.* That the centred canonical RKHS consists of zero mean function,  $\mathbb{E} f_\beta(X) = 0$ , consider the following argument:

$$\begin{aligned} \mathbb{E} f_\beta(X) &= \mathbb{E} \langle X - \mu, \beta \rangle_{\mathcal{X}} \\ &= \mathbb{E} \langle X, \beta \rangle_{\mathcal{X}} - \langle \mu, \beta \rangle_{\mathcal{X}}, \end{aligned}$$

and since  $\mathbb{E} \langle X, \beta \rangle_{\mathcal{X}} = \langle \mu, \beta \rangle_{\mathcal{X}}$  for any  $\beta \in \mathcal{X}$ , the results follows.  $\square$

*Remark 2.3.* In practice, the probability measure  $P$  over  $\mathcal{X}$  is unknown, so we find it useful to use the empirical distribution over  $\mathcal{X}$  instead, so that  $\mathcal{X}$  is centred by the sample mean  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ .

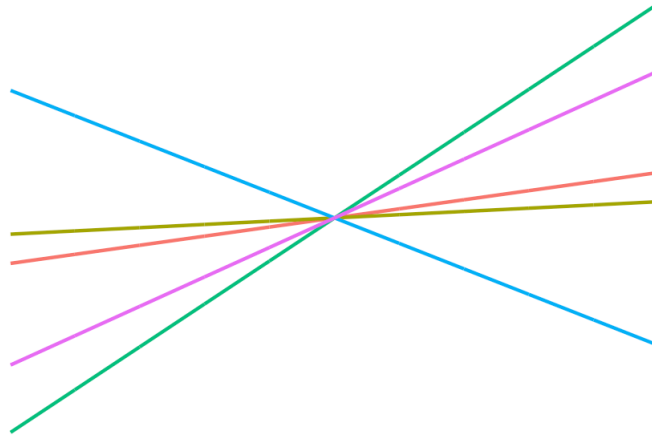


Figure 2.3: Sample paths from the canonical RKHS.

### 2.4.3 The fractional Brownian motion RKHS

Brownian motion, which also goes by the name Wiener process, has been an inquisitive subject in the mathematical sciences, and here, we describe a function space influenced by a generalised version of Brownian motion paths.



Suppose  $B_\gamma(t)$  is a continuous-time Gaussian process on  $[0, T]$ , i.e. for any finite set of indices  $t_1, \dots, t_k$ , where each  $t_j \in [0, T]$ ,  $(B_\gamma(t_1), \dots, B_\gamma(t_k))$  is a multivariate normal random variable.  $B_\gamma(t)$  is said to be a *fractional Brownian motion* (fBm) if  $\mathbb{E} B_\gamma(t) = 0$  for all  $t \in [0, T]$  and

$$\text{Cov}(B_\gamma(t), B_\gamma(s)) = \frac{1}{2}(|t|^{2\gamma} + |s|^{2\gamma} - |t-s|^{2\gamma}) \quad \forall t, s \in [0, T],$$

where  $\gamma \in (0, 1)$  is called the *Hurst index*, *Hurst parameter* or even *Hurst coefficient*. Introduced by [mandelbrot1968fractional](#), fBms are a generalisation of Brownian motion. The Hurst parameter plays two roles: 1) It describes the raggedness of the resultant motion, with higher values leading to smoother motion; and 2) it determines the type of process the fBm is, as past increments of  $B_\gamma(t)$  are weighted by  $(t-s)^{\gamma-1/2}$ . When  $\gamma = 1/2$  exactly, then the fBm is a standard Brownian motion and its increments are independent; when  $\gamma > 1/2$  (resp.  $\gamma < 1/2$ ) its increments are positively (resp. negatively) correlated.

Now let  $\mathcal{X}$  be a Hilbert space. [schoenberg1937](#) has shown that, for  $0 < \gamma \leq 1$ , there exists a Hilbert space  $\mathcal{V}$  and a function  $\phi_\gamma : \mathcal{X} \rightarrow \mathcal{V}$  such that  $\forall x, x' \in \mathcal{X}$ ,

$$\|\phi_\gamma(x) - \phi_\gamma(x')\|_{\mathcal{V}} = \|x - x'\|_{\mathcal{X}}^\gamma.$$

Using the polarisation identity, we find that the kernel of the RKHS with feature space  $\mathcal{V}$  and feature map  $\phi_\gamma$  defines a kernel function  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  identical to the fBm covariance kernel.

**Definition 2.28** (Fractional Brownian motion RKHS). The fractional Brownian motion (fBm) RKHS  $\mathcal{F}$  is the space of functions on the Hilbert space  $\mathcal{X}$  possessing the reproducing kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  defined by

$$h_\gamma(x, x') = \langle \phi_\gamma(x), \phi_\gamma(x') \rangle_{\mathcal{V}} = \frac{1}{2}(\|x\|_{\mathcal{X}}^{2\gamma} + \|x'\|_{\mathcal{X}}^{2\gamma} - \|x - x'\|_{\mathcal{X}}^{2\gamma}),$$

which depends on the Hurst coefficient  $\gamma \in (0, 1)$ . We shall reference this space as the fBm- $\gamma$  RKHS.

*Remark 2.4.* When  $\gamma = 1$ , by the polarisation identity we get  $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ , which is the (reproducing) kernel of the canonical RKHS.

From its construction, it is clear that the fBm kernel is positive definite, and thus defines an RKHS. That the fBm RKHS describes a space of functions is proved in

def:fbmrkhs

[cohen2002](#), who studied this space in depth. It is also noted in the collection of examples of [berlinet2011reproducing](#).

The Hurst coefficient  $\gamma$  controls the “smoothness” of the functions in the RKHS. We can talk about smoothness in the context of Hölder continuity of functions.

**Definition 2.29** (Hölder condition). A function  $f$  over a set  $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$  is said to be *Hölder continuous* of order  $0 < \gamma \leq 1$  if there exists a  $C > 0$  such that  $\forall x, x' \in \mathcal{X}$ ,

$$|f(x) - f(x')| \leq C\|x - x'\|^{\gamma}.$$

Functions in the Hölder space  $C^{k,\gamma}(\mathcal{X})$ , where  $k \geq 0$  is an integer, consists of those functions over  $\mathcal{X}$  having continuous derivatives up to order  $k$  and such that the  $k$ th partial derivatives are Hölder continuous of order  $\gamma$ . Unlike realisations of actual fBm paths with Hurst index  $\gamma$ , which are well-known to be almost surely Hölder continuous of order less than  $\gamma$  ([embrechts2002selfsimilar](#)), functions in its namesake RKHS are strictly smoother.

**Claim 2.10.** *The fBm- $\gamma$  RKHS  $\mathcal{F}$  of functions over  $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$  are Hölder continuous of order  $\gamma$ .*

*Proof.* For some  $f \in \mathcal{F}$  we have  $f(x) = \langle f, h(\cdot, x) \rangle_{\mathcal{F}}$  by the reproducing property of the kernel  $h$  of  $\mathcal{F}$ . It follows from the Cauchy-Schwarz inequality that for any  $x, x' \in \mathcal{X}$ ,

$$\begin{aligned} |f(x) - f(x')| &= |\langle f, h(\cdot, x) - h(\cdot, x') \rangle_{\mathcal{F}}| \\ &\leq \|f\|_{\mathcal{F}} \cdot \|h(\cdot, x) - h(\cdot, x')\|_{\mathcal{F}} \\ &= \|f\|_{\mathcal{F}} \cdot \|x - x'\|_{\mathcal{X}}^{\gamma}, \end{aligned}$$

since

$$\begin{aligned} \|h(\cdot, x) - h(\cdot, x')\|_{\mathcal{F}}^2 &= \|h(\cdot, x)\|_{\mathcal{F}}^2 + \|h(\cdot, x')\|_{\mathcal{F}}^2 - 2\langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}} \\ &= h(x, x) + h(x', x') - 2h(x, x') \\ &= \|x - x'\|_{\mathcal{X}}^{2\gamma}, \end{aligned}$$

and thus proving the claim. □

The fBm- $\gamma$  RKHS is spanned by the functions  $h(\cdot, x)$ , which means that  $f(0) = 0$  for all  $f \in \mathcal{F}$ , which may be undesirable. We define the centred fBm RKHS as follows.

4. This is the same for any RKHS?

**Definition 2.30** (Centred fBm RKHS). Let  $\mathcal{X}$  be a Hilbert space,  $P$  be a probability measure over  $\mathcal{X}$ , and  $\mu \in \mathcal{X}$  be the mean with respect to this probability measure. The kernel  $\bar{h} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  defined by

$$\bar{h}(x, x') = \frac{1}{2} \mathbb{E} \left[ \|x - X\|_{\mathcal{X}}^{2\gamma} + \|x' - X'\|_{\mathcal{X}}^{2\gamma} - \|x - x'\|_{\mathcal{X}}^{2\gamma} - \|X - X'\|_{\mathcal{X}}^{2\gamma} \right]$$

is the reproducing kernel of the *centred* fBm- $\gamma$  RKHS, which consists of functions  $f$  in the fBm- $\gamma$  RKHS such that  $\mathbb{E} f(X) = 0$ . In the above definition,  $X, X' \sim P$  are two independent copies of a random vector  $X \in \mathcal{X}$ .

*Remark 2.5.* Again, when  $\gamma = 1$ , we get the reduction

$$\begin{aligned} \bar{h}(x, x') &= \frac{1}{2} \mathbb{E} [\|x - X\|_{\mathcal{X}}^2 + \|x' - X'\|_{\mathcal{X}}^2 - \|x - x'\|_{\mathcal{X}}^2 - \|X - X'\|_{\mathcal{X}}^2] \\ &= \frac{1}{2} \mathbb{E} [\langle X, X \rangle_{\mathcal{X}} + \langle X', X' \rangle_{\mathcal{X}} + 2\langle x, x' \rangle_{\mathcal{X}} - 2\langle x, X \rangle_{\mathcal{X}} - 2\langle x', X' \rangle_{\mathcal{X}}] \\ &= \langle \mu, \mu \rangle_{\mathcal{X}} + \langle x, x' \rangle_{\mathcal{X}} - \langle x, \mu \rangle_{\mathcal{X}} - \langle \mu, x' \rangle_{\mathcal{X}} \\ &= \langle x - \mu, x' - \mu \rangle_{\mathcal{X}}, \end{aligned}$$

which is the (reproducing) kernel of the centred canonical RKHS.

#### 2.4.4 The squared exponential RKHS

The [squared exponential \(SE\)](#) kernel function is indeed known to be the default kernel used for Gaussian process regression in machine learning. It is a positive definite function, and hence defines an RKHS. The definition of the [SE](#) RKHS is as follows.

**Definition 2.31** (Squared exponential RKHS). The squared exponential (SE) RKHS  $\mathcal{F}$  of functions over some set  $\mathcal{X} \subseteq \mathbb{R}^p$  equipped with the 2-norm  $\|\cdot\|_2$  is defined by the positive definite kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

$$h(x, x') = \exp \left( -\frac{\|x - x'\|_2^2}{2l^2} \right).$$

The real-valued parameter  $l > 0$  is called the *lengthscale* parameter, and is a smoothing parameter for the functions in the RKHS.

It is known by many other names, including the Gaussian kernel, due to its semblance to the kernel of the Gaussian pdf. Especially in the machine learning literature, the term

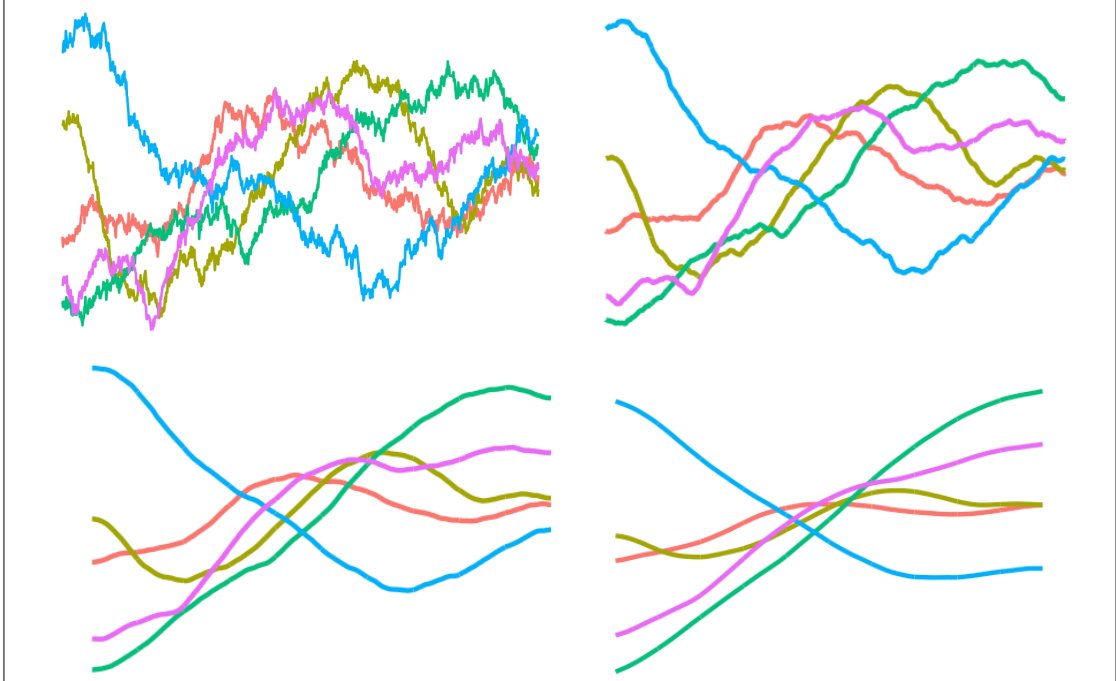


Figure 2.4: Sample paths from the fBm RKHS with varying Hurst coefficients.

Gaussian radial basis functions (RBF) is used, and commonly the simpler parameterisation  $\gamma = 1/2l^2$  is utilised. [duvenaud2014automatic](#) remarks that “exponentiated quadratic” is a better fitting and descriptive name for this kernel.

Despite being used extensively for learning algorithms using kernels, an explicit study of the RKHS defined by the SE kernel was not done until recently by [steinwart2006explicit](#). In that work, the authors describe the nature of real-valued functions in the SE RKHS by considering a real restriction on the SE RKHS of functions over complex values. Their derivation of an orthonormal basis of such an RKHS proved the SE kernel to be the reproducing kernel for the SE RKHS.

SE kernels are known to be “universal”. That is, it satisfies the following definition of universal kernels due to [micchelli2006universal](#).

**Definition 2.32** (Universal kernel). Let  $C(\mathcal{X})$  is the space of all continuous, complex-valued functions  $f : \mathcal{X} \rightarrow \mathbb{C}$  equipped with the maximum norm  $\|\cdot\|_\infty$ , and denote  $\mathcal{K}(\mathcal{X})$  as the space of *kernel sections*  $\overline{\text{span}}\{h(\cdot, x) | x \in \mathcal{X}\}$ , where here,  $h$  is a complex-valued kernel function. A kernel  $h$  is said to be *universal* if given any compact subset  $\mathcal{Z} \subset \mathcal{X}$ ,

any positive number  $\epsilon$  and any function  $f \in C(\mathcal{Z})$ , there is a function  $g \in \mathcal{K}(\mathcal{Z})$  such that  $\|f - g\|_{\mathcal{Z}} \leq \epsilon$ .

The consequence of this universal property vis-à-vis regression modelling is that any (continuous) regression function  $f$  may be approximated very well by a function  $\hat{f}$  belonging to the SE RKHS, and these two functions can get arbitrarily close to each other in the max norm sense. This, together with some very convenient computational advantages that the SE kernel brings (more on this in a later chapter), is a testament to the popularity of SE kernels.

In a similar manner to the two previous subsections, we may also derive the *centred* SE RKHS.

**Definition 2.33** (Centred SE RKHS). Let  $\mathcal{X} \subseteq \mathbb{R}^p$  be equipped with the 2-norm  $\|\cdot\|_2$ , and let  $P$  denote the distribution over  $\mathcal{X}$ . Assuming integrability of  $h(x, X)$ , for any  $x \in \mathcal{X}$  and a random element  $X \in \mathcal{X}$ , the *centred* squared exponential (SE) RKHS (with lengthscale  $l$ ) of functions over  $\mathcal{X}$  is defined by the positive definite kernel  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

$$h(x, x') = e^{-\frac{\|x - x'\|_2^2}{2l^2}} - \mathbb{E} e^{-\frac{\|x - X'\|_2^2}{2l^2}} - \mathbb{E} e^{-\frac{\|X - x'\|_2^2}{2l^2}} + \mathbb{E} e^{-\frac{\|X - X'\|_2^2}{2l^2}},$$

where  $X, X' \sim P$  are two independent random elements of  $\mathcal{X}$ . This ensures that  $\mathbb{E} f(X) = 0$  for any  $f$  in this RKHS.

#### 2.4.5 The Pearson RKHS

In all of the previous RKHS of functions, the domain  $\mathcal{X}$  was taken to be some Euclidean space. The Pearson RKHS is a vector space of functions whose domain  $\mathcal{X}$  is a finite set. Let  $P$  be a probability measure over the finite set  $\mathcal{X}$ . The Pearson RKHS is defined as follows.

**Definition 2.34** (Pearson RKHS). The *Pearson RKHS* is the RKHS of functions over a finite set  $\mathcal{X}$  defined by the reproducing kernel

$$h(x, x') = \frac{\delta_{xx'}}{P(X = x)} - 1,$$

where  $X \sim P$  and  $\delta$  is the Kronecker delta.

def:pearson

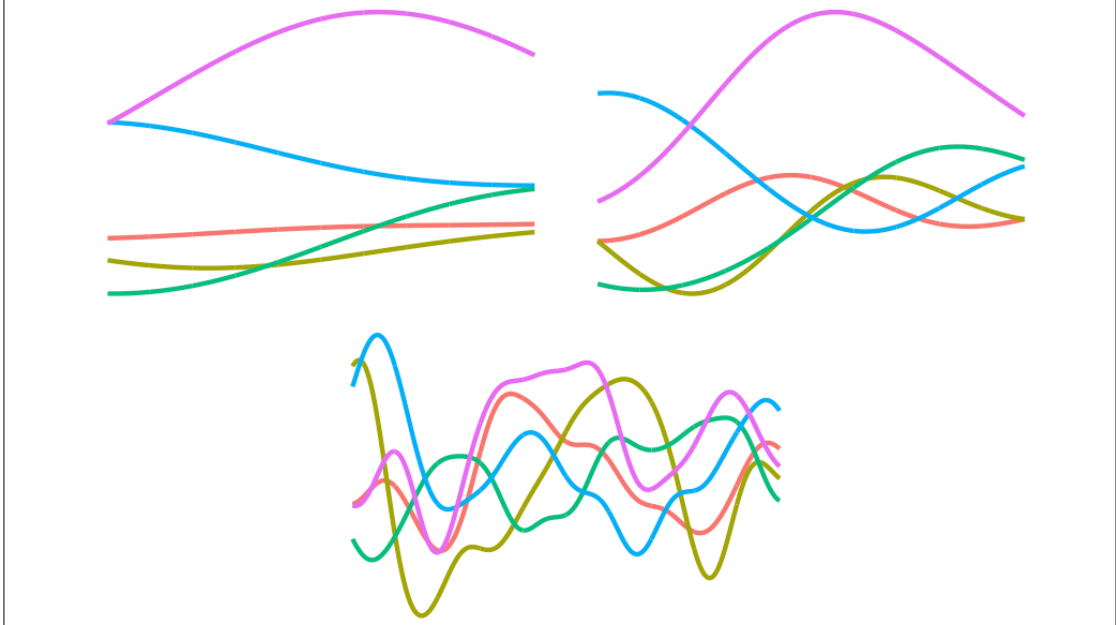


Figure 2.5: Sample paths from the SE RKHS with varying values for the lengthscale.

The Pearson RKHS contains functions which are centred, and has the desirable property that the contribution of  $f(x)^2$  to the squared norm of  $f$  is proportional to  $P(X = x)$ .

**Claim 2.11.** *Let  $\mathcal{F}$  be the Pearson RKHS of functions over a finite set  $\mathcal{X}$ . Then,*

$$\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathbb{R} \mid \mathbb{E} f(X) = 0\}$$

with

$$\|f\|_{\mathcal{F}}^2 = \text{Var } f(X) = \sum_{x \in \mathcal{X}} P(X = x) f(x)^2, \quad \forall f \in \mathcal{F}.$$

*Proof.* Write  $p_x = P(X = x)$ . The set of functions  $\{h(\cdot, x) \mid x \in \mathcal{X}\}$  form a basis for  $\mathcal{F}$ , and thus each  $f \in \mathcal{F}$  can be written as  $f(x) = \sum_{x' \in \mathcal{X}} w_{x'} h(x, x')$  for some scalars  $w_i \in \mathbb{R}$ ,  $i \in \mathcal{X}$ . But  $\mathbb{E} h(X, x') = \mathbb{E}[\delta_{Xx'}]/p_{x'} - 1 = p_{x'}/p_{x'} - 1 = 0$ , and thus  $\mathbb{E} f(X) = 0$ . Conversely, suppose  $f : \mathcal{X} \rightarrow \mathbb{R}$  is such that  $\mathbb{E} f(X) = 0$ . Taking  $w_x = p_x f(x)$ , we see

that

$$\begin{aligned} \sum_{x' \in \mathcal{X}} w_{x'} h(x, x') &= \frac{w_x}{p_x} - \sum_{x' \in \mathcal{X}} w_{x'} \\ &= \frac{f(x) \cancel{p_x}}{\cancel{p_x}} - \sum_{x' \in \mathcal{X}} \cancel{p_{x'}} f(x') \xrightarrow{E f(X) = 0} = f(x) \end{aligned}$$

and thus  $h(\cdot, x)$  spans  $\mathcal{F}$  so  $f \in \mathcal{F}$ . To provide the second part, noting that with the choice  $w_x = p_x f(x)$  and due to the reproducing property of  $h$  for the RKHS  $\mathcal{F}$ , the squared norm is

$$\begin{aligned} \langle f, f \rangle_{\mathcal{F}} &= \left\langle \sum_{x \in \mathcal{X}} w_x h(\cdot, x), \sum_{x' \in \mathcal{X}} w_{x'} h(\cdot, x') \right\rangle_{\mathcal{F}} \\ &= \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} w_x w_{x'} \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}} \\ &= \sum_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} w_x w_{x'} h(x, x') \\ &= \sum_{x \in \mathcal{X}} w_x f(x) \\ &= \sum_{x \in \mathcal{X}} P(X = x) f(x)^2, \end{aligned}$$

which is also the variance of  $f(X)$ . □

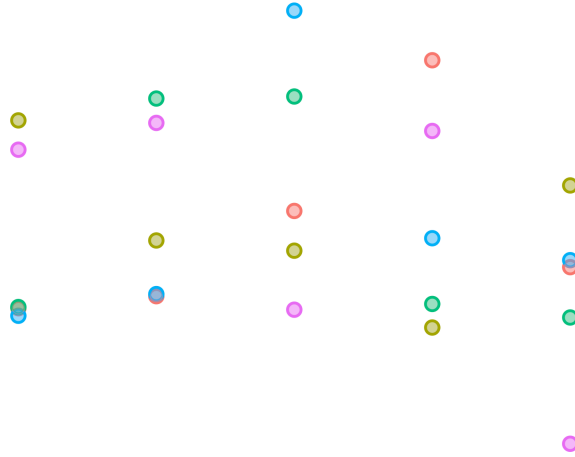


Figure 2.6: Sample “paths” from the Pearson RKHS. These are represented as points over a finite set.

## 2.5 Constructing RKKS from existing RKHS

sec:constru  
ctrkks

The previous section outlined all of the basic RKHSs of functions that will form the building blocks when constructing more complex function spaces. As previously mentioned in the preliminaries, sums of kernels are kernels and products of kernels are also kernels. This provides us a platform for constructing new RKHS from existing ones. To be more flexible in the specification of these new function spaces, we do not restrict ourselves to positive definite kernels only, thereby necessitating us to use the theory of RKKS.

### 2.5.1 Sums, products and scaling of RKHS

Sums of positive definite kernels are also positive definite kernels, and the product of positive definite kernel is a positive definite kernel. They each, in turn, are associated with a RKHS that is defined by the sum of kernels and product of kernels, respectively. The two lemmas below formalise these two facts.

thm:sumkern  
els

**Lemma 2.12** (Sum of kernels). *If  $h_1$  and  $h_2$  are kernels on  $\mathcal{X}_1$  and  $\mathcal{X}_2$  respectively, then  $h = h_1 + h_2$  is a kernel on  $\mathcal{X}_1 \times \mathcal{X}_2$ . Moreover, denote  $\mathcal{F}_1$  and  $\mathcal{F}_2$  the RKHS defined by  $h_1$  and  $h_2$  respectively. Then  $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2$  is an RKHS defined by  $h = h_1 + h_2$ , where*

$$\mathcal{F}_1 \oplus \mathcal{F}_2 = \{f : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \mathbb{R} \mid f = f_1 + f_2, f_1 \in \mathcal{F}_1 \text{ and } f_2 \in \mathcal{F}_2\}.$$

For all  $f \in \mathcal{F}$ ,

$$\|f\|_{\mathcal{F}}^2 = \min_{f_1+f_2=f} \{\|f_1\|_{\mathcal{F}_1}^2 + \|f_2\|_{\mathcal{F}_2}^2\}.$$

*Proof.* That  $h_1 + h_2$  is a kernel should be obvious, as the sum of two positive definite functions is also positive definite. For a proof of the remaining statements, see [berlinet2011reproducing](#).  $\square$

thm:prodker  
nels

**Lemma 2.13** (Products of kernels). *Let  $\mathcal{F}_1$  and  $\mathcal{F}_2$  be two RKHS of functions over  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , with respective reproducing kernels  $h_1$  and  $h_2$ . Then,  $h = h_1 h_2$  is a kernel on  $\mathcal{X}_1 \times \mathcal{X}_2$ . Moreover, the tensor product space  $\mathcal{F}_1 \otimes \mathcal{F}_2$  is an RKHS with reproducing kernel  $h$ .*

*Proof.* Fix  $n \in \mathbb{N}$ , and let  $\mathbf{H}_1$  and  $\mathbf{H}_2$  be the kernel matrices for  $h_1$  and  $h_2$  respectively. Since these kernel matrices are symmetric and positive-definite by virtue of  $h_1$  and  $h_2$



being symmetric and positive-definite functions, we can write  $\mathbf{H}_1 = \mathbf{A}^\top \mathbf{A}$  and  $\mathbf{H}_2 = \mathbf{B}^\top \mathbf{B}$  for some matrices  $\mathbf{A}$  and  $\mathbf{B}$ : perform an (orthogonal) eigendecomposition of each of the kernel matrices, and take square roots of the eigenvalues. Let  $\mathbf{H}$  be the kernel matrix for  $h = h_1 h_2$ . With  $x_i = (x_{i1}, x_{i2})$ , its  $(i, j)$  entries are

$$\begin{aligned} h(x_i, x_j) &= h_1(x_{i1}, x_{j1}) h_2(x_{i2}, x_{j2}) \\ &= (\mathbf{A}^\top \mathbf{A})_{ij} \cdot (\mathbf{B}^\top \mathbf{B})_{ij} \\ &= \sum_{k=1}^n a_{ik} a_{jk} \sum_{l=1}^n b_{il} b_{jl}, \end{aligned}$$

where we have denoted  $b_{ij}$  and  $c_{ij}$  to be the  $(i, j)$ th entries of  $\mathbf{B}$  and  $\mathbf{C}$  respectively. Then,

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n h(x_i, x_j) &= \sum_{k=1}^n \sum_{l=1}^n \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j a_{ik} a_{jk} b_{il} b_{jl} \\ &= \sum_{k=1}^n \sum_{l=1}^n \left( \sum_{i=1}^n \lambda_i a_{ik} b_{il} \right) \left( \sum_{j=1}^n \lambda_j a_{jk} b_{jl} \right) \\ &= \sum_{k=1}^n \sum_{l=1}^n \left( \sum_{i=1}^n \lambda_i a_{ik} b_{il} \right)^2 \\ &\geq 0 \end{aligned}$$

Again, for the remainder of the statement in the lemma, we refer to [berlinet2011reproducing](#). □

A familiar fact from linear algebra is realised here from [Lemmas 2.12](#) and [2.13](#): 1) the addition of positive definite matrices is a positive definite matrix; and 2) the *Hadamard product*<sup>3</sup> of two positive definite matrices is a positive definite matrix.

The scale of an RKHS of functions  $\mathcal{F}$  over a set  $\mathcal{X}$  with kernel  $h$  may be arbitrary. To resolve this issue, a scale parameter  $\lambda \in \mathbb{R}$  for the kernel  $h$  may be introduced, which will typically need to be estimated from the data. If  $h$  is a positive definite kernel on  $\mathcal{X} \times \mathcal{X}$ , and  $\lambda \geq 0$  a scalar, then this yields a scaled RKHS  $\mathcal{F}_\lambda = \{\lambda f \mid f \in \mathcal{F}\}$  with reproducing kernel  $\lambda h$ , where  $\mathcal{F}$  is the RKHS defined by  $h$ .

<sup>3</sup>The Hadamard product is an element-wise multiplication of two matrices  $\mathbf{A}$  and  $\mathbf{B}$  of identical dimensions, denoted  $\mathbf{A} \circ \mathbf{B}$ . That is,  $(\mathbf{A} \circ \mathbf{B})_{ij} = \mathbf{A}_{ij} \mathbf{B}_{ij}$ .

Restricting  $\lambda$  to the positive reals is arbitrary and unnecessarily restrictive. Especially when considering sums and products of scaled RKHSs, having negative scale parameters also give additional flexibility. The resulting kernels from summation and/or multiplication with negative kernels may no longer be positive-definite, and in such cases, they give rise to RKKS instead.

*Remark 2.6.* Recall that a RKKS  $\mathcal{F}$  of functions over  $\mathcal{X}$  can be uniquely decomposed as the difference between two RKHSs  $\mathcal{F}_+$  and  $\mathcal{F}_-$ , and its associated Hilbert space  $\mathcal{F}_{\mathcal{H}}$  is the RKHS  $\mathcal{F}_+ \oplus \mathcal{F}_-$ . It is important to note that both  $\mathcal{F}$  and  $\mathcal{F}_{\mathcal{H}}$  contain identical functions over  $\mathcal{X}$ , but their topologies are different. That is to say, functions that are close with respect to the norm of  $\mathcal{F}$  may not be close to each other in the norm of  $\mathcal{F}_{\mathcal{H}}$ .

## 2.5.2 The polynomial RKKS

A polynomial construction based on a particular RKHS building block is considered here. For example, using the canonical RKHS in the polynomial construction would allow us to easily add higher order effects of the covariates  $x \in \mathcal{X}$ . In particular, we only require a single scale parameter in polynomial kernel construction.

**Definition 2.35** (The polynomial RKKS). Let  $\mathcal{X}$  be a Hilbert space. The kernel function  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  obtained through the  $d$ -degree polynomial construction of linear kernels is

$$h_{\lambda}(x, x') = (\lambda \cdot \langle x, x' \rangle_{\mathcal{X}} + c)^d,$$

where  $\lambda \in \mathbb{R}$  is a scale parameter for the linear kernel, and  $c \in \mathbb{R}$  is a real constant called the *offset*. This kernel defined the *polynomial RKKS* of degree  $d$ .

Write

$$h_{\lambda}(x, x')_{\mathcal{F}} = \sum_{k=0}^d \frac{d!}{k!(d-k)!} c^{k-d} \lambda^k \langle x, x' \rangle_{\mathcal{X}}^k.$$

Evidently, as the name suggests, this is a polynomial involving the canonical kernel. In particular, each of the  $k$ -powered kernels (i.e.,  $\langle x, x' \rangle_{\mathcal{X}}^k$ ) defines an RKHS of their own (since these are merely products of kernels), and therefore the sum of these  $k$ -powered kernels define the polynomial RKHS.

The offset parameter influences trade-off between the higher-order versus lower-order terms in the polynomial. It is sometimes known as the bias term.

**Claim 2.14.** *The polynomial RKKS of functions over  $\mathbb{R}$ , denoted  $\mathcal{F}$ , contains polynomial functions of the form  $f(x) = \sum_{k=0}^d \beta_k x^k$ .*

*Proof.* By construction,  $\mathcal{F} = \mathcal{F}_0 \oplus \bigoplus_{i=1}^d \bigotimes_{j=1}^i \mathcal{F}_j$ , where each  $\mathcal{F}_j, j \neq 0$  is the canonical RKHS, and  $\mathcal{F}_0$  is the RKHS of constant functions. Each  $g \in \mathcal{F}$  can therefore be written as  $g = \beta_0 + \sum_{i=1}^d \prod_{j=1}^i f_j$ , and  $f_j(x) = b_j x$  from before, where  $b_j$  is a constant. Therefore,  $g(x) = \sum_{k=0}^d \beta_k x^k$ .  $\square$

*Remark 2.7.* We may opt to use other RKHSs as the building blocks of the polynomial RKHS. In particular, using the centred canonical kernel seems natural, so that each of the functions in the constituents of the direct sum of spaces is centred. However, the polynomial RKKS itself will not be centred.

### 2.5.3 The ANOVA RKKS

We find it useful to begin this subsection by spending some time to elaborate on the classical analysis of variance (ANOVA) decomposition, and the associated notions of main effects and interactions. This will go a long way in understanding the thinking behind constructing an ANOVA-like RKKS of functions.

#### The classical ANOVA decomposition

The standard one-way ANOVA is essentially a linear regression model which allows comparison of means from two or more samples. Given sets of observations  $y_j = \{y_{1j}, \dots, y_{n_j j}\}$ ,  $j = 1, \dots, m$ , we consider the linear model  $y_{ij} = \mu_j + \epsilon_{ij}$ , where  $\epsilon_{ij}$  are independent, univariate normal random variables with a common variance. This covariate-less model is used to make inferences about the *treatment means*  $\mu_j$ . Often, the model is written in the *overparameterised* form by substituting  $\mu_j = \mu + \tau_j$ . This gives a different, arguably better, interpretability to the model: The  $\tau_j$ 's, referred to as the *treatment effects*, now represent the amount of deviation from the grand, *overall mean*  $\mu$ . Estimating all  $\tau_j$ 's and  $\mu$  separately is not possible because there is one degree of freedom that needs to be addressed in the model: There are  $p+1$  mean parameters to estimate but only information from  $p$  means. A common fix to the identifiability issue is to set one of the  $\mu_j$ 's, say the first one  $\mu_1$ , to zero, or impose the restriction  $\sum_{j=1}^m \mu_j = 0$ . The former treats one of the  $m$  levels as the control, while the latter treats all treatment effects symmetrically.

Now write the ANOVA model slightly differently, as  $y_i = f(x_i) + \epsilon_i$ , where  $f$  is defined on the discrete domain  $\mathcal{X} = \{1, \dots, m\}$ , and  $i$  indexes all of the  $n := \sum_{j=1}^m n_j$  observations. Here,  $f$  represents the group-level mean, returning  $\mu_j$  for some  $j \in \mathcal{X}$ . In a similar manner, we can perform the ANOVA decomposition on  $f$  as

$$f = Af + (I - A)f = f_o + f_t,$$

where  $A$  is an averaging operator that “averages out” its argument  $x$  and returns a constant, and  $I$  is the identity operator.  $f_o = Af$  is a constant function representing the *overall mean*, whereas  $f_t = (I - A)f$  is a function representing the *treatment effects*  $\tau_j$ . Here are two choices of  $A$ :

- $Af(x) = f(1) = \mu_1$ . This implies  $f(x) = f(1) + (f(x) - f(1))$ . The overall mean  $\mu$  is the group mean  $\mu_1$ , which corresponds to setting the restriction  $\mu_1 = 0$ .
- $Af(x) = \sum_{x=1}^m f(x)/m =: \bar{\alpha}$ . This implies  $f(x) = \bar{\alpha} + (f(x) - \bar{\alpha})$ . The overall mean is  $\mu = \sum_{j=1}^m \alpha_j/m$ , which corresponds to the restriction  $\sum_{j=1}^m \mu_j = 0$ .

By definition,  $AAf = A^2f = Af$ , because averaging a constant returns that constant [Side note: This idempotent property of the linear operator  $A$  on  $f$  speaks to the possibility of it being an *orthogonal projection*, and indeed this is so—we shall return to this point later when we describe functional ANOVA decomposition]. We must have that  $Af_t = A(I - A)f = Af - A^2f = 0$ . The choice of  $A$  is arbitrary, as is the choice of restriction, so long as it satisfies the condition that  $Af_c = 0$ .

The multiway ANOVA can be motivated in a similar fashion. Let  $x = (x_1, \dots, x_p) \in \prod_{k=1}^p \mathcal{X}_k$ , and consider functions that map  $\prod_{k=1}^p \mathcal{X}_j$  to  $\mathbb{R}$ . Let  $A_j$  be an averaging operator on  $\mathcal{X}_k$  that averages the  $k$ th component of  $x$  from the active argument list, i.e.  $A_k f$  is constant on the  $\mathcal{X}_k$  axis but not necessarily an overall constant function. An ANOVA decomposition of  $f$  is

$$f = \left( \prod_{k=1}^p (A_k + I - A_k) \right) f = \sum_{\mathcal{K} \in \mathcal{P}_p} \left( \prod_{k \in \mathcal{K}} (I - A_k) \prod_{k \notin \mathcal{K}} A_k \right) f = \sum_{\mathcal{K} \in \mathcal{P}_p} f_{\mathcal{K}}$$

where we had denoted  $\mathcal{P}_p = \mathcal{P}(\{1, \dots, p\})$  to be the power set of  $\{1, \dots, p\}$  whose cardinality is  $2^p$ . The summands  $f_{\mathcal{K}}$  will compose of the overall effect, main effects, two-way interaction terms, and so on. Each of the terms will satisfy the condition  $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p$ .

**Example 2.1** (Two-way ANOVA decomposition). Let  $p = 2$ ,  $\mathcal{X}_1 = \{1, \dots, m_1\}$ , and  $\mathcal{X}_2 = \{1, \dots, m_2\}$ . The power set  $\mathcal{P}_2$  is  $\{\{\}, \{1\}, \{2\}, \{1, 2\}\}$ . The ANOVA decomposition of  $f$  is

$$f = f_0 + f_1 + f_2 + f_{12}.$$

Here are two choices for the averaging operator  $A_k$  analogous to the previous illustration in the one-way ANOVA.

- Let  $A_1 f(x) = f(1, x_2)$  and  $A_2 f(x) = f(x_1, 1)$ . Then,

$$\begin{aligned} f_0 &= A_1 A_2 f &&= f(1, 1) \\ f_1 &= (I - A_1) A_2 f &&= f(x_1, 1) - f(1, 1) \\ f_2 &= A_1 (I - A_2) f &&= f(1, x_2) - f(1, 1) \\ f_{12} &= (I - A_1)(I - A_2) f &&= f(x_1, x_2) - f(x_1, 1) - f(1, x_2) + f(1, 1). \end{aligned}$$

- Let  $A_k f(x) = \sum_{x_k=1}^{m_k} f(x_1, x_2) / m_k, k = 1, 2$ . Then,

$$\begin{aligned} f_0 &= A_1 A_2 f &&= f_{..} \\ f_1 &= (I - A_1) A_2 f &&= f_{x_1 \cdot} - f_{..} \\ f_2 &= A_1 (I - A_2) f &&= f_{\cdot x_2} - f_{..} \\ f_{12} &= (I - A_1)(I - A_2) f &&= f - f_{x_1 \cdot} - f_{\cdot x_2} + f_{..}, \end{aligned}$$

where  $f_{..} = \sum_{x_1, x_2} f(x_1, x_2) / m_1 m_2$ ,  $f_{x_1 \cdot} = \sum_{x_2} f(x_1, x_2) / m_2$ , and  $f_{\cdot x_1} = \sum_{x_1} f(x_1, x_2) / m_1$ .

### Functional ANOVA decomposition

Let us now extend the ANOVA decomposition idea to a general function  $f : \mathcal{X} \rightarrow \mathbb{R}$  in some vector space  $\mathcal{F}$ . Specifically, we shall consider the (Hilbert) space of square integrable functions over  $\mathcal{X}$  with measure  $\nu$ ,  $\mathcal{F} \equiv L^2(\mathcal{X}, \nu)$ . We shall jump straight into the multiway ANOVA analogue for functional decomposition, and to that end, consider  $x = (x_1, \dots, x_p) \in \prod_{k=1}^p \mathcal{X}_k =: \mathcal{X}$  a measurable space, where each of the spaces  $\mathcal{X}_k$  has measure  $\nu_k$ , and  $\nu = \nu_1 \times \dots \times \nu_d$  is the product measure on  $\mathcal{X}$ . As  $\mathcal{X}$  need not necessarily be a collection of finite sets, we need to figure out a suitable linear operator that performs an “averaging” of some sort.

Consider the linear operator  $A_k : \mathcal{F} \rightarrow \mathcal{F}_{-k}$ , where  $\mathcal{F}_{-k}$  is a vector space of functions for which the  $k$ th component is constant over  $\mathcal{X}$ , defined by

$$A_k f = \int_{\mathcal{X}_k} f(x_1, \dots, x_p) d\nu(x_k). \quad (2.2)$$

Thus, for the one-way ANOVA ( $p = 1$ ), we get

$$f = \overbrace{\int_{\mathcal{X}} f(x) d\nu(x)}^{f_0} + \overbrace{\left( f - \int_{\mathcal{X}} f(x) d\nu(x) \right)}^{f_1} \quad (2.3)$$

and for the two-way ANOVA ( $p = 2$ ), we have  $f = f_0 + f_1 + f_2 + f_{12}$ , with

$$\begin{aligned} f_0 &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) d\nu(x_1) d\nu(x_2) \\ f_1 &= \int_{\mathcal{X}_2} \left( f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) d\nu(x_1) \right) d\nu(x_2) \\ f_2 &= \int_{\mathcal{X}_1} \left( f(x_1, x_2) - \int_{\mathcal{X}_2} f(x_1, x_2) d\nu(x_2) \right) d\nu(x_1) \\ f_{12} &= f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) d\nu(x_1) - \int_{\mathcal{X}_2} f(x_1, x_2) d\nu(x_2) \\ &\quad + \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) d\nu(x_1) d\nu(x_2). \end{aligned}$$

As a remark, the averaging operator  $A_k$  defined in (2.2) is indeed true to its name, in that it calculates the mean function of  $f$  over the  $k$ th coordinate. For comparison, this is identical to the second type of restriction we considered in the classical ANOVA previously (i.e., setting  $\sum_j \mu_j = 0$ ). We must also have, as before, that  $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p$ . For the one-way functional ANOVA decomposition in (2.3), it must be that  $f_1$  is a zero-mean function. As for the two-way ANOVA, it is the case that  $\int_{\mathcal{X}_k} f_1(x_1, x_2) d\nu(x_k) = 0, k = 1, 2$ , and  $\int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f_{12}(x_1, x_2) d\nu(x_1) d\nu(x_2) = 0$ .

We notice that the decomposition in (2.3) is orthogonal:

**Claim 2.15.** *For the ANOVA decomposition in (2.3),  $f_0$  and  $f_1$  are orthogonal for the usual  $L^2$  inner product.*

*Proof.* Note that  $f_0$  is a constant function, and that  $f_1 = f - f_0$ . Thus,

$$\begin{aligned}\langle f_0, f_1 \rangle &= \int f_0 f_1 d\nu \\ &= f_0 \int (f - f_0) d\nu \\ &= f_0(f_0 - f_0) = 0.\end{aligned}$$

□

In fact, for  $k = 1$ , any  $f \in \mathcal{F}$  can be decomposed as a sum of a constant plus a zero mean function, so we have the geometric decomposition of the vector space  $\mathcal{F} = \mathcal{F}_0 \oplus^\perp \bar{\mathcal{F}}_1$ , where  $\mathcal{F}_0$  is a vector space of constant functions, and  $\bar{\mathcal{F}}_1$  a vector space of zero-mean functions over  $\mathcal{X}_1$ . For  $k \geq 2$  we can argue something similar. The space  $\mathcal{F}$  has the tensor product structure<sup>4</sup>  $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$ , and considered individually, each  $\mathcal{F}_k$  can be decomposed orthogonally  $\mathcal{F}_k = \mathcal{F}_0 \oplus^\perp \bar{\mathcal{F}}_k$ . Note that  $\mathcal{F}_k$  consists of functions  $f : \mathcal{X}_k \rightarrow \mathbb{R}$ . Expanding out under the distributivity rule of tensor products and rearranging slightly, we obtain

$$\begin{aligned}\mathcal{F} &= (\mathcal{F}_0 \oplus^\perp \bar{\mathcal{F}}_1) \otimes \cdots \otimes (\mathcal{F}_0 \oplus^\perp \bar{\mathcal{F}}_1) \\ &= \mathcal{F}_0^{\otimes p} \oplus^\perp \bigoplus_{j=1}^p \left( \mathcal{F}_0^{\otimes(p-1)} \otimes \bar{\mathcal{F}}_j \right) \oplus^\perp \bigoplus_{\substack{j,k=1 \\ j < k}}^p \left( \mathcal{F}_0^{\otimes(p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k \right) \quad (2.4) \\ &\quad \oplus^\perp \cdots \oplus^\perp \left( \bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p \right).\end{aligned}$$

{eq:funcano  
vaspace}

To clarify,

- $\mathcal{F}_0^{\otimes p}$  is the space of constant functions  $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \rightarrow \mathbb{R}$ .
- $\left( \mathcal{F}_0^{\otimes(p-1)} \otimes \bar{\mathcal{F}}_j \right)$  is the space of functions that are constant on all coordinates except the  $j$ th coordinate of  $x$ . Further, the functions are centred on the  $j$ th coordinate.
- $\left( \mathcal{F}_0^{\otimes(p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k \right)$  is the space of functions that are constant on all coordinates except the  $j$ th and  $k$ th coordinate of  $x$ . Further, the functions are centred on these two coordinates.
- $\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p$  is the space of zero-mean functions  $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \rightarrow \mathbb{R}$ .
- Similarly for the rest of the spaces in the summand, of which there are  $2^p$  members all together.

Therefore, given an arbitrary function  $f \in \mathcal{F}$ , the projection of  $f$  onto the above respective orthogonal spaces in (2.4) leads to the *functional ANOVA representation*

$$f(x) = \mu + \sum_{j=1}^p f_j(x_j) + \sum_{\substack{j,k=1 \\ j < k}}^p f_{jk}(x_j, x_k) + \cdots + f_{1\dots p}(x). \quad (2.5)$$

{eq:functionalanova2}

**Definition 2.36** (Functional ANOVA representation). Let  $\mathcal{P}_d = \mathcal{P}(\{1, \dots, d\})$ , the power set of  $\{1, \dots, d\}$ . For any function  $f \in \mathcal{F} \equiv L^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d, \nu_1 \otimes \cdots \otimes \nu_d)$ , the formula for  $f$  in (2.5) is known as the *functional ANOVA representation* of  $f$  if  $\forall k \in \mathcal{K} \in \mathcal{P}_p$ ,

$$A_k f_{\mathcal{K}} = \int_{\mathcal{X}_{\mathcal{K}}} f_{\mathcal{K}}(x_{\mathcal{K}}) d\nu_{\mathcal{K}}(x_{\mathcal{K}}) = 0, \quad (2.6)$$

{eq:functionanovaorth}

where  $\mathcal{X}_{\mathcal{K}} = \prod_{k \in \mathcal{K}} \mathcal{X}_k$ , and  $x_{\mathcal{K}} = \{x_k, k \in \mathcal{K}\}$  is an element of this space. In other words, the integral of  $f_{\mathcal{K}}$  with respect to any of the variables indexed by the elements in  $\mathcal{K}$  (itself an element of the power set), is zero. The requirement (2.6) ensures orthogonality of the summands in (2.5).

For the constant term, main effects, and two-way interaction terms, the familiar classical expressions are obtained:

$$\begin{aligned} f_0 &= \int f d\nu \\ f_j &= \int f \prod_{i \neq j} d\nu_i - f_0 \\ f_{jk} &= \int f \prod_{i \neq j, k} d\nu_i - f_j - f_k - f_0. \end{aligned}$$

*Remark 2.8.* Not all of the higher order terms need to be included. There may even be a model motivated reason for dropping certain main effects or interaction effects.

## The ANOVA kernel

At last, we come to the section of deriving the ANOVA RKKS, and, rest assured, the preceding long build-up will prove to be not in vain. The main idea is to construct an

<sup>4</sup>There is an isomorphism  $L^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d, \nu_1 \otimes \cdots \otimes \nu_d) \cong L^2(\mathcal{X}_1, \nu_1) \otimes \cdots \otimes L^2(\mathcal{X}_d, \nu_d)$ . See, for example, [reed1972methods](#); [kree1974produits](#).



RKKS such that the functions that lie in them will have the ANOVA representation in (2.5). The bulk of the work has been done, and in fact we know exactly how this ANOVA RKKS should be structured—it is the space as specified in (2.4). The ANOVA RKKS will be constructed by a similar manipulation of the individual kernels representing the RKHS building blocks.

**Definition 2.37** (The ANOVA RKKS). For  $k = 1, \dots, p$ , let  $\mathcal{F}_k$  be a centred RKHS of functions over the set  $\mathcal{X}_k$  with kernel  $h_k : \mathcal{X}_k \times \mathcal{X}_k \rightarrow \mathbb{R}$ . Let  $\lambda_k, k = 1, \dots, p$  be real-valued scale parameters. The ANOVA RKKS of functions  $f : \mathcal{X}_1 \times \dots \times \mathcal{X}_p \rightarrow \mathbb{R}$  is specified by the ANOVA kernel, defined by

$$h_\lambda(x, x') = \prod_{k=1}^p (1 + \lambda_k h_k(x_k, x'_k)). \quad (2.7)$$

The construction an ANOVA RKKS is very very simple in through multiplication of univariate kernels. Expanding out equations (2.7), we see that it is in fact a sum of products of kernels with increasing orders of interaction:

$$\begin{aligned} h_\lambda(x, x') &= 1 + \sum_{j=1}^p \lambda_j h_j(x_j, x'_j) + \sum_{\substack{j,k=1 \\ j < k}}^p \lambda_j \lambda_k h_j(x_j, x'_j) h_k(x_k, x'_k) \\ &\quad + \dots + \prod_{j=1}^p \lambda_j h_j(x_j, x'_j). \end{aligned}$$

It is now clear from the expansion that the ANOVA RKKS yields functions that resemble those with the ANOVA representation in (2.5): The mean value of the function stems from the ‘1’, i.e. it lies in an RKHS of constant functions; the main effects are represented by the sum of the individual kernels; the two-way interaction terms are represented by the second-order kernel interactions; and so on.

**Example 2.2.** Consider two RKKSs  $\mathcal{F}_k$  with kernel  $\lambda_k h_k$ ,  $k = 1, 2$ . The ANOVA kernel defining the ANOVA RKKS  $\mathcal{F}$  is

$$h_\lambda((x_1, x_2), (x'_1, x'_2)) = 1 + \lambda_1 h_1(x_1, x'_1) + \lambda_2 h_2(x_2, x'_2) + \lambda_1 \lambda_2 h_1(x_1, x'_1) h_2(x_2, x'_2).$$

Suppose that  $\mathcal{F}_1$  and  $\mathcal{F}_2$  are the centred canonical RKKS of functions over  $\mathbb{R}$ . Then, functions in  $\mathcal{F} = \mathcal{F}_0 \oplus \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus (\mathcal{F}_1 \otimes \mathcal{F}_2)$  are of the form

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

As remarked in the previous subsection, not all of the components of the ANOVA RKKS need to be included in the construction. The selective exclusion of certain interactions characterises many interesting statistical models. Excluding certain terms of the ANOVA RKKS is equivalent to setting the scale parameter for those relevant components to be zero, i.e., they play no role in the decomposition of the function. With this in mind, the ANOVA RKKS then gives us an objective way of model-building, from linear regression, to multilevel models, longitudinal models, and so on.

## 2.6 Summary

The brief notes on functional analysis allow us to describe the theory of reproducing kernel Hilbert and Kreĭn spaces. These are of great interest to us because the topology endowed on such spaces gives great assurances—in particular, all evaluation functionals are continuous in these spaces. Moreover, RKHS and RKKS can be specified completely through kernel functions, with new and complex function spaces built simply by manipulation of these kernel functions. Of particular importance is the ANOVA functional decomposition, for which we realise provides an objective way of constructing various statistical models (such models will be described later on in detail in Chapter 4).

An annotated collection of bibliographical references used for this chapter is as follows.

- **Functional analysis.** On the introductory material relating to functional analysis in Section 2.1, the lecture notes by [sejdinovic2012](#) is recommended, and forms the basis for most of the material described. Additionally, [rudin1987real](#) provides a complementary reading.
- **RKHS theory.** There are certainly no shortages of introductory texts relating to the theory of RKHS: [steinwart2008support](#), [berlinet2011reproducing](#), and [gu2013smoothing](#) to name a few. The concise sketch proof for the Moore-Aronszajn theorem was mostly inspired by [hein2004kernels](#)

- **RKKS theory.** The innovation of indefinite inner product spaces perhaps started in mathematical physics literature, for which the theory of special relativity depends. Four-dimensional space-time is an often cited example. In any case, we referred to mainly [ong2004learning](#), which gives an overview in the context of learning using indefinite kernels. [alpay1991some](#) and [zafeiriou2012subspace](#) were also useful for understanding the fundamental concepts of RKKS.
- **RKHS building blocks.** The main building block RKHS, i.e. the canonical RKHS, the fBm RKHS and the Pearson RKHS are described in the manuscript of [bergsma2017](#).
- **ANOVA and functional ANOVA.** Classical ANOVA is pretty much existent in every fundamental statistical textbook. These texts have extremely well written introductions to this very important concept: [casella2002statistical](#), [dean1999design](#). On the relation between classical ANOVA and functional ANOVA decomposition, [gu2013smoothing](#) offers novel insights. There is diverse literature concerning functional ANOVA, namely from the fields of statistical learning (e.g. [wahba1990spline](#)), applied mathematics (e.g. [kuo2010decompositions](#)), and sensitivity analysis (e.g. [sobol2001global](#); [durrande2013anova](#)). What is interesting is that several authors who simply set out to obtain a suitable functional decomposition, all ended up somewhat independently recovering the ANOVA decomposition as being “optimal” in some sense. This speaks largely to this classical idea that is ANOVA.

## 2.7 Miscellanea

### 2.7.1 A vector space... of ‘functions’?

At first glance, this may seem strange, that the notion of functions (as mappings from input to output space) and vector spaces are somehow equatable. Upon further thought, one realises that firstly, two functions of a similar, particular form may be added together (in some meaningful way) resulting in a function in that same form. Secondly, multiplication of a function by a scalar  $c$  can be thought of as  $c$  times the output of that function. Indeed, running through the checklist of what constitutes a vector space, we find that a “space of functions” satisfies them all. In modern linear algebra texts, this

checklist is the eight axioms of vector spaces over a field  $\mathbb{F}$ : The vectors forms an abelian group under addition, and this group has an  $\mathbb{F}$ -module structure.