Vector-valued reproducing kernel Banach spaces with applications to multi-task learning[☆]

Haizhang Zhang [a,b,*], Jun Zhang [c]

[a] *School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou 510275, PR China*
[b] *Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou 510275, PR China*
[c] *Department of Psychology, University of Michigan, Ann Arbor, MI 48109, USA*

ARTICLE INFO

ABSTRACT

Motivated by multi-task machine learning with Banach spaces, we propose the notion of vector-valued reproducing kernel Banach spaces (RKBSs). Basic properties of the spaces and the associated reproducing kernels are investigated. We also present feature map constructions and several concrete examples of vector-valued RKBSs. The theory is then applied to multi-task machine learning. Especially, the representer theorem and characterization equations for the minimizer of regularized learning schemes in vector-valued RKBSs are established.

1. Introduction

The purpose of this paper is to establish the notion of vector-valued reproducing kernel Banach spaces and demonstrate its applications to multi-task machine learning. Built on the theory of scalar-valued reproducing kernel Hilbert spaces (RKHSs) [3], kernel methods have been proven successful in single task machine learning [10,14,29,30,33]. Multi-task learning where the unknown target function to be learned from finite sample data is vector-valued appears more often in practice. Refs. [13,25] proposed the development of kernel methods for learning multiple related tasks simultaneously. The mathematical foundation used there was the theory of vector-valued RKHSs [5,27]. Recent progresses

in vector-valued RKHSs can be found in [7–9,38]. In such a framework, both the space of the candidate functions used for approximation and the output space are chosen as a Hilbert space.

There are some occasions where it might be desirable to select the space of candidate functions, the output space, or both as Banach spaces. Hilbert spaces constitute a special and limited class of Banach spaces. Any two Hilbert spaces over a common number field with the same dimension are isometrically isomorphic. By reaching out to other Banach spaces, one obtains more variety in geometric structures and norms that are potentially useful for learning and approximation. Moreover, training data might come with intrinsic structures that make them impossible or inappropriate to be embedded into a Hilbert space. Learning schemes based on features in a Hilbert space may not work well for them. Finally, in some applications, a Banach space norm is engaged for some particular purpose. A typical example is the linear programming regularization in coefficient based regularization for machine learning [29], where the ℓ_1 norm is employed to obtain sparsity in the resulting minimizer.

There has been considerable work in learning a single task with Banach spaces (see, for example, [4,6,12,15,17,20,24,26,34,36,42]). The difficulty in mapping patterns into a Banach space and making use of these features for learning mainly lies in the lack of an inner product in Banach spaces. In particular, without an appropriate correspondence of the Riesz representation of continuous linear functionals, point evaluations do not have a kernel representation in these studies. Semi-inner products, a mathematical tool discovered by Lumer [23] for the purpose of extending Hilbert space type arguments to Banach spaces, seem to be a natural substitute for inner products in Banach spaces. An illustrative example is that we were able to extend the classical theory of frames and Riesz bases to Banach spaces via semi-inner products [40]. Semi-inner products were first used to machine learning by Der and Lee [12] for the study of large margin classification by hyperplanes in a Banach space. With this tool, we established the notion of scalar-valued reproducing kernel Banach spaces (RKBSs) and investigated regularized learning schemes in RKBSs [37,39]. There has been increasing interest in the application of this new theory [19,31,32,41].

We attempt to build a mathematical foundation for multi-task learning with Banach spaces. Specifically, we shall propose a definition of vector-valued RKBSs and investigate its fundamental properties in the next section. Feature map representations and several concrete examples of vector-valued RKBSs will be presented in Sections 3 and 4, respectively. In Section 5, we investigate regularized learning schemes in vector-valued RKBSs.

2. Definition and basic properties

We are concerned with spaces of functions from a fixed set to a vector space. We shall allow the space of functions and the range space both to be a Banach space. Our key tool in dealing with a general Banach space is the semi-inner product [16,23]. Recall that a semi-inner product on a Banach space V is a function from $V \times V$ to $\mathbb{C}$, denoted by $[\cdot, \cdot]_V$, such that for all $f, g, h \in V$ and $\alpha, \beta \in \mathbb{C}$

1. (linearity with respect to the first variable) $[\alpha f + \beta g, h]_V = \alpha[f, h]_V + \beta[g, h]_V$;
2. (positivity) $[f, f]_V > 0$ for $f \neq 0$;
3. (conjugate homogeneity with respect to the second variable) $[f, \alpha g]_V = \overline{\alpha}[f, g]_V$;
4. (Cauchy–Schwarz inequality) $|[f, g]_V| \leq [f, f]_V^{1/2}[g, g]_V^{1/2}$.

A semi-inner product $[\cdot, \cdot]_V$ on V is said to be *compatible* if

$$[f, f]_V^{1/2} = \|f\|_V \quad \text{for all } f \in V,$$

where $\| \cdot \|_V$ denotes the norm on V. Every Banach space has a compatible semi-inner product [16,23]. Let $[\cdot, \cdot]_V$ be a compatible semi-inner product on V. Then one sees by the Cauchy–Schwarz inequality that for each $f \in \mathcal{B}$, the linear functional f^* on V defined by

$$f^*(g) := [g, f]_V, \quad g \in V \tag{2.1}$$

is bounded on V. In other words, f^* lies in the dual space $\mathcal{B}^*$ of $\mathcal{B}$. Moreover, we have

$$\|f^*\|_{V^*} = \|f\|_V \tag{2.2}$$

and

$$f^*(f) = \|f\|_V \|f^*\|_{V^*}. \tag{2.3}$$

Introduce the *duality mapping* $\mathcal{J}_V$ from V to V^* by setting

$$\mathcal{J}_V(f) := f^*, \quad f \in V.$$

We desire to represent the continuous linear functionals on the vector-valued RKBS by the semi-inner product. However, the semi-inner product might not be able to fulfill this important role for an arbitrary Banach space. For instance, one verifies that the continuous linear functional

$$\mu(g) := \sum_{j=1}^{\infty} (-1)^j \frac{1}{2^j} g\left(\frac{1}{2^j}\right), \quad g \in C([0,1])$$

on $C([0,1])$ endowed with the usual maximum norm cannot be represented as

$$\mu(g) = [g, f], \quad g \in C([0,1])$$

for any compatible semi-inner product $[\cdot, \cdot]$ on $C([0,1])$ and any $f \in C([0,1])$.

The above example indicates that the duality mapping might not be surjective for a general Banach space. Other problems such as non-uniqueness of compatible semi-inner products and non-injectivity of the duality mapping may also occur. To overcome these difficulties, we shall focus on Banach spaces that are uniformly convex and uniformly Fréchet differentiable in this preliminary work on vector-valued RKBSs. A Banach space V is *uniformly convex* if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|f + g\|_V \leq 2 - \delta \quad \text{for all } f, g \in V \text{ with } \|f\|_V = \|g\|_V = 1 \text{ and } \|f - g\|_V \geq \varepsilon.$$

Uniform convexity ensures the injectivity of the duality mapping and the existence and uniqueness of the best approximation to a closed convex subset of V [16]. We also say that V is *uniformly Fréchet differentiable* if for all $f, g \in V$

$$\lim_{t \in \mathbb{R}, \, t \to 0} \frac{\|f + tg\|_V - \|f\|_V}{t} \tag{2.4}$$

exists and the limit is approached uniformly for all f, g in the unit ball of V. If V is uniformly Fréchet differentiable then it has a unique compatible semi-inner product [16]. The differentiability (2.4) of the norm is useful to derive characterization equations for the minimizer of regularized learning schemes in Banach spaces. For simplicity, we call a Banach space *uniform* if it is both uniformly convex and uniformly Fréchet differentiable. An analogue of the Riesz representation theorem holds for uniform Banach spaces.

Lemma 2.1 (*Giles [16]*)**.** *Let V be a uniform Banach space. Then it has a unique compatible semi-inner product $[\cdot, \cdot]_V$ and the duality mapping $\mathcal{J}_V$ is bijective from V to V^*. In other words, for each $\mu \in V^*$ there exists a unique $f \in V$ such that*

$$\mu(g) = [g, f]_V \quad \text{for all } g \in V.$$

In this case,

$$[f^*, g^*]_{\mathcal{B}^*} := [g, f]_{\mathcal{B}}, \quad f, g \in \mathcal{B} \tag{2.5}$$

defines a compatible semi-inner product on $\mathcal{B}^*$.

Let V be a uniform Banach space. We shall always denote by $[\cdot, \cdot]_V$ the unique compatible semi-inner product on V. By Lemma 2.1 and Eq. (2.2), the duality mapping is bijective and isometric from V to V^*. It is also conjugate homogeneous by property 3 of semi-inner products. However, it is non-additive unless V reduces to a Hilbert space. As a consequence, a compatible semi-inner product is in general conjugate homogeneous but non-additive with respect to its second variable. Namely,

$$[f, g + h]_V \neq [f, g]_V + [f, h]_V$$

in general.

We are ready to present the definition of vector-valued RKBSs. Let Λ be a Banach space which we shall sometimes call the *output space* and X be a prescribed set which is usually called the *input space*. A space $\mathcal{B}$ is called a *Banach space of Λ-valued functions* on X if it consists of certain functions from X to Λ and the norm on $\mathcal{B}$ is compatible with point evaluations in the sense that

$$\|f\|_{\mathcal{B}} = 0 \quad \text{if and only if} \quad f(x) = 0 \quad \text{for all } x \in X.$$

For instance, $L^p([0, 1]), p \geq 1$ is not a Banach space of functions while $C([0, 1])$ is. We restrict our consideration to Banach spaces of functions so that point evaluations (usually referred to as "sampling" in applications) are well-defined.

Definition 2.2. We call $\mathcal{B}$ a Λ-valued RKBS on X if both $\mathcal{B}$ and Λ are uniform and $\mathcal{B}$ is a Banach space of functions from X to Λ such that for every $x \in X$, the point evaluation $\delta_x : \mathcal{B} \to \Lambda$ defined by

$$\delta_x(f) := f(x), \quad f \in \mathcal{B}$$

is continuous from $\mathcal{B}$ to Λ.

We shall derive a reproducing kernel for so defined vector-valued RKBSs. Throughout the rest of the paper, we let $[\cdot, \cdot]_{\mathcal{B}}$ and $[\cdot, \cdot]_{\Lambda}$ be the unique semi-inner products and $\mathcal{J}_{\mathcal{B}}$ and $\mathcal{J}_{\Lambda}$ the associated duality mappings on $\mathcal{B}$ and Λ, respectively. For two Banach spaces V_1, V_2, we denote by $\mathcal{M}(V_1, V_2)$ the set of all the bounded operators from V_1 to V_2 and $\mathcal{L}(V_1, V_2)$ the subset of $\mathcal{M}(V_1, V_2)$ of those bounded operators that are also linear. When $V_1 = V_2$, $\mathcal{M}(V_1, V_2)$ is abbreviated as $\mathcal{M}(V_1)$. For each $T \in \mathcal{M}(V_1, V_2)$, we denote by $\|T\|_{\mathcal{M}(V_1, V_2)}$ the greatest lower bound of all the nonnegative constants α such that

$$\|Tu\|_{V_2} \leq \alpha \|u\|_{V_1} \quad \text{for all } u \in V_1.$$

When T is also linear, this quantity equals the operator norm $\|T\|_{\mathcal{L}(V_1, V_2)}$ of T in $\mathcal{L}(V_1, V_2)$. In those languages, we require that the point evaluation δ_x on a Λ-valued RKBS on X belongs to $\mathcal{L}(\mathcal{B}, \Lambda)$ for all $x \in X$.

Theorem 2.3. *Let $\mathcal{B}$ be a Λ-valued RKBS on X. Then there exists a unique function K from $X \times X$ to $\mathcal{M}(\Lambda)$ such that*

(1) *$K(x, \cdot)\xi \in \mathcal{B}$ for all $x \in X$ and $\xi \in \Lambda$,*
(2) *for all $f \in \mathcal{B}, x \in X$, and $\xi \in \Lambda$*

$$[f(x), \xi]_{\Lambda} = [f, K(x, \cdot)\xi]_{\mathcal{B}}, \tag{2.6}$$

(3) *for all $x, y \in X$*

$$\|K(x, y)\|_{\mathcal{M}(\Lambda)} \leq \|\delta_x\|_{\mathcal{L}(\mathcal{B}, \Lambda)} \|\delta_y\|_{\mathcal{L}(\mathcal{B}, \Lambda)}. \tag{2.7}$$

Proof. Let $x \in X$ and $\xi \in \Lambda$. As $\delta_x \in \mathcal{L}(\mathcal{B}, \Lambda)$, we see that

$$|[f(x), \xi]_{\Lambda}| \leq \|f(x)\|_{\Lambda} \|\xi\|_{\Lambda} \leq \|\delta_x\|_{\mathcal{L}(\mathcal{B}, \Lambda)} \|f\|_{\mathcal{B}} \|\xi\|_{\Lambda}. \tag{2.8}$$

The above inequality together with the linearity of the semi-inner product with respect to its first variable implies that

$$f \to [f(x), \xi]_{\Lambda}$$

is a bounded linear functional on $\mathcal{B}$. By Lemma 2.1, there exists a unique function $g_{x,\xi} \in \mathcal{B}$ such that

$$[f(x), \xi]_{\Lambda} = [f, g_{x,\xi}]_{\mathcal{B}}. \tag{2.9}$$

Define a function K from $X \times X$ to the set of operators from Λ to Λ by setting

$$K(x, y)\xi := g_{x,\xi}(y), \quad x, y \in X, \ \xi \in \Lambda.$$

Clearly, K satisfies the two requirements (1) and (2). It is also unique by the uniqueness of the function $g_{x,\xi}$ satisfying (2.9). It remains to show that it is bounded. To this end, we get by (2.8) that

$$\|K(x,\cdot)\xi\|_{\mathscr{B}} = \sup_{f\in\mathscr{B},\|f\|_{\mathscr{B}}\leq 1}|[f,K(x,\cdot)]_{\mathscr{B}}| = \sup_{f\in\mathscr{B},\|f\|_{\mathscr{B}}\leq 1}|[f(x),\xi]_{\Lambda}| \leq \|\delta_x\|_{\mathscr{L}(\mathscr{B},\Lambda)}\|\xi\|_{\Lambda}.$$

It follows that

$$\|K(x,y)\xi\|_{\mathscr{B}} \leq \|\delta_y\|_{\mathscr{L}(\mathscr{B},\Lambda)}\|K(x,\cdot)\xi\|_{\mathscr{B}} \leq \|\delta_x\|_{\mathscr{L}(\mathscr{B},\Lambda)}\|\delta_y\|_{\mathscr{L}(\mathscr{B},\Lambda)}\|\xi\|_{\Lambda},$$

which proves (2.7). □

We call the above function K the *reproducing kernel* of $\mathscr{B}$. It coincides with the usual reproducing kernel when $\mathscr{B}$ is a Hilbert space and $\Lambda = \mathbb{C}$, and with the vector-valued reproducing kernel when both $\mathscr{B}$ and Λ are Hilbert spaces. We explore basic properties of vector-valued RKBSs and its reproducing kernels for further investigation and applications.

Let $(\delta_x)^*$ be the adjoint operator of δ_x for all $x \in X$. Denote for a Banach space V by $(\cdot,\cdot)_V$ the bilinear form on $V \times V^*$ defined by

$$(v,\mu)_V := \mu(v), \quad v \in V, \ \mu \in V^*.$$

Thus, $(\delta_x)^*$ is defined by

$$(f,(\delta_x)^*\xi^*)_{\mathscr{B}} = (\delta(x)(f),\xi^*)_{\Lambda} = (f(x),\xi^*)_{\Lambda} = [f(x),\xi]_{\Lambda}, \quad f \in \mathscr{B}, \ \xi \in \Lambda. \tag{2.10}$$

Proposition 2.4. *Let $\mathscr{B}$ be a Λ-valued RKBS on X and K its reproducing kernel. Then there holds for all $x, y \in X$ and $\xi, \eta, \tau \in \Lambda$ that*

$$[K(x,x)\xi,\xi]_{\Lambda} \geq 0, \quad |[K(x,y)\xi,\eta]_{\Lambda}| \leq [K(x,x)\xi,\xi]_{\Lambda}^{1/2}[K(y,y)\eta,\eta]_{\Lambda}^{1/2}, \tag{2.11}$$

$$\|K(x,y)\|_{\mathscr{M}(\Lambda)} \leq \|K(x,x)\|_{\mathscr{M}(\Lambda)}^{1/2}\|K(y,y)\|_{\mathscr{M}(\Lambda)}^{1/2}, \tag{2.12}$$

$$K(x,\cdot)\xi = \mathscr{J}_{\mathscr{B}}^{-1}(\delta_x)^*\mathscr{J}_{\Lambda}(\xi), \tag{2.13}$$

$$K(x,y)(\alpha\xi) = \alpha K(x,y)\xi \quad \textit{for all } \alpha \in \mathbb{C}, \tag{2.14}$$

$$\|K(x,\cdot)\xi\|_{\mathscr{B}} \leq \|\delta_x\|_{\mathscr{L}(\mathscr{B},\Lambda)}\|\xi\|_{\Lambda}, \qquad \|K(x,\cdot)\xi\|_{\mathscr{B}} \leq \|K(x,x)\|_{\mathscr{M}(\Lambda)}^{1/2}\|\xi\|_{\Lambda}, \tag{2.15}$$

$$(K(x,\cdot)\xi)^* + (K(x,\cdot)\eta)^* = (K(x,\cdot)\tau)^* \quad \textit{whenever } \tau^* = \xi^* + \eta^*, \tag{2.16}$$

$$\text{span}\{(K(x,\cdot)\xi)^* : x \in X, \ \xi \in \Lambda\} \textit{ is dense in } \mathscr{B}^*. \tag{2.17}$$

Proof. By (2.6),

$$[K(x,x)\xi,\xi]_{\Lambda} = [K(x,\cdot)\xi,K(x,\cdot)\xi]_{\mathscr{B}} = \|K(x,\cdot)\xi\|_{\mathscr{B}}^2 \geq 0, \tag{2.18}$$

which proves the first inequality in Eq. (2.11). For the second one, we use the Cauchy–Schwarz inequality of semi-inner products to get that

$$|[K(x,y)\xi,\eta]_{\Lambda}| = |[K(x,\cdot)\xi,K(y,\cdot)\eta]_{\mathscr{B}}| \leq [K(x,\cdot)\xi,K(x,\cdot)\xi]_{\mathscr{B}}^{1/2}[K(y,\cdot)\eta,K(y,\cdot)\eta]_{\mathscr{B}}^{1/2}$$
$$= [K(x,x)\xi,\xi]_{\Lambda}^{1/2}[K(y,y)\eta,\eta]_{\Lambda}^{1/2}.$$

It follows from (2.11) that

$$|[K(x,y)\xi,\eta]_{\Lambda}| \leq \|K(x,x)\xi\|_{\Lambda}^{1/2}\|\xi\|_{\Lambda}^{1/2}\|K(y,y)\eta\|_{\Lambda}^{1/2}\|\eta\|_{\Lambda}^{1/2}$$
$$\leq \|K(x,x)\|_{\mathscr{M}(\Lambda)}^{1/2}\|K(y,y)\|_{\mathscr{M}(\Lambda)}^{1/2}\|\xi\|_{\Lambda}\|\eta\|_{\Lambda}.$$

Since $\|K(x,y)\xi\|_{\Lambda} = \sup\{|[K(x,y)\xi,\eta]_{\Lambda}| : \eta \in \Lambda, \|\eta\|_{\Lambda} = 1\}$, we have by the above equation that

$$\|K(x,y)\xi\|_{\Lambda} \leq \|K(x,x)\|_{\mathscr{M}(\Lambda)}^{1/2}\|K(y,y)\|_{\mathscr{M}(\Lambda)}^{1/2}\|\xi\|_{\Lambda},$$

which proves (2.12).

Turning to (2.13), we notice for each $f \in \mathcal{B}$ that

$$[f, \mathcal{J}_{\mathcal{B}}^{-1}(\delta_x)^* \mathcal{J}_\Lambda(\xi)]_{\mathcal{B}} = (f, (\delta_x)^* \mathcal{J}_\Lambda(\xi))_{\mathcal{B}} = (\delta_x(f), \xi^*)_\Lambda = (f(x), \xi^*)_\Lambda = [f(x), \xi]_\Lambda,$$

which together with (2.6) confirms (2.13). Since the duality mappings are conjugate homogeneous, we have by (2.13) that

$$K(x, \cdot)(\alpha\xi) = \mathcal{J}_{\mathcal{B}}^{-1}(\delta_x)^* \mathcal{J}_\Lambda(\alpha\xi) = \alpha \mathcal{J}_{\mathcal{B}}^{-1}(\delta_x)^* \mathcal{J}_\Lambda(\xi) = \alpha K(x, \cdot)\xi,$$

which implies (2.14).

Recall that the duality mappings $\mathcal{J}_{\mathcal{B}}$ and $\mathcal{J}_\Lambda$ are isometric. Note also that a bounded linear operator and its adjoint have equal operator norms. Using these two facts, we obtain from Eq. (2.13) that

$$\|K(x, \cdot)\xi\|_{\mathcal{B}} \leq \|(\delta_x)^*\|_{\mathcal{L}(\Lambda^*, \mathcal{B}^*)} \|\xi\|_\Lambda = \|\delta_x\|_{\mathcal{L}(\mathcal{B}, \Lambda)} \|\xi\|_\Lambda,$$

which is the first inequality in (2.15). The second one follows immediately from (2.18).

Let $\xi, \eta, \tau \in \Lambda$ be such that $\tau^* = \xi^* + \eta^*$. By (2.13),

$$(K(x, \cdot)\xi)^* + (K(x, \cdot)\eta)^* = (\delta_x)^* \xi^* + (\delta_x)^* \eta^* = (\delta_x)^*(\xi^* + \eta^*) = (\delta_x)^* \tau^* = (K(x, \cdot)\tau)^*.$$

Eq. (2.16) hence holds true.

For the last property, let us assume that there exists some $f \in \mathcal{B}$ that vanishes on span $\{(K(x, \cdot)\xi)^* : x \in X, \ \xi \in \Lambda\}$. Then

$$[f(x), \xi]_\Lambda = [f, K(x, \cdot)\xi]_{\mathcal{B}} = (f, (K(x, \cdot)\xi)^*)_{\mathcal{B}} = 0 \quad \text{for all } x \in X, \ \xi \in \Lambda,$$

which implies that $f(x) = 0$ for all $x \in X$. As $\mathcal{B}$ is a Banach space of functions, $f = 0$ as a vector in the Banach space $\mathcal{B}$. Therefore, (2.17) is true. The proof is complete. $\square$

We observe by the above proposition that the reproducing kernel of a vector-valued RKBS enjoys many properties similar to those of the reproducing kernel of a vector-valued RKHS. However, there are many significant differences due to the nature of a semi-inner product. First, although for all $x, y \in X$, $K(x, y)$ remains a homogeneous bounded operator on Λ, it is generally non-additive. This can be seen from (2.13), where $\mathcal{J}_\Lambda$ or $\mathcal{J}_{\mathcal{B}}^{-1}$ is non-additive. Second, it is well-known that when Λ is a Hilbert space, a function $K : X \times X \to \mathcal{L}(\Lambda)$ is the reproducing kernel of some Λ-valued RKHS on X if and only if for all finite $\xi_j \in \Lambda$ and pairwise distinct $x_j \in X, j = 1, 2, \ldots, m$,

$$\sum_{j=1}^{m} \sum_{k=1}^{m} [K(x_j, x_k)\xi_j, \xi_k]_\Lambda \geq 0. \tag{2.19}$$

Although (2.19) still holds for the reproducing kernel of a vector-valued RKBS when $m \leq 2$ and the number field is $\mathbb{R}$, it may cease to be true once the number of sampling points m exceeds 2. An example will be constructed in the next section. Finally, the denseness property (2.17) in the dual space $\mathcal{B}^*$ does not necessarily imply that

$$\overline{\text{span}}\{K(x, \cdot)\xi : x \in X, \ \xi \in \Lambda\} = \mathcal{B}. \tag{2.20}$$

A negative example will also be given in the next section after we present a construction of vector-valued RKBSs through feature maps. Before that, we present another important property of a vector-valued RKBS.

Proposition 2.5. *Let $\mathcal{B}$ be a Λ-valued RKBS on X. Suppose that $f_n \in \mathcal{B}$, $n \in \mathbb{N}$ converges to some $f_0 \in \mathcal{B}$ then $f_n(x)$ converges to $f_0(x)$ in the topology of Λ for each $x \in X$. The convergence is uniform on the set where $\|K(x, x)\|_{\mathcal{M}(\Lambda)}$ is bounded.*

Proof. Suppose that $\|f_n - f\|_{\mathcal{B}}$ converges 0 as n tends to infinity. We get by (2.15) that

$$
\begin{aligned}
\|f_n(x) - f(x)\|_\Lambda &= \sup_{\xi \in \Lambda, \|\xi\|_\Lambda = 1} |[f_n(x) - f(x), \xi]_\Lambda| \\
&= \sup_{\xi \in \Lambda, \|\xi\|_\Lambda = 1} |[f_n - f, K(x, \cdot)\xi]_{\mathcal{B}}| \leq \sup_{\xi \in \Lambda, \|\xi\|_\Lambda = 1} \|f_n - f\|_{\mathcal{B}} \|K(x, \cdot)\xi\|_{\mathcal{B}} \\
&\leq \|f_n - f\|_{\mathcal{B}} \|K(x, x)\|_{\mathcal{M}(\Lambda)}^{1/2}.
\end{aligned}
$$

Therefore, $f_n(x)$ converges pointwise to $f(x)$ on X and the convergence is uniform on the set where $\|K(x, x)\|_{\mathcal{M}(\Lambda)}$ is bounded. $\quad\square$

3. Feature map representations

Feature map representations form the most important way of expressing reproducing kernels. To introduce feature maps for the reproducing kernel of a vector-valued RKBS, we need the notion of the generalized adjoint [22] of a bounded linear operator between Banach spaces. Let V_1, V_2 be two uniform Banach spaces with the compatible semi-inner products $[\cdot, \cdot]_{V_1}$ and $[\cdot, \cdot]_{V_2}$, respectively. The *generalized adjoint* $T^\dagger$ of a $T \in \mathcal{L}(V_1, V_2)$ is an operator in $\mathcal{M}(V_2, V_1)$ defined by

$$[Tu, v]_{V_2} = [u, T^\dagger v]_{V_1}, \quad u \in V_1, \ v \in V_2.$$

It can be verified that

$$T^\dagger = \mathcal{J}_{V_1}^{-1} T^* \mathcal{J}_{V_2}.$$

Thus, $T^\dagger$ is indeed bounded as

$$\|T^\dagger\|_{\mathcal{M}(V_2, V_1)} = \|T^*\|_{\mathcal{L}(V_2^*, V_1^*)} = \|T\|_{\mathcal{L}(V_1, V_2)}.$$

We are in a position to present a characterization of the reproducing kernel of a vector-valued RKBS.

Theorem 3.1. *A function $K : X \times X \to \mathcal{M}(\Lambda)$ is the reproducing kernel of some Λ-valued RKBS on X if and only if there exist a uniform Banach space $\mathcal{W}$ and a mapping $\Phi : X \to \mathcal{L}(\mathcal{W}, \Lambda)$ such that*

$$K(x, y) = \Phi(y)\Phi^\dagger(x), \quad x, y \in X, \tag{3.1}$$

and

$$\overline{\mathrm{span}}\{(\Phi^\dagger(x)\xi)^* : x \in X, \ \xi \in \Lambda\} = \mathcal{W}^*. \tag{3.2}$$

Here $\Phi^\dagger$ is the function from X to $\mathcal{M}(\Lambda, \mathcal{W})$ defined by $\Phi^\dagger(x) := (\Phi(x))^\dagger, x \in X$.

Proof. Suppose that K is the reproducing kernel of some Λ-valued RKBS $\mathcal{B}$ on X. Set $\mathcal{W} := \mathcal{B}$ and define $\Phi : X \to \mathcal{L}(\mathcal{W}, \Lambda)$ by

$$(\Phi(x))(f) := f(x), \quad f \in \mathcal{B}, \ x \in X.$$

To identify $\Phi^\dagger$, we observe by the reproducing property (2.6) for all $\xi \in \Lambda$ and $f \in \mathcal{B}$ that

$$[f, \Phi^\dagger(x)\xi]_{\mathcal{B}} = [(\Phi(x))f, \xi]_\Lambda = [f(x), \xi]_\Lambda = [f, K(x, \cdot)\xi]_{\mathcal{B}}, \quad x \in X, \ \xi \in \Lambda,$$

which implies that $\Phi^\dagger(x)\xi = K(x, \cdot)\xi$ for all $x \in X$ and $\xi \in \Lambda$. Requirement (3.2) is fulfilled by (2.17). By the forms of Φ and $\Phi^\dagger$, we obtain that

$$\Phi(y)\Phi^\dagger(x)\xi = \Phi(y)(K(x, \cdot)\xi) = K(x, y)\xi,$$

which proves (3.1).

On the other hand, suppose that K is of the form (3.1) in terms of some mapping Φ satisfying the denseness condition (3.2). We shall construct the RKBS that takes K as its reproducing kernel. For this purpose, we let $\mathcal{B}$ be composed of functions from X to Λ of the following form

$$f_u(x) := \Phi(x)u, \quad x \in X \text{ for some } u \in \mathcal{W}.$$

Since each $\Phi(x)$ is a linear operator, $\mathcal{B}$ is a linear vector space. We impose a norm on $\mathcal{B}$ by setting

$$\|f_u\|_{\mathcal{B}} := \|u\|_{\mathcal{W}}, \quad u \in \mathcal{W}.$$

To verify that this is a well-defined norm, it suffices to show that the representer u of a function $f_u \in \mathcal{B}$ is unique. Assume that $f_u = 0$. Then for all $x \in X$ and $\xi \in \Lambda$,

$$(u, (\Phi^\dagger(x)\xi)^*)_{\mathcal{W}} = [u, \Phi^\dagger(x)\xi]_{\mathcal{W}} = [\Phi(x)u, \xi]_\Lambda = [0, \xi]_\Lambda = 0,$$

which combined with (3.2) implies that $u = 0$. The arguments also show that $\mathcal{B}$ is a Banach space of functions. Moreover, it is a uniform Banach space as it is isometrically isomorphic to $\mathcal{W}$. Clearly, we have for each $x \in X$ and $u \in \mathcal{W}$ that

$$\|f_u(x)\|_\Lambda = \|\Phi(x)u\|_\Lambda \leq \|\Phi(x)\|_{\mathcal{L}(\mathcal{W},\Lambda)}\|u\|_\mathcal{W} = \|\Phi(x)\|_{\mathcal{L}(\mathcal{W},\Lambda)}\|f_u\|_\mathcal{B},$$

which shows that point evaluations are bounded on $\mathcal{B}$. We conclude that $\mathcal{B}$ is a Λ-valued RKBS on X. It remains to prove that K is the reproducing kernel of $\mathcal{B}$. To this end, we identify the unique compatible semi-inner product on $\mathcal{B}$ as

$$[f_u, f_v]_\mathcal{B} := [u, v]_\mathcal{W}, \quad u, v \in \mathcal{W},$$

and observe for all $u \in \mathcal{W}$ and $x \in X$ that

$$[f_u, K(x, \cdot)\xi]_\mathcal{B} = [f_u, \Phi(\cdot)\Phi^\dagger(x)\xi]_\mathcal{B} = [u, \Phi^\dagger(x)\xi]_\mathcal{W} = [\Phi(x)u, \xi]_\Lambda = [f_u(x), \xi]_\Lambda,$$

which is what we want. The proof is complete. $\quad\square$

We call the Banach space $\mathcal{W}$ and the mapping Φ in Theorem 3.1 a pair of *feature space* and *feature map* for K, respectively. The proof of Theorem 3.1 contains a construction of vector-valued RKBSs by feature maps, which we pull out separately as a corollary below.

Corollary 3.2. *Let $\mathcal{W}$ be a uniform Banach space and $\Phi : X \to \mathcal{L}(\mathcal{W}, \Lambda)$ be a feature map of K that satisfies (3.1) and (3.2). Then the linear vector space*

$$\mathcal{B} := \{\Phi(\cdot)u : u \in \mathcal{W}\}$$

endowed with the norm

$$\|\Phi(\cdot)u\|_\mathcal{B} := \|u\|_\mathcal{W}, \quad u \in \mathcal{W}$$

and compatible semi-inner product

$$[\Phi(\cdot)u, \Phi(\cdot)v]_\mathcal{B} := [u, v]_\mathcal{W}, \quad u, v \in \mathcal{W}$$

is a Λ-valued RKBS on X with the reproducing kernel K given by (3.1).

As an interesting application of Corollary 3.2, we shall show that a vector-valued RKBS is always isometrically isomorphic to a scalar-valued RKBS on a different input space.

Corollary 3.3. *If $\mathcal{B}$ is a Λ-valued RKBS on X then the following linear vector space $\tilde{\mathcal{B}}$ of complex-valued functions $\tilde{f}$ on $\tilde{X} := X \times \Lambda$ of the form*

$$\tilde{f}(x, \xi) := [f(x), \xi]_\Lambda, \quad x \in X, \ \xi \in \Lambda, \ f \in \mathcal{B}$$

is an RKBS on $\tilde{X}$ with the norm

$$\|\tilde{f}\|_{\tilde{\mathcal{B}}} := \|f\|_\mathcal{B}, \quad f \in \mathcal{B}$$

and the compatible semi-inner product

$$[\tilde{f}, \tilde{g}]_{\tilde{\mathcal{B}}} := [f, g]_\mathcal{B}, \quad f, g \in \mathcal{B}.$$

The reproducing kernel $\tilde{K}$ of $\tilde{\mathcal{B}}$ is

$$\tilde{K}((x, \xi), (y, \eta)) := [K(x, y)\xi, \eta]_\Lambda, \quad x, y \in X, \ \xi, \eta \in \Lambda.$$

Proof. It suffices to point out that $\tilde{\mathcal{B}}$ is constructed by Corollary 3.2 via the choices

$$\Lambda := \mathbb{C}, \qquad \mathcal{W} := \mathcal{B}, \qquad \Phi(x, \xi) := (K(x, \cdot)\xi)^*, \quad (x, \xi) \in \tilde{X}.$$

The feature map satisfies the denseness condition by (2.17). $\quad\square$

We shall next construct by Corollary 3.2 a simple vector-valued RKBS to show that the reproducing kernel of a general vector-valued RKBS might not satisfy (2.19) or (2.20). Let $p, q, r, s \in (1, +\infty)$ satisfy that

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + \frac{1}{s} = 1. \tag{3.3}$$

Here, for the sake of convenience in enumerating elements from a finite set, we set $\mathbb{N}_l := \{1, 2, \ldots, l\}$ for $l \in \mathbb{N}$. For each $\gamma \in (1, +\infty)$ and $l \in \mathbb{N}$, ℓ_γ^l denotes the Banach space of all vectors $u = (u_j : j \in \mathbb{N}_l) \in \mathbb{C}^l$ with the norm

$$\|u\|_{\ell_\gamma^l} := \left(\sum_{j=1}^l |u_j|^\gamma \right)^{1/\gamma} < +\infty.$$

The space ℓ_γ^l is a uniform Banach space with the compatible semi-inner product

$$[u, v]_{\ell_\gamma^l} := \sum_{j=1}^l \frac{u_j \overline{v_j} |v_j|^{\gamma-2}}{\|v\|_{\ell_\gamma^l}^{\gamma-2}}, \quad u, v \in \ell_\gamma^l.$$

The dual element u^* of $u \in \ell_\gamma^l$ is hence given by

$$u^* := \left(\frac{\overline{v_j} |v_j|^{\gamma-2}}{\|v\|_{\ell_\gamma^l}^{\gamma-2}} : j \in \mathbb{N}_l \right), \quad u \in \ell_\gamma^l. \tag{3.4}$$

Non-completeness of the linear span of the reproducing kernel in $\mathcal{B}$. We give a counterexample of (2.20) first. Let $m, n \in \mathbb{N}$. We choose the output space Λ and feature space $\mathcal{W}$ as ℓ_p^n and ℓ_r^m, respectively. Thus, we have that $\Lambda^* = \ell_q^n$ and $\mathcal{W}^* = \ell_s^m$. The input space will be chosen as a set of m discrete points $X := \{x_j : j \in \mathbb{N}_m\}$. A feature map $\Phi : X \to \mathcal{L}(\mathcal{W}, \Lambda)$ should satisfy the denseness condition (3.2). We note by the definition of the generalized adjoint that this condition is equivalent to

$$\overline{\text{span}}\{\Phi^*(x)\xi^* : x \in X, \xi \in \Lambda\} = \mathcal{W}^*, \tag{3.5}$$

where $\Phi^*(x) := (\Phi(x))^*$ for all $x \in X$.

Let us take a close look at Eq. (2.20). By Corollary 3.2, a general function in $\mathcal{B}$ is of the form $f_u := \Phi(\cdot)u$ for some $u \in \mathcal{W}$. Eq. (2.20) does not hold true if and only if there exists a nontrivial $u \in \mathcal{W}$ such that

$$[K(x, \cdot)\xi, f_u]_{\mathcal{B}} = [\Phi(\cdot)\Phi^\dagger(x)\xi, \Phi(\cdot)u]_{\mathcal{B}} = [\Phi^\dagger(x)\xi, u]_{\mathcal{W}} = 0,$$

which in turn is equivalent to that $\text{span}\{\Phi^\dagger(x)\xi : x \in X, \xi \in \Lambda\}$ is not dense in $\mathcal{W}$. We conclude that to construct a Λ-valued RKBS for which (2.20) is not true, it suffices to find a feature map $\Phi : X \to \mathcal{L}(\mathcal{W}, \Lambda)$ that satisfies (3.5) but

$$\overline{\text{span}}\{\Phi^\dagger(x)\xi : x \in X, \xi \in \Lambda\} \subsetneqq \mathcal{W}. \tag{3.6}$$

To this end, we find a sequence of vectors $w_j \in \mathbb{C}^m$ and set

$$\Phi^*(x_j)\xi^* := (\xi^*)_1 w_j, \quad j \in \mathbb{N}_m, \tag{3.7}$$

where $(\xi^*)_1$ is the first component of the vector $\xi^* \in \mathbb{C}^n$. Since for each $j \in \mathbb{N}_m$, $\Phi^*(x_j)$ is a linear operator from Λ^* to $\mathcal{W}^*$ and both the spaces are finite-dimensional, $\Phi^*(x_j)$ is bounded. We reformulate (3.5) and (3.6) to get that they are respectively equivalent to

$$\text{span}\{w_j : j \in \mathbb{N}_m\} = \mathbb{C}^m \tag{3.8}$$

and

$$\text{span}\{\mathcal{J}_{\mathcal{W}}^{-1} w_j : j \in \mathbb{N}_m\} \subsetneqq \mathbb{C}^m. \tag{3.9}$$

Here for a vector $u = (u_j : j \in \mathbb{N}_m) \in \mathbb{C}^m$, we get by (3.4) that

$$\mathcal{J}_w^{-1} u = \left(\frac{\overline{u_j} |u_j|^{s-2}}{\|u\|_{\ell_s^m}^{s-2}} : j \in \mathbb{N}_m \right).$$

Therefore, the task reduces to the searching of an $m \times m$ nonsingular matrix A that becomes singular when we apply the function $t \to \overline{t}|t|^{s-2}$ to each of its components. We find two such matrices as shown below

$$m = 4, \ s = 4, \ A_1 := \begin{bmatrix} 0 & 8 & 2 & 4 \\ 5 & 0 & 5 & 1 \\ 5 & 4 & 6 & 9 \\ 0 & 9 & 4 & 8 \end{bmatrix}, \quad \text{and} \quad m = 4, \ s = 5, \ A_2 := \begin{bmatrix} 9 & 9 & 9 & 9 \\ 8 & 6 & 0 & 2 \\ 6 & 9 & 2 & 1 \\ 7 & 4 & 9 & 9 \end{bmatrix}.$$

Non-positive-definiteness of the reproducing kernel of $\mathcal{B}$. We shall give an example to show that (2.19) might not hold true for the reproducing kernel of a vector-valued RKBS when the number m of sampling points exceeds 2. In fact, we let $m = 3$ and $\mathcal{B}$ be constructed as in the above example with $\{w_j : j \in \mathbb{N}_3\}$ to be appropriately chosen in the definition (3.7) of Φ^*. Our purpose is to find $w_j \in \mathbb{C}^3$ and $\xi_j \in \Lambda, j \in \mathbb{N}_3$ such that

$$\sum_{j=1}^{3} \sum_{k=1}^{3} [K(x_j, x_k)\xi_j, \xi_k]_{\mathcal{B}} < 0. \tag{3.10}$$

We first note for all $j, k \in \mathbb{N}_3$ that

$$[K(x_j, x_k)\xi_j, \xi_k]_{\Lambda} = [\Phi(x_k)\Phi^{\dagger}(x_j)\xi_j, \xi_k]_{\Lambda} = [\Phi^{\dagger}(x_j)\xi_j, \Phi^{\dagger}(x_k)\xi_k]_{\Lambda}$$
$$= [(\Phi^{\dagger}(x_k)\xi_k)^*, (\Phi^{\dagger}(x_j)\xi_j)^*]_{\Lambda^*} = [\Phi^*(x_k)(\xi_k)^*, \Phi^*(x_j)(\xi_j)^*]_{\Lambda^*}.$$

We shall choose $\xi_j \in \Lambda$ so that $((\xi_j)^*)_1 = 1$ for each $j \in \mathbb{N}_3$. With the choice, we obtain by (3.7) and the above equation that

$$\sum_{j=1}^{3} \sum_{k=1}^{3} [K(x_j, x_k)\xi_j, \xi_k]_{\mathcal{B}} = \sum_{j=1}^{3} \sum_{k=1}^{3} [w_k, w_j]_{\ell_s^3}.$$

The conclusion is that for (3.10) to hold, it suffices to find $w_j \in \mathbb{C}^3, j \in \mathbb{N}_3$ that form a basis for $\mathbb{C}^3$ but

$$\sum_{j=1}^{3} \sum_{k=1}^{3} [w_k, w_j]_{\ell_s^3} < 0.$$

Two examples are shown below

$$s = 4, \ [w_1, w_2, w_3] = \begin{bmatrix} 4 & -2 & -3 \\ 3 & -5 & 4 \\ 1 & -1 & 1 \end{bmatrix}, \quad \text{and} \quad s = 5, \ [w_1, w_2, w_3] = \begin{bmatrix} 3 & 2 & -3 \\ 2 & -3 & 3 \\ -5 & 0 & 4 \end{bmatrix}.$$

4. Examples of vector-valued RKBSs

We present several examples of vector-valued RKBSs in this section. The first one of them is applicable to learning a sensing matrix.

4.1. The space of sensing matrices

Spaces involved in this example are all over the field $\mathbb{R}$ of real numbers. The input space and the output space are chosen by $X := \mathbb{R}^d$ and $\Lambda := \mathbb{R}^n$. The vector-valued RKBS $\mathcal{B}$ consists of all the $n \times d$ real matrices. Each $A \in \mathcal{B}$ is considered to be a function from $\mathbb{R}^d$ to $\mathbb{R}^n$ with the point evaluation

$$A(x) := Ax, \quad x \in \mathbb{R}^d.$$

To find a norm that makes $\mathcal{B}$ a uniform Banach space, we first point out that a finite-dimensional Banach space V is uniform if and only if its norm is strictly convex. For a proof of this simple fact, see,

for example, [40]. Recall that $\| \cdot \|_V$ is said to be *strictly convex* if for all $u, v \in V \setminus \{0\}$, $\|u + v\|_V = \|u\|_V + \|v\|_V$ always implies that $u = \alpha v$ for some $\alpha > 0$. Strictly convex norms on $\mathscr{B}$ include

- column-wise norms:

$$\|A\|_{\mathscr{B}} := G(\|a_1\|_1, \|a_2\|_2, \ldots, \|a_d\|_d), \quad A \in \mathscr{B}, \tag{4.1}$$

where for each $j \in \mathbb{N}_d$, a_j is the j-th column of A and $\| \cdot \|_j$ is a strictly convex norm on $\mathbb{R}^n$, and G is a strictly convex function from $\mathbb{R}_+^d$ to $\mathbb{R}_+ := [0, \infty)$ that is strictly increasing with respect to each of its variables and is homogeneous in the sense that

$$G(\alpha x) = \alpha G(x) \quad \text{for all } x \in \mathbb{R}_+^d \text{ and } \alpha \in \mathbb{R}_+.$$

It is straightforward to verify that under the above conditions, (4.1) is indeed a strictly convex norm on $\mathscr{B}$. An explicit instance is

$$\|A\|_{\mathscr{B}} := \|(\|a_j\|_{\ell_p^n} : j \in \mathbb{N}_d)\|_{\ell_r^d}, \quad A \in \mathscr{B}, \tag{4.2}$$

where $p, r \in (1, +\infty)$. One can easily transform a column-wise norm $\| \cdot \|_{\mathscr{B}}$ into a row-wise norm by equipping $A \in \mathscr{B}$ with $\|A^T\|_{\mathscr{B}}$, where A^T is the transpose of A.

- the p-th Schatten norm (see, Section 3.5 of [18]):

$$\|A\|_{\mathscr{B}} := \left(\sum_{j=1}^{\min(n,d)} (\sigma_j(A))^p \right)^{1/p}, \quad A \in \mathscr{B}, \ p \in (1, +\infty),$$

where $\sigma_j(A)$ is the j-th singular value of A. The p-th Schatten norm belongs to the class of matrix norms that are invariant under multiplication by unitary matrices.

We shall look at the reproducing kernel of $\mathscr{B}$ when it is endowed with the norm (4.2) and the output space $\mathbb{R}^n$ is equipped with the norm of ℓ_γ^n for some $\gamma \in (1, +\infty)$. Let q, s be the conjugate numbers of p and r, respectively. In other words, they satisfy (3.3). We proceed by (2.6) that

$$(Ax, \xi^*)_{\ell_\gamma^n} = [A, K(x, \cdot)\xi]_{\mathscr{B}} = (A, (K(x, \cdot)\xi)^*)_{\mathscr{B}}, \quad A \in \mathscr{B}, \ x \in \mathbb{R}^d, \ \xi \in \mathbb{R}^n,$$

which implies that

$$(K(x, \cdot)\xi)^* = \xi^* x^T, \quad x \in \mathbb{R}^d, \ \xi \in \mathbb{R}^n. \tag{4.3}$$

The dual element of $A \in \mathscr{B}$ is given by

$$A^* = \frac{1}{\|A\|_{\mathscr{B}}^{r-2}} \left[a_j^* \|a_j\|_{\ell_p^n}^{r-2} : j \in \mathbb{N}_d \right],$$

where a_j^* is the dual vector of a_j in ℓ_p^n. The reproducing kernel of $\mathscr{B}$ can be derived from the above two equation. Its explicit form is too complicated to be presented. We shall see from the study of regularized learning schemes in vector-valued RKBSs that the identification (4.3) of its dual is usually more important.

4.2. Tensor products of scalar-valued RKBSs

Let $n \in \mathbb{N}$ and $\mathscr{B}_j, j \in \mathbb{N}_n$ be scalar-valued RKBSs on an input space X. We let $\mathscr{B}$ be the tensor product of $\mathscr{B}_j, j \in \mathbb{N}_n$. Thus, it consists of $\mathbb{C}^n$-valued functions of the form $f = (f_j \in \mathscr{B}_j : j \in \mathbb{N}_n)$. To define a norm on $\mathscr{B}$, we choose functions $\mathscr{N}, \mathscr{N}^*$ from $\mathbb{R}_+^n$ to $\mathbb{R}_+$ that are strictly convex, strictly increasing with respect to each of the variables, homogeneous, and satisfy that $x \to \mathscr{N}^*(|x|)$ is the dual norm of $x \to \mathscr{N}(|x|)$ on $\mathbb{R}^n$. Here, $|x| := (|x_j| : j \in \mathbb{N}_n)$ for each $x \in \mathbb{R}^n$. An example is

$$\mathscr{N}(x) := \|x\|_{\ell_p^n}, \qquad \mathscr{N}^*(x) := \|x\|_{\ell_q^n}, \quad x \in \mathbb{R}_+^n,$$

where p, q are a pair of conjugate numbers in $(1, +\infty)$. With such two gauge functions, we impose the following norm on $\mathscr{B}$

$$\|f\|_{\mathscr{B}} := \mathscr{N}(\|f_1\|_{\mathscr{B}_1}, \|f_2\|_{\mathscr{B}_2}, \ldots, \|f_n\|_{\mathscr{B}_n}), \quad f \in \mathscr{B}. \tag{4.4}$$

Proposition 4.1. *The tensor product space $\mathscr{B}$ with the norm (4.4) is a uniform Banach space.*

Proof. We first show that (4.4) defines a uniform convex norm on $\mathcal{B}$. It is straightforward to verify that it is a norm. Let ε be a fixed positive number and $f, g \in \mathcal{B}$ be such that $\|f\|_{\mathcal{B}} = \|g\|_{\mathcal{B}} = 1$ and $\|f - g\|_{\mathcal{B}} \geq \varepsilon$. We have that

$$\mathcal{N}(\|f_1 + g_1\|_{\mathcal{B}_1}, \ldots, \|f_n + g_n\|_{\mathcal{B}_n}) \leq \mathcal{N}(\|f_1\|_{\mathcal{B}_1} + \|g_1\|_{\mathcal{B}_1}, \ldots, \|f_n\|_{\mathcal{B}_n} + \|g_n\|_{\mathcal{B}_n})$$
$$\leq \mathcal{N}(\|f_1\|_{\mathcal{B}_1}, \ldots, \|f_n\|_{\mathcal{B}_n}) + \mathcal{N}(\|g_1\|_{\mathcal{B}_1}, \ldots, \|g_n\|_{\mathcal{B}_n}).$$

As all the norms on $\mathbb{R}^n$ are equivalent, $\mathcal{N}$ is continuous on $\mathbb{R}^n_+$, and vectors $x \in \mathbb{R}^n_+$ satisfying $\mathcal{N}(|x|) = 1$ form a compact subset in $\mathbb{R}^n$. We also recall that $\mathcal{N}$ is strictly increasing with respect to each of its variables and $|x| \to \mathcal{N}(|x|)$ is a strictly convex norm on $\mathbb{R}^n$. We conclude from these two facts and the above equation that $\mathcal{B}$ is uniform convex if there exists some positive constant ε' independent of f, g such that

$$\max\{\|f_j\|_{\mathcal{B}_j} + \|g_j\|_{\mathcal{B}_j} - \|f_j + g_j\|_{\mathcal{B}_j} : j \in \mathbb{N}_n\} \geq \varepsilon'$$

or

$$\max\{|\|f_j\|_{\mathcal{B}_j} - \|g_j\|_{\mathcal{B}_j}| : j \in \mathbb{N}_n\} \geq \varepsilon'.$$

Assume to the contrary that such a positive constant does not exist. It implies that for all $\beta > 0$, there exist $f, g \in \mathcal{B}$ that satisfy $\|f - g\|_{\mathcal{B}} \geq \varepsilon$ and

$$\|f_j\|_{\mathcal{B}_j} + \|g_j\|_{\mathcal{B}_j} - \|f_j + g_j\|_{\mathcal{B}_j} < \beta, \qquad |\|f_j\|_{\mathcal{B}_j} - \|g_j\|_{\mathcal{B}_j}| < \beta \quad \text{for all } j \in \mathbb{N}_n.$$

Again, as any two norms on $\mathbb{R}^n$ are equivalent, the inequality $\|f - g\|_{\mathcal{B}} \geq \varepsilon$ implies that $\|f_k - g_k\|_{\mathcal{B}_k} \geq \varepsilon_0 > 0$ for some $k \in \mathbb{N}_n$ and some positive constant ε_0 independent of f, g. The conclusion is that there exist some $k \in \mathbb{N}_n$ and some positive constants $M, \varepsilon_0 > 0$ such that for all $\beta > 0$, there exist $u, v \in \mathcal{B}_k$ such that $\|u\|_{\mathcal{B}_k} \leq M, \|v\|_{\mathcal{B}_k} \leq M$ and

$$\|u - v\|_{\mathcal{B}_k} \geq \varepsilon_0, \qquad |\|u\|_{\mathcal{B}_k} - \|v\|_{\mathcal{B}_k}| < \beta, \qquad \|u\|_{\mathcal{B}_k} + \|v\|_{\mathcal{B}_k} - \|u + v\|_{\mathcal{B}_k} < \beta. \qquad (4.5)$$

We shall show that the above equation contradicts the uniform convexity of $\mathcal{B}_k$. We may choose β so small that $\beta < \varepsilon_0/4$. It follows from the first two inequalities of (4.5) that

$$\|u\|_{\mathcal{B}_k} \geq \frac{\varepsilon_0}{4}, \qquad \|v\|_{\mathcal{B}_k} \geq \frac{\varepsilon_0}{4}. \qquad (4.6)$$

To proceed, we estimate that

$$\left\| \frac{u}{\|u\|_{\mathcal{B}_k}} - \frac{v}{\|v\|_{\mathcal{B}_k}} \right\|_{\mathcal{B}_k} = \left\| \frac{u}{\|u\|_{\mathcal{B}_k}} - \frac{v}{\|u\|_{\mathcal{B}_k}} + \frac{v}{\|u\|_{\mathcal{B}_k}} - \frac{v}{\|v\|_{\mathcal{B}_k}} \right\|_{\mathcal{B}_k}$$
$$\geq \frac{1}{\|u\|_{\mathcal{B}_k}} \|u - v\|_{\mathcal{B}_k} - \|v\|_{\mathcal{B}_k} \left| \frac{1}{\|u\|_{\mathcal{B}_k}} - \frac{1}{\|v\|_{\mathcal{B}_k}} \right|$$
$$\geq \frac{\varepsilon_0 - \beta}{\|u\|_{\mathcal{B}_k}} \geq \frac{3\varepsilon_0}{4M}.$$

By the uniform convexity of $\mathcal{B}_k$, there exist a positive constant δ dependent on ε_0, M and the space $\mathcal{B}_k$ only such that

$$\left\| \frac{u}{\|u\|_{\mathcal{B}_k}} + \frac{v}{\|v\|_{\mathcal{B}_k}} \right\|_{\mathcal{B}_k} < 2 - \delta. \qquad (4.7)$$

Finally, we get by (4.5)–(4.7) that

$$\|u\|_{\mathcal{B}_k} + \|v\|_{\mathcal{B}_k} - \|u + v\|_{\mathcal{B}_k}$$
$$= \|u\|_{\mathcal{B}_k} + \|v\|_{\mathcal{B}_k} - \|u\|_{\mathcal{B}_k} \left\| \frac{u}{\|u\|_{\mathcal{B}_k}} + \frac{v}{\|v\|_{\mathcal{B}_k}} + \frac{v}{\|u\|_{\mathcal{B}_k}} - \frac{v}{\|v\|_{\mathcal{B}_k}} \right\|_{\mathcal{B}_k}$$

$$\geq \|u\|_{\mathcal{B}_k} + \|v\|_{\mathcal{B}_k} - (2 - \delta)\|u\|_{\mathcal{B}_k} - \|u\|_{\mathcal{B}_k}\|v\|_{\mathcal{B}_k} \left| \frac{1}{\|u\|_{\mathcal{B}_k}} - \frac{1}{\|v\|_{\mathcal{B}_k}} \right|$$

$$\geq \|u\|_{\mathcal{B}_k} + \|v\|_{\mathcal{B}_k} + \left| \|u\|_{\mathcal{B}_k} - \|v\|_{\mathcal{B}_k} \right| - (2 - \delta)\|u\|_{\mathcal{B}_k}$$

$$\geq \delta\|u\|_{\mathcal{B}_k} \geq \frac{\varepsilon_0 \delta}{4},$$

which contradicts to the third inequality of (4.5) as β can be arbitrarily small.

It is clear that $\mathcal{B}^* = \{(f_j^* : j \in \mathbb{N}_n) : f \in \mathcal{B}\}$ with the norm

$$\|(f_j^* : j \in \mathbb{N}_n)\|_{\mathcal{B}^*} = \mathcal{N}^*(\|f_1^*\|_{\mathcal{B}_1^*}, \ldots, \|f_n^*\|_{\mathcal{B}_n^*}).$$

Similar arguments to those above prove that $\mathcal{B}^*$ is uniformly convex. By the fact (see [11]) that a Banach space is uniformly Fréchet differentiable if and only if its dual is uniformly convex, $\mathcal{B}$ is uniform. $\square$

We next identify the reproducing kernel of $\mathcal{B}$ with the following norm

$$\|f\|_{\mathcal{B}} := \left(\sum_{j=1}^{n} \|f_j\|_{\mathcal{B}_j}^p \right)^{1/p}, \quad f \in \mathcal{B}.$$

Let the output space $\mathbb{C}^n$ be equipped with the norm of ℓ_r^n and let K_j be the reproducing kernel of $\mathcal{B}_j, j \in \mathbb{N}_n$. The unique compatible semi-inner product on $\mathcal{B}$ is given by

$$[f, g]_{\mathcal{B}} := \frac{1}{\|g\|_{\mathcal{B}}^{p-2}} \sum_{j=1}^{n} [f_j, g_j]_{\mathcal{B}_j} \|g_j\|_{\mathcal{B}_j}^{p-2}, \quad f, g \in \mathcal{B}.$$

The duality mapping on $\mathcal{B}$ is hence of the form

$$f^* := \left(\frac{f_j^* \|f_j\|_{\mathcal{B}_j}^{p-2}}{\|f\|_{\mathcal{B}}^{p-2}} : j \in \mathbb{N}_n \right), \quad f \in \mathcal{B}. \tag{4.8}$$

To find an expression for $(K(x, \cdot)\xi)^*$ for $x \in X$ and $\xi \in \mathbb{C}^n$, we deduce that

$$[f(x), \xi]_{\ell_r^n} = \frac{1}{\|\xi\|_{\ell_r^n}^{r-2}} \sum_{j=1}^{n} \overline{\xi}_j |\xi_j|^{r-2} f_j(x) = \frac{1}{\|\xi\|_{\ell_r^n}^{r-2}} \sum_{j=1}^{n} \overline{\xi}_j |\xi_j|^{r-2} [f_j, K_j(x, \cdot)]_{\mathcal{B}_j}.$$

It follows that

$$(K(x, \cdot)\xi)^* = \left(\frac{\overline{\xi}_j |\xi_j|^{r-2}}{\|\xi\|_{\ell_r^n}^{r-2}} (K_j(x, \cdot))^* : j \in \mathbb{N}_n \right), \quad x \in X, \ \xi \in \mathbb{C}^n. \tag{4.9}$$

By Eqs. (4.8) and (4.9),

$$\|K(x, \cdot)\xi\|_{\mathcal{B}} = \frac{1}{\|\xi\|_{\ell_r^n}^{r-2}} \left(\sum_{j=1}^{n} \left(|\xi_j|^{r-1} \sqrt{K_j(x, x)} \right)^q \right)^{1/q}, \quad x \in X, \ \xi \in \mathbb{C}^n$$

and

$$K(x, y)\xi = \left(\frac{\xi_j}{|\xi_j|} K_j(x, y) \left(\frac{\|K(x, \cdot)\xi\|_{\mathcal{B}}^{p-2} |\xi_j|^{r-1}}{\|\xi\|_{\ell_r^n}^{r-2} K_j(x, x)^{\frac{p-2}{2}}} \right)^{1/(p-1)} : j \in \mathbb{N}_n \right), \quad x, y \in X, \ \xi \in \mathbb{C}^n.$$

4.3. Translation invariant vector-valued RKBSs

A $\mathbb{C}^n$-valued RKBS $\mathcal{B}$ on $\mathbb{R}^d$ is said to be *translation invariant* if translations are isometric on $\mathcal{B}$, namely, if for each $f \in \mathcal{B}$ and $x \in \mathbb{R}^d, f(\cdot + x) \in \mathcal{B}$ and $\|f(\cdot + x)\|_{\mathcal{B}} = \|f\|_{\mathcal{B}}$. It was proved in [35]

that a scalar-valued RKHS is translation invariant if and only if its reproducing kernel is of the form $\psi(x-y)$ for some scalar-valued function ψ. For the Banach space case, as a reproducing kernel alone does not determine its RKBS, we do not have such a characterization. Our purpose in this subsection is to construct a class of translation invariant vector-valued RKBSs by the Fourier transform.

Denote by $L^1(\mathbb{R}^d)$ the Banach space of Lebesgue measurable functions f on $\mathbb{R}^d$ equipped with the norm

$$\|f\|_{L^1(\mathbb{R}^d)} := \int_{\mathbb{R}^d} |f(x)|dx.$$

For $\varphi \in L^1(\mathbb{R}^d)$, its Fourier transform $\hat{\varphi}$ and inverse Fourier transform $\check{\varphi}$ are respectively given by

$$\hat{\varphi}(t) := \frac{1}{(\sqrt{2\pi})^d} \int_{\mathbb{R}^d} \varphi(x)e^{-ix\cdot t}dx, \quad t \in \mathbb{R}^d$$

and

$$\check{\varphi}(t) := \frac{1}{(\sqrt{2\pi})^d} \int_{\mathbb{R}^d} \varphi(x)e^{ix\cdot t}dx, \quad t \in \mathbb{R}^d.$$

Here $x \cdot t$ is the standard inner product on $\mathbb{R}^d$.

To start the construction, we let ϕ be a nonnegative function in $L^1(\mathbb{R}^d)$ with $\int_{\mathbb{R}^d} \phi(x)dx = 1$ and denote by $L^p(\mathbb{R}^d, d\phi)$, $p \in (1, +\infty)$, the Banach space of Lebesgue measurable functions f on $\mathbb{R}^d$ with the norm

$$\|f\|_{L^p(\mathbb{R}^d, d\phi)} := \left(\int_{\mathbb{R}^d} |f(x)|^p \phi(x)dx \right)^{1/p} < +\infty.$$

The feature space $\mathcal{W}$ is chosen as

$$\mathcal{W} := \{u = (u_1, \ldots, u_n) : u_j \in L^p(\mathbb{R}^d, d\phi), \, j \in \mathbb{N}_n\}$$

endowed with the norm

$$\|u\|_{\mathcal{W}} := \left(\sum_{j=1}^n \|u_j\|_{L^p(\mathbb{R}^d, d\phi)}^p \right)^{1/p}.$$

Its dual space $\mathcal{W}^*$ is given by

$$\mathcal{W}^* = \{w = (w_1, \ldots, w_n) : w_j \in L^q(\mathbb{R}^d, d\phi), \, j \in \mathbb{N}_n\}$$

with the norm

$$\|w\|_{\mathcal{W}^*} := \left(\sum_{j=1}^n \|w_j\|_{L^q(\mathbb{R}^d, d\phi)}^q \right)^{1/q}.$$

The bilinear form on $\mathcal{W} \times \mathcal{W}^*$ is

$$(u, w)_{\mathcal{W}} = \sum_{j=1}^n \int_{\mathbb{R}^d} u_j(x)w_j(x)\phi(x)dx, \quad u \in \mathcal{W}, \, w \in \mathcal{W}^*.$$

Moreover, the dual element of $u \in \mathcal{W}$ is

$$u^* = \left(\frac{u_j^* \|u_j\|_{L^p(\mathbb{R}^d, d\phi)}^{p-2}}{\|u\|_{\mathcal{W}}^{p-2}} : j \in \mathbb{N}_n \right).$$

By Proposition 4.1, $\mathcal{W}$ is a uniform Banach space. Our feature map $\Phi : \mathbb{R}^d \to \mathcal{L}(\mathcal{W}, \mathbb{C}^n)$ is then defined by

$$\Phi(x)u := S(u\phi)\hat{\ }(x), \quad x \in \mathbb{R}^d, \, u \in \mathcal{W},$$

where S is an invertible $n \times n$ matrix and $(u\phi)\hat{} := ((u_j\phi)\hat{} : j \in \mathbb{N}_n)$. The map Φ is well-defined as $f\phi \in L^1(\mathbb{R}^d)$ for all $f \in L^p(\mathbb{R}^d, d\phi)$ by the Hölder inequality. We also notice that $\Phi(x)$ is continuous from $\mathcal{W}$ to $\mathbb{C}^n$ for each $x \in \mathbb{R}^d$ by the fact that

$$|(f\phi)\hat{}(x)| \leq \|f\phi\|_{L^1(\mathbb{R}^d)} \leq \|f\|_{L^p(\mathbb{R}^d, d\phi)} \quad \text{for all } f \in L^p(\mathbb{R}^d, d\phi).$$

One sees that the adjoint operator $\Phi^* : \mathbb{R}^d \to \mathcal{L}(\mathbb{C}^n, \mathcal{W}^*)$ is given by

$$\Phi^*(x)(\eta) = \frac{e^{-ix \cdot t}}{(\sqrt{2\pi})^d} S^T \eta, \quad x \in \mathbb{R}^d, \ \eta \in \mathbb{C}^n.$$

Clearly, the denseness condition (3.5) is satisfied. The equivalent condition (3.2) hence holds true. We obtain by Corollary 3.2 that

$$\mathcal{B} := \{f_u := S(u\phi)\hat{} : u \in \mathcal{W}\}$$

with the norm $\|f_u\|_{\mathcal{B}} := \|u\|_{\mathcal{W}}$ and compatible semi-inner product

$$[S(u\phi)\hat{}, S(v\phi)\hat{}]_{\mathcal{B}} = [u, v]_{\mathcal{W}} = \sum_{j=1}^{n} \frac{1}{\|v\|_{\mathcal{W}}^{p-2}} \int_{\mathbb{R}^d} u_j(x)\overline{v_j(x)}|v_j(x)|^{p-2}\phi(x)dx$$

is a $\mathbb{C}^n$-valued RKBS. It is translation invariant because for all $y \in \mathbb{R}^d$ and $u \in \mathcal{W}$

$$\|S(u\phi)\hat{}(\cdot + y)\|_{\mathcal{B}} = \|S(e^{-iy \cdot t}u\phi)\hat{}\|_{\mathcal{B}} = \|e^{-y \cdot t}u\|_{\mathcal{W}} = \|u\|_{\mathcal{W}} = \|S(u\phi)\hat{}\|_{\mathcal{B}}.$$

To understand the reproducing kernel of $\mathcal{B}$, we present the dual space of $\mathcal{B}$

$$\mathcal{B}^* = \{S(u^*\phi)\check{} : u \in \mathcal{W}\}$$

with the norm, compatible semi-inner product and bilinear form

$$\|S(u^*\phi)\check{}\|_{\mathcal{B}^*} = \|u^*\|_{\mathcal{W}^*}, \qquad [S(u^*\phi)\check{}, S(v^*\phi)\check{}]_{\mathcal{B}^*} = [v, u]_{\mathcal{W}},$$
$$(S(u\phi)\hat{}, S(v^*\phi)\check{})_{\mathcal{B}} = (u, v^*)_{\mathcal{W}}.$$

With these preparations, we identify by (2.6) that

$$(K(x, \cdot)\xi)^* = S(v_{x,\xi}^*\phi)\check{}, \quad x \in \mathbb{R}^d, \ \xi \in \mathbb{C}^n,$$

where

$$v_{x,\xi}^*(t) := \frac{e^{-ix \cdot t}}{(\sqrt{2\pi})^d} S^T \xi^*, \quad t \in \mathbb{R}^d$$

and ξ^* is the dual element of ξ in $\mathbb{C}^n$ under a strictly convex norm. By the above two equations,

$$(K(x, \cdot)\xi)^*(y) = \frac{1}{(\sqrt{2\pi})^d} SS^T \xi^* \hat{\phi}(x - y), \quad x, y \in \mathbb{R}^d, \ \xi \in \mathbb{C}^n.$$

We also derive that

$$K(x, y)\xi = \frac{\|S^T\xi^*\|_{\ell_q^n}^{\frac{p-2}{p-1}}}{(\sqrt{2\pi})^d} S \left(\frac{\overline{(S^T\xi^*)_j}}{|((S^T\xi^*)_j)|^{\frac{p-2}{p-1}}} : j \in \mathbb{N}_n \right)^T \hat{\phi}(y - x), \quad x, y \in \mathbb{R}^d, \ \xi \in \mathbb{C}^n.$$

We remark that when $p = 2$, $\mathbb{C}^n$ is endowed with the standard Euclidean norm $\| \cdot \|$, and ϕ is the Gaussian function, K becomes the Gaussian kernel for $\mathbb{C}^n$-valued RKHS

$$K(x, y) = SS^* \exp\left(-\frac{\|x - y\|^2}{2}\right), \quad x, y \in \mathbb{R}^d,$$

which confirms the validity of the above construction.

5. Multi-task learning with Banach spaces

We discuss the applications of vector-valued RKBSs to the learning of vector-valued functions from finite samples. Specifically, suppose that the unknown target function is from the input space X to an output space Λ and the observations of the function on given sampling points $\{x_j : j \in \mathbb{N}_m\} \subseteq X$ are available. The observation at $x_j, j \in \mathbb{N}_m$ could be $f(x_j)$ or the application of some continuous linear functional in Λ^* on $f(x_j)$. And it is usually corrupted by noise in practice. To handle the noise and have a good generalization error, we shall follow the regularization methodology. For notational simplicity, let $\mathbf{x} := (x_j : j \in \mathbb{N}_m) \in X^m$ and $f(\mathbf{x}) := (f(x_j) : j \in \mathbb{N}_m) \in \Lambda^m$. A general learning scheme has the following form

$$\inf_{f \in \mathcal{B}} Q(f(\mathbf{x})) + \lambda \Psi(\|f\|_{\mathcal{B}}), \tag{5.1}$$

where $\mathcal{B}$ is a chosen Λ-valued RKBS on $X, Q : \Lambda^m \to \mathbb{R}_+$ is a loss function, λ is a positive regularization parameter, and $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ is called a *regularizer*. We are concerned with the existence and uniqueness, representation, and solving of the minimizer of (5.1). Before moving on to these topics, let us see some examples of learning schemes of the form (5.1):

– Regularization networks

$$Q(f(\mathbf{x})) := \sum_{j=1}^{m} \|f(x_j) - \xi_j\|_{\Lambda}^2, \quad \Psi(\|f\|_{\mathcal{B}}) := \|f\|_{\mathcal{B}}^2, \tag{5.2}$$

where $\xi_j \in \Lambda, j \in \mathbb{N}_m$ are observed outputs of f at $\mathbf{x}$. In general, one may use

$$Q(f(\mathbf{x})) = P(\|f(x_1) - \xi_1\|_{\Lambda}, \ldots, \|f(x_m) - \xi_m\|_{\Lambda}), \tag{5.3}$$

where P is a function from $\mathbb{R}_+^m \to \mathbb{R}_+$. A particular choice of P leads to the support vector machine regression.

– Support vector machine regression

$$\Lambda := \mathbb{R}^n, \qquad Q(f(\mathbf{x})) = \sum_{j=1}^{m} \max(0, \|f(x_j) - \xi_j\|_{\ell_1^n} - \varepsilon),$$

where ε is a positive constant standing for the tolerance level.
– Spectral learning: when $\mathcal{B}$ is the space of sensing matrices introduced in the last section with a unitarily invariant matrix norm, (5.1) is the special spectral learning considered in [2].

5.1. Existence and uniqueness

The weak topology is the weakest topology on a Banach space V such that elements in V^* remain continuous on V. A sequence $u_n \in V, n \in \mathbb{N}$, is said to converge weakly to $u_0 \in V$ if for each $\mu \in V^*, \mu(u_n)$ converges to $\mu(u_0)$. We call a regularizer $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ *admissible* if it is continuous and nondecreasing on $\mathbb{R}_+$ with

$$\lim_{t \to \infty} \Psi(t) = +\infty. \tag{5.4}$$

Proposition 5.1. *If $Q : \Lambda^m \to \mathbb{R}_+$ is continuous with respect to each of its variables under the weak topology on Λ and Ψ is an admissible regularizer then (5.1) has at least a minimizer.*

Proof. Arguments similar to those in the proof of Proposition 4 in [39] still apply to the vector-valued case considered here.   □

When Λ is finite-dimensional, any two topologies on it are equivalent. Thus, continuity under the weak topology is equivalent to continuity with respect to the norm of Λ.

Corollary 5.2. *Let $\mathcal{B}$ be finite-dimensional. If $Q : \Lambda^m \to \mathbb{R}_+$ is continuous with respect to each of its variables and Ψ is an admissible regularizer then (5.1) has at least a minimizer.*

We next deal with the case when the loss function has the form (5.3).

Proposition 5.3. *If $P : \mathbb{R}_+^m \to \mathbb{R}_+$ is continuous on $\mathbb{R}_+^m$ and nondecreasing with respect to each of its variables and the regularizer Ψ is admissible then*

$$\inf_{f \in \mathcal{B}} P(\|f(x_1) - \xi_1\|_\Lambda, \ldots, \|f(x_m) - \xi_m\|_\Lambda) + \lambda \Psi(\|f\|_\mathcal{B}) \tag{5.5}$$

has a minimizer.

Proof. Set

$$\mathcal{E}(f) := P(\|f(x_1) - \xi_1\|_\Lambda, \ldots, \|f(x_m) - \xi_m\|_\Lambda) + \lambda \Psi(\|f\|_\mathcal{B}), \quad f \in \mathcal{B}$$

and $\varepsilon_0 := \inf_{f \in \mathcal{B}} \mathcal{E}(f)$. Using the arguments similar to those in [39], we can find a sequence $f_n \in \mathcal{B}$, $n \in \mathbb{N}$ that is weakly convergent to some $f_0 \in \mathcal{B}$, and some $\alpha > 0$ such that $\|f_0\|_\mathcal{B} \leq \alpha$ and $\|f_n\|_\mathcal{B} \leq \alpha$ for all $n \in \mathbb{N}$. Moreover, for any $\epsilon > 0$ there exists some $N \in \mathbb{N}$ such that for $n > N$,

$$\Psi(\|f_n\|_\mathcal{B}) \geq \Psi(\|f_0\|_\mathcal{B}) - \epsilon. \tag{5.6}$$

Since f_n converges weakly to f_0, by (2.6)

$$\lim_{n \to \infty} [f_n(x_j) - \xi_j, f_0(x_j) - \xi_j]_\Lambda = [f_0(x_j) - \xi_j, f_0(x_j) - \xi_j]_\Lambda \quad \text{for all } j \in \mathbb{N}_m.$$

It implies by the Cauchy–Schwarz inequality of semi-inner products that for any $\delta > 0$ there exists some $N' \in \mathbb{N}$ such that for $n > N'$

$$\|f_n(x_j) - \xi_j\|_\mathcal{B} \geq \|f_0(x_j) - \xi_j\|_\mathcal{B} - \delta \quad \text{for all } j \in \mathbb{N}_m. \tag{5.7}$$

Since

$$\|f_0(x_j) - \xi_j\|_\mathcal{B}, \qquad \|f_n(x_j) - \xi_j\|_\mathcal{B} \leq \max\{\alpha\|\delta_{x_j}\|_{\mathcal{L}(\mathcal{B},\Lambda)} + \|\xi_j\|_\Lambda : j \in \mathbb{N}_m\}$$

and Ψ is uniformly continuous on compact subsets of $\mathbb{R}_+^m$ and is nondecreasing with respect to each of its variables, we get by (5.7) that

$$P(\|f_n(x_1) - \xi_1\|_\Lambda, \ldots, \|f_n(x_m) - \xi_m\|_\Lambda) \geq P(\|f_0(x_1) - \xi_1\|_\Lambda, \ldots, \|f_0(x_m) - \xi_m\|_\Lambda) - \epsilon$$

for sufficiently large n. This combined with (5.6) proves that f_0 is a minimizer of (5.5). $\quad\square$

For uniqueness of the minimizer, we have the following routine result.

Proposition 5.4. *If Q is convex on Λ^m and Ψ is strictly increasing and strictly convex then (5.1) has at most one minimizer.*

Proof. It is straightforward that the function mapping $f \in \mathcal{B}$ to $Q(f(\mathbf{x})) + \lambda \Psi(\|f\|_\mathcal{B})$ is strictly convex on $\mathcal{B}$. $\quad\square$

We close this subsection with the following corollary to the above propositions.

Corollary 5.5. *Let $\mathcal{B}$ be a Λ-valued RKBS on X. Then $\inf_{f \in \mathcal{B}} \mathcal{E}(f)$ has a unique minimizer for the following choices of regularization functionals:*

$$\mathcal{E}(f) = \sum_{j=1}^m \|f(x_j) - \xi_j\|_\Lambda^p + \lambda\|f\|_\mathcal{B}^r, \quad p \in [1, +\infty), \ r \in (1, +\infty),$$

$$\mathcal{E}(f) = \sum_{j=1}^m \max(0, \|f(x_j) - \xi_j\|_\Lambda - \varepsilon) + \lambda\|f\|_\mathcal{B}^r, \quad r \in (1, +\infty), \ \varepsilon > 0.$$

5.2. The representer theorem

We study the representation of the minimizer of (5.1) by the reproducing kernel K of $\mathcal{B}$. The result, known as the representer theorem in the scalar-valued and vector-valued RKHS cases, was due to [21] and [25], respectively. For more references on this subject for the RKHS case, see [1,28]

and the references cited therein. We established the representer theorem for scalar-valued RKBSs in [37,39]. The representer theorem is closely related to the minimal norm interpolation. We start with examining the latter problem.

Let $\mathbf{x} := (x_j : j \in \mathbb{N}_m) \in X^m$ be a fixed set of sampling points. Denote for each $\mathbf{z} := (\eta_j : j \in \mathbb{N}_m) \in \Lambda^m$ by $\mathit{l}_{\mathbf{z}}$ the set of functions $f \in \mathcal{B}$ that satisfy the interpolation condition $f(\mathbf{x}) = \mathbf{z}$. We need two notations for the proof of the representer theorem for the minimal norm interpolation. For a subset A of Banach space V, $A^\perp$ stands for the set of all the continuous linear functionals on V that vanish on A, and for $B \subseteq V^*$, $^\perp B := \{u \in V : \mu(u) = 0 \text{ for all } \mu \in B\}$.

Lemma 5.6. *Let $\mathbf{z} \in \Lambda^m$. If $\mathit{l}_{\mathbf{z}}$ is nonempty then the minimal norm interpolation problem*

$$\inf\{\|f\|_{\mathcal{B}} : f \in \mathit{l}_{\mathbf{z}}\} \tag{5.8}$$

has a unique minimizer. A function $f_0 \in \mathcal{B}$ is the minimizer of (5.8) if and only if $f(\mathbf{x}) = \mathbf{z}$ and

$$f_0^* \in \overline{\text{span}} \left\{ (K(x_j, \cdot)\xi)^* : j \in \mathbb{N}_m, \ \xi \in \Lambda \right\}. \tag{5.9}$$

Proof. Clearly, $\mathit{l}_{\mathbf{z}}$ is a closed convex subset of $\mathcal{B}$. A minimizer of (5.8) is the best approximation in $\mathit{l}_{\mathbf{z}}$ to the origin 0 of $\mathcal{B}$. It is well-known that a closed convex subset in a uniform convex Banach space has a unique best approximation to a point in the same space. By this fact, (5.8) has a unique minimizer. It is also trivial that $f_0 \in \mathit{l}_{\mathbf{z}}$ is the minimizer if and only if

$$\|f_0 + g\|_{\mathcal{B}} \geq \|f_0\|_{\mathcal{B}} \quad \text{for all } g \in \mathit{l}_0.$$

By the characterization of best approximation by the semi-inner product established in [16], the above equation holds if and only if

$$[g, f_0] = 0 \quad \text{for all } g \in \mathit{l}_0,$$

which can be equivalently expressed as $f_0^* \in (\mathit{l}_0)^\perp$. Note that $g \in \mathit{l}_0$ if and only if

$$[g, K(x_j, \cdot)\xi]_{\mathcal{B}} = [g(x_j), \xi]_\Lambda = 0 \quad \text{for all } j \in \mathbb{N}_m \text{ and } \xi \in \Lambda,$$

which is equivalent to that

$$g \in {}^\perp \left\{ (K(x_j, \cdot)\xi)^* : j \in \mathbb{N}_m, \ \xi \in \Lambda \right\}.$$

We conclude that $f_0 \in \mathit{l}_{\mathbf{z}}$ is the minimizer of (5.8) if and only if

$$f_0^* \in \left({}^\perp \left\{ (K(x_j, \cdot)\xi)^* : j \in \mathbb{N}_m, \ \xi \in \Lambda \right\} \right)^\perp.$$

By the Hahn–Banach theorem, for each $B \in \mathcal{B}^*$, $({}^\perp B)^\perp = \overline{\text{span}} \, B$. The proof is hence complete. $\quad \square$

The above lemma enables us to prove the main result of the section without much effort.

Theorem 5.7. *Suppose that (5.1) has at least a minimizer. If the regularizer is nondecreasing then (5.1) has a minimizer that satisfies (5.9). If Ψ is strictly increasing then every minimizer of (5.1) must satisfy (5.9).*

Proof. Let $f \in \mathcal{B}$ be a minimizer of (5.1). We let f_0 be the minimizer of

$$\min\{\|g\|_{\mathcal{B}} : g \in \mathit{l}_{f(\mathbf{x})}\}. \tag{5.10}$$

Then $\|f_0\|_{\mathcal{B}} \leq \|f\|_{\mathcal{B}}$ and $f_0(\mathbf{x}) = f(\mathbf{x})$. It follows that $Q(f_0(\mathbf{x})) = Q(f(\mathbf{x}))$ while $\Psi(\|f_0\|_{\mathcal{B}}) \leq \Psi(\|f\|_{\mathcal{B}})$ as Ψ is nondecreasing. Therefore, f_0 is a minimizer of (5.1). By Lemma 5.6, f_0 satisfies (5.9).

Suppose that Ψ is strictly increasing and $f \in \mathcal{B}$ does not satisfy (5.9). Again, we let $f_0 \in \mathcal{B}$ be the minimizer of (5.10). As f does not satisfy (5.9), $f \neq f_0$ by Lemma 5.6. Thus, $\|f\|_{\mathcal{B}} > \|f_0\|_{\mathcal{B}}$. The consequence is that while $Q(f(\mathbf{x})) = Q(f_0(\mathbf{x}))$, $\Psi(\|f\|_{\mathcal{B}}) > \Psi(\|f_0\|_{\mathcal{B}})$ because Ψ is strictly increasing. Therefore, f cannot be the minimizer of (5.1). The proof is complete. $\quad \square$

5.3. Characterization equations

We consider the solving of the regularized learning scheme (5.1) in this subsection. We try to make use of the representer theorem. To this end, we note that the output space Λ is usually finite-dimensional in practice. Let us assume that (5.1) has a unique minimizer f_0, $\dim(\Lambda) = n < +\infty$, and $\{e_l^* : l \in \mathbb{N}_n\}$ is a basis for $\mathcal{B}^*$. In this case, we see by property (2.16) of the reproducing kernel K that f_0 has the form

$$f_0^* = \sum_{j=1}^m (K(x_j, \cdot)\eta_j)^* \tag{5.11}$$

for some $\eta_j \in \Lambda, j \in \mathbb{N}_m$. It hence suffices to find the finite model parameters η_j's in order to obtain f_0. To this end, one may substitute (5.11) into (5.1) to convert the original minimization problem in a potentially infinite-dimensional Banach space into one about the finitely many parameters η_j's. We next show how the reformulation can be done under the finite-dimensionality assumption on Λ. As each $\xi \in \Lambda$ is uniquely determined by $\{[\xi, e_l]_\Lambda : l \in \mathbb{N}_n\}$. We may rewrite the regularization functional as

$$\min_{f \in \mathcal{B}} \mathcal{R}(([f(\xi_j), e_l]_\Lambda : j \in \mathbb{N}_m,\ l \in \mathbb{N}_n)) + \lambda \Psi(\|f\|_{\mathcal{B}}) \tag{5.12}$$

for some function $\mathcal{R} : \mathbb{C}^{m \times n} \to \mathbb{R}_+$. By (2.5) and (2.6)

$$[f(\xi_j), e_l]_\Lambda = [f, K(x_j, \cdot)e_l]_{\mathcal{B}} = [(K(x_j, \cdot)e_l)^*, f^*]_{\mathcal{B}^*}.$$

For the regularizer part, we have by (2.2) that $\|f\|_{\mathcal{B}} = \|f^*\|_{\mathcal{B}^*}$. Therefore, the parameters η_j's in (5.11) are the minimizer of

$$\min_{\tau \in \Lambda^m} \mathcal{R}\left(\left(\left[(K(x_j, \cdot)e_l)^*, \sum_{k=1}^m (K(x_k, \cdot)\tau_k)^*\right]_{\mathcal{B}^*} : j \in \mathbb{N}_m,\ l \in \mathbb{N}_n\right)\right)$$
$$+ \lambda \Psi\left(\left\|\sum_{j=1}^m (K(x_j, \cdot)\tau_j)^*\right\|_{\mathcal{B}^*}\right).$$

Unlike the RKHS case, the above minimization problem is usually non-convex with respect to τ_j^* or τ_j even when $\mathcal{R}$ and Ψ are both convex. The reason is that a semi-inner product is generally non-additive with respect to its second variable.

In some occasions, one is able to derive a characterization equation for the minimization problem (5.1), which together with the representer theorem constitutes a powerful tool in converting the minimization into a system of equations about the model parameters in the representer theorem. We shall derive characterization equations for the particular example of (5.1)

$$\min_{f \in \mathcal{B}} \sum_{j=1}^m \varphi(\|f(x_j) - \xi_j\|_\Lambda) + \lambda \Psi(\|f\|_{\mathcal{B}}), \tag{5.13}$$

where ξ_j stands for the observation of the target function at x_j for $j \in \mathbb{N}_m$, and φ is a chosen loss function from $\mathbb{R}_+$ to $\mathbb{R}_+$. We shall assume that both φ and Ψ are continuously differentiable and

$$\lim_{t \to 0^+} \frac{\varphi'(t)}{t} = 0. \tag{5.14}$$

For convenience, we make the convention that $0/0 := 0$. The next two results hold for any Λ regardless of its dimension.

Theorem 5.8. *Let Ψ and φ be continuously differentiable on $\mathbb{R}_+$ with (5.14). A function $f_0 \neq 0$ is the minimizer of (5.13) if and only if*

$$\lambda \frac{\Psi'(\|f_0\|_{\mathcal{B}})}{\|f_0\|_{\mathcal{B}}} f_0^* + \sum_{j=1}^m \frac{\varphi'(\|f_0(x_j) - \xi_j\|_{\mathcal{B}})}{\|f_0(x_j) - \xi_j\|_{\mathcal{B}}} (K(x_j, \cdot)(f_0(x_j) - \xi_j))^* = 0. \tag{5.15}$$

The zero function is the minimizer of (5.13) *if and only if*

$$\|T\|_{\mathcal{B}^*} \leq \lambda \Psi'(0), \tag{5.16}$$

where

$$T := \sum_{j=1}^{m} \frac{\varphi'(\|\xi_j\|_{\Lambda})}{\|\xi_j\|_{\Lambda}} (K(x_j, \cdot)\xi_j)^*.$$

Proof. The proof is similar to that for the scalar-valued RKBS case in [39]. One only needs to handle the semi-inner product in vector-valued RKBSs carefully. $\square$

In the sequel, we discuss the application of the above theorem to the regularization networks

$$\min_{f \in \mathcal{B}} \sum_{j=1}^{m} \|f(x_j) - \xi_j\|_{\Lambda}^2 + \lambda \|f\|_{\mathcal{B}}^2. \tag{5.17}$$

To this end, we say that the point evaluations on $\mathcal{B}$ at $x_j, j \in \mathbb{N}_m$ are *essentially linearly independent* if for all $\eta_j \in \Lambda, j \in \mathbb{N}_m$

$$\sum_{j=1}^{m} [f(x_j), \eta_j]_{\Lambda} = 0 \quad \text{for all } f \in \mathcal{B}$$

necessitates that $\eta_j = 0$ for each $j \in \mathbb{N}_m$. By (2.6), $\delta_{x_j}, j \in \mathbb{N}_m$ are essentially linearly independent if and only if

$$\sum_{j=1}^{m} (K(x_j, \cdot)\eta_j)^* = 0$$

implies that $\eta_j = 0$ for each $j \in \mathbb{N}_m$.

Corollary 5.9. *Suppose that the point evaluations on $\mathcal{B}$ at $x_j, j \in \mathbb{N}_m$ are essentially linearly independent. Then f_0 is the minimizer of the regularization network* (5.17) *if and only if it is of the form* (5.11) *where the parameters η_j's satisfy*

$$\lambda \eta_j + f_0(x_j) - \xi_j = 0 \quad \text{for all } j \in \mathbb{N}_m. \tag{5.18}$$

Proof. For the regularization network (5.17), (5.15) and (5.16) are equivalent to each other when $f_0 = 0$. By Theorem 5.8, f_0 is the minimizer of (5.17) if and only if

$$\lambda f_0^* + \sum_{j=1}^{m} \left(K(x_j, \cdot)(f_0(x_j) - \xi_j)\right)^* = 0. \tag{5.19}$$

Thus, f_0 has the form (5.11). Since $\delta_{x_j}, j \in \mathbb{N}_m$ are essentially linearly independent, (5.19) is equivalent to that the parameters η_j's in (5.11) satisfy (5.18). The proof is complete. $\square$

Similarly, one may substitute the representer theorem into the characterization equations (5.15) and (5.18) to reduce the minimization problem to the solving of a system of equations about the parameters η_j's. Again, due to the non-additivity of a semi-inner product with respect to its second variable, the resulting equations are generally nonlinear about the parameters. We conduct the reformulation when Λ is of finite dimension $n \in \mathbb{N}$ and $\{e_l^* : l \in \mathbb{N}_n\}$ forms a basis for Λ^*. In this case, (5.18) can be reformulated as

$$\lambda [\eta_j, e_l]_{\Lambda} + \left[(K(x_j, \cdot)e_l)^*, \sum_{k=1}^{m} (K(x_k, \cdot)\eta_k)^* \right]_{\mathcal{B}^*} = [\xi_j, e_l], \quad j \in \mathbb{N}_m, \, l \in \mathbb{N}_n.$$

We shall leave the solution of the resulting non-convex minimization problem and nonlinear equations about the parameters in the representer theorem for future study.

References

[1] A. Argyriou, C.A. Micchelli, M. Pontil, When is there a representer theorem? Vector versus matrix regularizers, J. Mach. Learn. Res. 10 (2009) 2507–2529.
[2] A. Argyriou, C.A. Micchelli, M. Pontil, On spectral learning, J. Mach. Learn. Res. 11 (2010) 935–953.
[3] N. Aronszajn, Theory of reproducing Kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[4] K.P. Bennett, E.J. Bredensteiner, Duality and geometry in SVM classifiers, in: P. Langley (Ed.), Proceeding of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 2000, pp. 57–64.
[5] J. Burbea, P. Masani, Banach and Hilbert Spaces of Vector-valued Functions, in: Pitman Research Notes in Mathematics, vol. 90, Pitman, Boston, MA, 1984.
[6] S. Canu, X. Mary, A. Rakotomamonjy, Functional learning through Kernel, in: J. Suykens, G. Horvath, S. Basu, C. Micchelli, J. Vandewalle (Eds.), Advances in Learning Theory: Methods, Models and Applications, in: NATO Science Series III: Computer and Systems Sciences, vol. 190, IOS Press, Amsterdam, 2003, pp. 89–110.
[7] A. Caponnetto, C.A. Micchelli, M. Pontil, Y. Ying, Universal multi-task Kernels, J. Mach. Learn. Res. 9 (2008) 1615–1646.
[8] C. Carmeli, E. De Vito, A. Toigo, Vector valued reproducing Kernel Hilbert spaces of integrable functions and Mercer theorem, Anal. Appl. 4 (2006) 377–408.
[9] C. Carmeli, E. De Vito, A. Toigo, V. Umanita, Vector valued reproducing Kernel Hilbert spaces and universality, Anal. Appl. 8 (2010) 19–61.
[10] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2002) 1–49.
[11] D.F. Cudia, On the localization and directionalization of uniform convexity, Bull. Amer. Math. Soc. 69 (1963) 265–267.
[12] R. Der, D. Lee, Large-margin classification in Banach spaces, in: JMLR Workshop and Conference Proceedings, 2: AISTATS, 2007, pp. 91–98.
[13] T. Evgeniou, C.A. Micchelli, M. Pontil, Learning multiple tasks with Kernel methods, J. Mach. Learn. Res. 6 (2005) 615–637.
[14] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[15] C. Gentile, A new approximate maximal margin classification algorithm, J. Mach. Learn. Res. 2 (2001) 213–242.
[16] J.R. Giles, Classes of semi-inner-product spaces, Trans. Amer. Math. Soc. 129 (1967) 436–446.
[17] M. Hein, O. Bousquet, B. Schölkopf, Maximal margin classification for metric spaces, J. Comput. System Sci. 71 (2005) 333–359.
[18] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1991.
[19] P.E.T. Jorgensen, E.P.J. Pearse, Gel'fand triples and boundaries of infinite networks, New York J. Math. 17 (2011) 745–781.
[20] D. Kimber, P.M. Long, On-line learning of smooth functions of a single variable, Theoret. Comput. Sci. 148 (1995) 141–156.
[21] G. Kimeldorf, G. Wahba, Some results on Tchebycheffian spline functions, J. Math. Anal. Appl. 33 (1971) 82–95.
[22] D.O. Koehler, A note on some operator theory in certain semi-inner-product spaces, Proc. Amer. Math. Soc. 30 (1971) 363–366.
[23] G. Lumer, Semi-inner-product spaces, Trans. Amer. Math. Soc. 100 (1961) 29–43.
[24] C.A. Micchelli, M. Pontil, A function representation for learning in Banach spaces, in: Learning Theory, in: Lecture Notes in Computer Science, vol. 3120, Springer, Berlin, 2004, pp. 255–269.
[25] C.A. Micchelli, M. Pontil, On learning vector-valued functions, Neural Comput. 17 (2005) 177–204.
[26] C.A. Micchelli, M. Pontil, Feature space perspectives for learning the Kernel, Mach. Learn. 66 (2007) 297–319.
[27] G.B. Pedrick, Theory of reproducing kernels for Hilbert spaces of vector valued functions, Technical Report 19, University of Kansas, 1957.
[28] B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, in: Proceeding of the Fourteenth Annual Conference on Computational Learning Theory and the Fifth European Conference on Computational Learning Theory, Springer-Verlag, London, 2001, pp. 416–426.
[29] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, Mass, 2002.
[30] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.
[31] G. Song, H. Zhang, Reproducing kernel Banach spaces with the ℓ^1 norm II: error analysis for regularized least square regression, Neural Comput. 23 (2011) 2713–2729.
[32] B. Sriperumbudur, K. Fukumizu, G. Lanckriet, Learning in Hilbert vs. Banach spaces: a measure embedding viewpoint, in: Advances in Neural Information Processing Systems, vol. 24, MIT Press, Cambridge, 2011.
[33] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[34] U. von Luxburg, O. Bousquet, Distance-based classification with Lipschitz functions, J. Mach. Learn. Res. 5 (2004) 669–695.
[35] Y. Xu, H. Zhang, Refinement of reproducing Kernels, J. Mach. Learn. Res. 10 (2009) 107–140.
[36] T. Zhang, On the dual formulation of regularized linear systems with convex risks, Mach. Learn. 46 (2002) 91–129.
[37] H. Zhang, Y. Xu, J. Zhang, Reproducing Kernel Banach spaces for machine learning, J. Mach. Learn. Res. 10 (2009) 2741–2775.
[38] H. Zhang, Y. Xu, Q. Zhang, Refinement of operator-valued reproducing Kernels, J. Mach. Learn. Res. 13 (2012) 91–136.
[39] H. Zhang, J. Zhang, Regularized learning in Banach spaces as an optimization problem: representer theorems, J. Global Optim. (in press).
[40] H. Zhang, J. Zhang, Frames, Riesz bases, and sampling expansions in Banach spaces via semi-inner products, Appl. Comput. Harmon. Anal. 31 (2011) 1–25.
[41] F. Zhdanov, Theory and applications of competitive prediction, Ph.D. Thesis, University of London, 2011.
[42] D. Zhou, B. Xiao, H. Zhou, R. Dai, Global geometry of SVM classifiers, Technical Report 30-5-02, Institute of Automation, Chinese Academy of Sciences, 2002.