Journal of Machine Learning Research 13 (2012) 91-136 Submitted 12/10; Revised 8/11; Published 1/12

Refinement of Operator-valued Reproducing Kernels

Haizhang Zhang[*] ZHHAIZH2@SYSU.EDU.CN
School of Mathematics and Computational Science
Sun Yat-sen University
Guangzhou 510275, P. R. China

Yuesheng Xu[†] YXU06@SYR.EDU
Department of Mathematics
Syracuse University
Syracuse, NY 13244, USA

Qinghui Zhang ZHQINGH@MAIL2.SYSU.EDU.CN
School of Mathematics and Computational Science
Sun Yat-sen University
Guangzhou 510275, P. R. China

Editor: John Shawe-Taylor

Abstract

This paper studies the construction of a *refinement* kernel for a given operator-valued reproducing kernel such that the vector-valued reproducing kernel Hilbert space of the refinement kernel contains that of the given kernel as a subspace. The study is motivated from the need of updating the current operator-valued reproducing kernel in multi-task learning when underfitting or overfitting occurs. Numerical simulations confirm that the established refinement kernel method is able to meet this need. Various characterizations are provided based on feature maps and vector-valued integral representations of operator-valued reproducing kernels. Concrete examples of refining translation invariant and finite Hilbert-Schmidt operator-valued reproducing kernels are provided. Other examples include refinement of Hessian of scalar-valued translation-invariant kernels and transformation kernels. Existence and properties of operator-valued reproducing kernels preserved during the refinement process are also investigated.

Keywords: vector-valued reproducing kernel Hilbert spaces, operator-valued reproducing kernels, refinement, embedding, translation invariant kernels, Hessian of Gaussian kernels, Hilbert-Schmidt kernels, numerical experiments

1. Introduction

Machine learning designs algorithms for the purpose of inferring from finite empirical data a function dependency which can then be used to understand or predict generation of new data. Past research has mainly focused on single task learning problems where the function to be learned is scalar-valued. Built upon the theory of scalar-valued reproducing kernels (Aronszajn, 1950), kernel methods have proven useful in single task learning (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1998). The approach might be justified in three ways. Firstly, as inputs for

∗. Also in the Guangdong Province Key Laboratory of Computational Science.

†. Also at Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou 510275, P. R. China.

learning algorithms are sample data, requiring the sampling process to be stable seems inevitable. Thanks to the existence of an inner product, Hilbert spaces are the class of normed vector spaces that we can handle best. These two considerations lead immediately to the notion of reproducing kernel Hilbert spaces (RKHS). Secondly, a reasonable learning scheme is expected to make use of the similarity between a new input and the existing inputs for prediction. Inner products provide a natural measurement of similarities. It is well-known that a bivariate function is a scalar-valued reproducing kernel if and only if it is representable as some inner product of the feature of inputs (Schölkopf and Smola, 2002). Finally, finding a feature map and taking the inner product of the feature of two inputs are equivalent to choosing a scalar-valued reproducing kernel and performing function evaluations of it. This brings computational efficiency and gives birth to the important "kernel trick" (Schölkopf and Smola, 2002) in machine learning. For references on single task learning and scalar-valued RKHS, we recommend Aronszajn (1950), Cucker and Smale (2002), Cucker and Zhou (2007), Evgeniou et al. (2000), Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004) and Vapnik (1998); Zhang et al. (2009).

In this paper, we are concerned with multi-task learning where the function to be reconstructed from finite sample data takes range in a finite-dimensional Euclidean space, or more generally, a Hilbert space. Motivated by the success of kernel methods in single task learning, it was proposed in Evgeniou et al. (2005) and Micchelli and Pontil (2005) to develop algorithms for multi-task learning in the framework of vector-valued RKHS. We attempt to contribute to the theory of vector-valued RKHS by studying a special embedding relationship between two vector-valued RKHS. We shall briefly review existing work on vector-valued RKHS and the associated operator-valued reproducing kernels. The study of vector-valued RKHS dates back to Pedrick (1957). The notion of matrix-valued or operator-valued reproducing kernels was also obtained in Burbea and Masani (1984). References Mukherjee and Wu (2006), Mukherjee and Zhou (2006) and Ying and Campbell (2008) were devoted to learning a multi-variate function and its gradient simultaneously. Reference Carmeli et al. (2006) established the Mercer theorem for vector-valued RKHS and characterized those spaces with elements being p-integrable vector-valued functions. Various characterizations and examples of universal operator-valued reproducing kernels were provided in Caponnetto et al. (2008) and Carmeli et al. (2010). The latter (Carmeli et al., 2010) also examined basic operations of operator-valued reproducing kernels and extended the Bochner characterization of translation invariant reproducing kernels to the operator-valued case.

The purpose of this paper is to study the refinement relationship of two vector-valued reproducing kernels. We say that a vector-valued reproducing kernel is a refinement of another kernel of such type if the RKHS of the first kernel contains that of the latter one as a linear subspace and their norms coincide on the smaller space. The precise definition will be given in the next section after we provide necessary preliminaries on vector-valued RKHS. The study is motivated by the need of updating a vector-valued reproducing kernel for multi-task machine learning when underfitting or overfitting occurs. Detailed explanations of this motivation will be presented in the next section. Mathematically, a thorough understanding of the refinement relationship is essential to the establishment of a multi-scale decomposition of vector-valued RKHS, which in turn is the foundation for extending multi-scale analysis (Daubechies, 1992; Mallat, 1989) to kernel methods. In fact, a special refinement method by a bijective mapping from the input space to itself provides such a decomposition. As the procedure is similar to the scalar-valued case, we refer interested authors to Xu and Zhang (2007) for the details. The notion of refinement of scalar-valued kernels was initiated and extensively investigated by the first two authors (Xu and Zhang, 2007, 2009). Therefore, a gen-

eral principle we shall follow is to briefly mention or even completely omit arguments that are not essentially different from the scalar-valued case. As we proceed with the study, it will become clear that nontrivial obstacles in extending the scalar-valued theory to vector-valued RKHS are mainly caused by the complexity in the vector-valued integral representation of the operator-valued reproducing kernels under investigation, by the complicated form of the feature map involved, which is also operator-valued, and by the infinite-dimensionality of the output space in some occasions.

To be more specific, we would personally regard the following results to be mathematically nontrivial: Theorem 11 of characterizing the refinement of kernels defined by the integral of scalar-valued kernels with respect to an operator-valued measure, Proposition 10 of studying the refinement of positive operators, Lemma 13 of proving the disjointness of the RKHS of translation-invariant kernels of different types, and Theorem 21 about the refinement of finite Hilbert-Schmidt kernels. Besides, compared to the scalar-valued case in Xu and Zhang (2009), Sections 5.2 and 5.3 about the refinement of Hessian kernels and transformation kernels are unique, and Section 7 of numerical experiments is novel. By contrast, the discussion of general characterizations and finite-dimensional RKHS in Section 3, refinement of kernels defined by the integral of operator-value kernels with respect to a scalar-valued measure in Section 4.1, and Section 6 about the existence of refinement and properties preserved by the refinement process can be viewed as either trivial extensions or not of sufficient mathematical depth. We also remark that every vector-valued RKHS is isometrically isomorphic to a scalar-valued RKHS on an extended input space (see Proposition 6 below). However, this does not mean that the question of studying refinement of operator-valued kernels can be trivially reduced to that about scalar-valued kernels. The isometry procedure will usually make the resulting scalar-valued kernel and extended input space complex and difficult to analyze. Moreover, favorable properties such as translation invariance and Hilbert-Schmidt structure of the original kernels are generally lost in the process. Therefore, an independent study of the refinement of operator-valued kernels is necessary and challenging.

This paper is organized as follows. We shall introduce necessary preliminaries on vector-valued RKHS and motivate our study from multi-tasking learning in the next section. In Section 3, we shall present three general characterizations of the refinement relationship by examining the difference of two given kernels, the feature map representation of kernels, and the associated kernels on the extended input space. Recall that most scalar-valued reproducing kernels are represented by integrals. In the operator-valued case, we have two types of integral representations: the integral of operator-valued reproducing kernels with respect to a scalar-valued measure, and the integral of scalar-valued reproducing kernels with respect to an operator-valued measure. As a key part of this paper, we shall investigate in Section 4 specifications of the general characterizations when the operator-valued reproducing kernels are given by such integrals. In Section 5, we present concrete examples of refinement by looking into translation-invariant operator-valued kernels, Hessian of a scalar-valued kernels, Hilbert-Schmidt kernels, etc. Section 6 treats specially the existence of nontrivial refinements and desirable properties of operator-valued reproducing kernels that can be preserved during the refinement process. In Section 7, we perform three numerical simulations to show the effect of the refinement kernel method in updating operator-valued reproducing kernels for multi-task learning. Finally, we conclude the paper in Section 8.

2. Kernel Refinement

To explain our motivation from multi-task learning in details, we first recall the definition of operator-valued reproducing kernels. Throughout the paper, we let X and Λ denote a prescribed set and a separable Hilbert space, respectively. We shall call X the input space and Λ the output space. To avoid confusion, elements in X and Λ will be denoted by x, y, and ξ, η, respectively. Unless specifically mentioned, all the normed vector spaces in the paper are over the field $\mathbb{C}$ of complex numbers. Let $\mathcal{L}(\Lambda)$ be the set of all the bounded linear operators from Λ to Λ, and $\mathcal{L}_+(\Lambda)$ its subset of those linear operators A that are self-adjoint and positive, namely,

$$(A\xi, \xi)_\Lambda \geq 0 \text{ for all } \xi \in \Lambda,$$

where $(\cdot,\cdot)_\Lambda$ is the inner product on Λ. The adjoint of $A \in \mathcal{L}(\Lambda)$ is denoted by A^*. An $\mathcal{L}(\Lambda)$-*valued reproducing kernel* on X is a function $K : X \times X \to \mathcal{L}(\Lambda)$ such that $K(x,y) = K(y,x)^*$ for all $x, y \in X$, and such that for all $x_j \in X$, $\xi_j \in \Lambda$, $j \in \mathbb{N}_n := \{1, 2, \ldots, n\}$, $n \in \mathbb{N}$,

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

For each $\mathcal{L}(\Lambda)$-valued reproducing kernel K on X, there exists a unique Hilbert space, denoted by $\mathcal{H}_K$, consisting of Λ-valued functions on X such that

$$K(x,\cdot)\xi \in \mathcal{H}_K \text{ for all } x \in X \text{ and } \xi \in \Lambda \tag{2}$$

and

$$(f(x), \xi)_\Lambda = (f, K(x,\cdot)\xi)_{\mathcal{H}_K} \text{ for all } f \in \mathcal{H}_K,\ x \in X, \text{ and } \xi \in \Lambda. \tag{3}$$

It is implied by the above two properties that the point evaluation at each $x \in X$:

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

is continuous from $\mathcal{H}_K$ to Λ. In other words, $\mathcal{H}_K$ is a Λ-valued RKHS. We call it the RKHS of K. Conversely, for each Λ-valued RKHS on X, there exists a unique $\mathcal{L}(\Lambda)$-valued reproducing kernel K on X that satisfies (2) and (3). For this reason, we also call K the reproducing kernel (or kernel for short) of $\mathcal{H}_K$. The bijective correspondence between $\mathcal{L}(\Lambda)$-valued reproducing kernels and Λ-valued RKHS is central to the theory of vector-valued RKHS.

Given two $\mathcal{L}(\Lambda)$-valued reproducing kernels K, G on X, we shall investigate in this paper the fundamental embedding relationship $\mathcal{H}_K \preceq \mathcal{H}_G$ in the sense that $\mathcal{H}_K \subseteq \mathcal{H}_G$ and for all $f \in \mathcal{H}_K$, $\|f\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_G}$. Here, $\|\cdot\|_{\mathcal{W}}$ denotes the norm of a normed vector space $\mathcal{W}$. We call G a *refinement* of K if there does hold $\mathcal{H}_K \preceq \mathcal{H}_G$. Such a refinement is said to be nontrivial if $G \neq K$.

We motivate this study from the kernel methods for multi-task learning and from the multi-scale decomposition of vector-valued RKHS. Let $\mathbf{z} := \{(x_j, \xi_j) : j \in \mathbb{N}_n\} \subseteq X \times \Lambda$ be given sample data. A typical kernel method infers from $\mathbf{z}$ the minimizer $f_{\mathbf{z}}$ of

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{j=1}^{n} C(x_j, \xi_j, f(x_j)) + \sigma\phi(\|f\|_{\mathcal{H}_K}), \tag{4}$$

where K is a selected $\mathcal{L}(\Lambda)$-valued reproducing kernel on X, C a prescribed loss function, σ a positive regularization parameter, and ϕ a regularizer. The ideal predictor $f_0 : X \to \Lambda$ that we are pursuing is the one that minimizes

$$\mathcal{E}(f) := \int_{X\times\Lambda} C(x,\xi,f(\xi))dP$$

among all possible functions f from X to Λ. Here P is an unknown probability measure on $X \times \Lambda$ that dominates the generation of data from $X \times \Lambda$. We wish that $\mathcal{E}(f_z) - \mathcal{E}(f_0)$ can converge to zero in probability as the number n of sampling points tends to infinity. Whether this will happen depends heavily on the choice of the kernel K. The error $\mathcal{E}(f_z) - \mathcal{E}(f_0)$ can be decomposed into the sum of the *approximation error* and *sampling error* (Schölkopf and Smola, 2002; Vapnik, 1998). The approximation error occurs as we search the minimizer in a restricted set of candidate functions, namely, $\mathcal{H}_K$. It becomes smaller as $\mathcal{H}_K$ enlarges. The sampling error is caused by replacing the expectation $\mathcal{E}(f)$ of the loss function $C(x,\xi,f(\xi))$ with the sample mean

$$\frac{1}{n}\sum_{j=1}^{n} C(x_j,\xi_j,f(x_j)).$$

By the law of large numbers, the sample mean converges to the expectation in probability as $n \to \infty$ for a fixed $f \in \mathcal{H}_K$. However, as $f_{\mathbf{z}}$ varies according to changes in the sample data $\mathbf{z}$, we need a uniform version of the law of large number on $\mathcal{H}_K$ in order to well control the sampling error. Therefore, the sampling error usually increases as $\mathcal{H}_K$ enlarges, or to be more precisely, as the *capacity* of $\mathcal{H}_K$ increases.

By the above analysis, we might encounter two situations after the choice of an $\mathcal{L}(\Lambda)$-valued reproducing kernel K:

— overfitting, which occurs when the capacity of $\mathcal{H}_K$ is too large, forcing the minimizer obtained from (4) to imitate artificial function dependency in the sample data, and thus causing the sampling error to be out of control;

— underfitting, which occurs when $\mathcal{H}_K$ is too small for the minimizer of (4) to describe the desired function dependency implied in the data, and thus failing in bounding the approximation error.

When one of the above situations happens, a remedy is to modify the reproducing kernel. Specifically, one might want to find another $\mathcal{L}(\Lambda)$-valued reproducing kernel G such that $\mathcal{H}_K \preceq \mathcal{H}_G$ when there is underfitting, or such that $\mathcal{H}_G \preceq \mathcal{H}_K$ when there is overfitting. We see that in either case, we need to make use of the refinement relationship. We shall verify in the last section through extensive numerical simulations that the refinement kernel method is indeed able to provide an appropriate update of an operator-valued reproducing kernel when underfitting or overfitting occurs.

Before moving on to the characterization of refinement of operator-valued reproducing kernels, we collect here notations that will be frequently used in the rest of the paper. They will also be (or have been) defined when first used.

– X: a general input space,

– Λ: a Hilbert space, serving as the output space,

– $\|\cdot\|_\Lambda$: the norm on a Hilbert or Banach space Λ,

– $\mathcal{W}$: a Hilbert space, usually serving as the feature space of reproducing kernels,

– $\mathcal{L}(\Lambda)$: the space of bounded linear operators from Λ to Λ,

– $\mathcal{L}_+(\Lambda)$: the set of self-adjoint and positive bounded linear operators from Λ to Λ,

– $\mathcal{L}(\Lambda,\mathcal{W})$: the space of bounded linear operators from Λ to $\mathcal{W}$,

– K,G: $\mathcal{L}(\Lambda)$-valued reproducing kernels,

– $\mathcal{H}_K,\mathcal{H}_G$: the RKHS of kernels K,G, respectively,

– $\mathcal{H}_K \preceq \mathcal{H}_G$: G is a refinement of K, namely, $\mathcal{H}_K \subseteq \mathcal{H}_G$ and $\|f\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_G}$ for all $f \in \mathcal{H}_K$,

– $\tilde{X}$: the extended input space $X \times \Lambda$,

– $\tilde{K}$: the scalar-valued kernel (11) associated with an $\mathcal{L}(\Lambda)$-valued kernel K,

– μ,ν: scalar-valued or operator-valued measures,

– $|\mu|$: the variation (19) of a measure μ,

– $(\Omega,\mathcal{F},\mu)$: a measure space,

– $\mu \preceq \nu$: means that μ is the restriction of ν on some measurable set,

– $L^2(\Omega,\mathcal{B},d\mu)$: the Hilbert space (16) of square integrable $\mathcal{B}$-valued functions on Ω with respect to the measure μ,

– $L^2(\Omega,d\mu)$: the Hilbert space of scalar-valued square integrable functions on Ω with respect to the measure μ,

– $L^\infty(\Omega,d\mu)$: the Banach space of essentially bounded measurable functions on Ω with respect to the measure μ,

– $A \preceq B$: see (29) for this refinement relation of two positive operators,

– $\mathcal{B}(\mathbb{R}^d,\Lambda)$: the set of all the $\mathcal{L}_+(\Lambda)$-valued measures of bounded variation on the σ-algebra of Borel subsets in $\mathbb{R}^d$,

– γ_c,γ_s: the continuous part γ_c and singular part γ_s in the Lebesgue decomposition (38) of a Borel measure γ,

– L_c,L_s: the continuous and singular parts (39) of a translation-invariant kernel,

– $\Lambda \otimes \mathcal{W}$: the tensor product of two Hilbert spaces Λ and $\mathcal{W}$,

– $\sqrt{A}$: the square root of a positive bounded linear operator A.

3. General Characterizations

The relationship between the RKHS of the sum of two operator-valued reproducing kernels and those of the summand kernels has been made clear in Theorem 1 on page 44 of Pedrick (1957). Our first characterization of refinement is a direct consequence of this result.

Proposition 1 *Let K, G be two $\mathcal{L}(\Lambda)$-valued reproducing kernels on X. Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $G-K$ is an $\mathcal{L}(\Lambda)$-valued reproducing kernel on X and $\mathcal{H}_K \cap \mathcal{H}_{G-K} = \{0\}$. If $\mathcal{H}_K \preceq \mathcal{H}_G$ then $\mathcal{H}_{G-K}$ is the orthogonal complement of $\mathcal{H}_K$ in $\mathcal{H}_G$.*

Every reproducing kernel has a feature map representation. Specifically, K is an $\mathcal{L}(\Lambda)$-valued reproducing kernel on X if and only if there exists a Hilbert space $\mathcal{W}$ and a mapping $\Phi : X \to \mathcal{L}(\Lambda, \mathcal{W})$ such that

$$K(x,y) = \Phi(y)^*\Phi(x), \quad x,y \in X, \tag{5}$$

where $\mathcal{L}(\Lambda, \mathcal{W})$ denotes the set of bounded linear operators from Λ to $\mathcal{W}$, and $\Phi(y)^*$ is the adjoint operator of $\Phi(y)$. We call Φ a *feature map* of K. The following lemma is useful in identifying the RKHS of a reproducing kernel given by a feature map representation (5).

Lemma 2 *If K is an $\mathcal{L}(\Lambda)$-valued reproducing kernel on X given by (5) then*

$$\mathcal{H}_K = \{\Phi(\cdot)^* u : u \in \mathcal{W}\}$$

with inner product

$$(\Phi(\cdot)^* u, \Phi(\cdot)^* v)_{\mathcal{H}_K} := (P_\Phi u, P_\Phi v)_{\mathcal{W}}, \quad u, v \in \mathcal{W},$$

where P_Φ is the orthogonal projection of $\mathcal{W}$ onto

$$\mathcal{W}_\Phi := \overline{span}\{\Phi(x)\xi : x \in X, \xi \in \Lambda\}.$$

The second characterization can be proved using Lemma 2 and the same arguments with those for the scalar-valued case (Xu and Zhang, 2007).

Theorem 3 *Suppose that $\mathcal{L}(\Lambda)$-valued reproducing kernels K and G are given by the feature maps $\Phi : X \to \mathcal{L}(\Lambda, \mathcal{W})$ and $\Phi' : X \to \mathcal{L}(\Lambda, \mathcal{W}')$, respectively. Assume that $\mathcal{W}_\Phi = \mathcal{W}$ and $\mathcal{W}'_{\Phi'} = \mathcal{W}'$. Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if there exists a bounded linear operator T from $\mathcal{W}'$ to $\mathcal{W}$ such that*

$$T\Phi'(x) = \Phi(x) \text{ for all } x \in X,$$

and the adjoint operator $T^* : \mathcal{W} \to \mathcal{W}'$ is isometric. In this case, G is a nontrivial refinement of K if and only if T is not injective.

To illustrate the above useful results, we shall present a concrete example aiming at refining $\mathcal{L}(\Lambda)$-valued reproducing kernels K with a finite-dimensional RKHS. A simple observation is made regarding such a kernel.

Proposition 4 *A Λ-valued RKHS $\mathcal{H}_K$ is of finite dimension $n \in \mathbb{N}$ if and only if there exists an $n \times n$ hermitian and strictly positive-definite matrix A and n linearly independent functions $\phi_j : X \to \Lambda$, $j \in \mathbb{N}_n$ such that*

$$K(x,y)\xi = \sum_{j=1}^{n}\sum_{k=1}^{n} A_{jk}(\xi, \phi_j(x))_\Lambda \phi_k(y), \quad x,y \in X,\ \xi \in \Lambda. \tag{6}$$

Proof Assume that $\mathcal{H}_K$ is n dimensional with orthogonal basis $\{\phi_j : j \in \mathbb{N}_n\}$. As $K(x,\cdot)\xi \in \mathcal{H}_K$ for all $x \in X$, $\xi \in \Lambda$, there exist functions $c_j : X \times \Lambda \to \mathbb{C}$ such that

$$K(x,y)\xi = \sum_{j=1}^{n} c_j(\xi,x)\phi_j(y), \;\; x,y \in X,\; \xi \in \Lambda.$$

Since $\{\phi_j : j \in \mathbb{N}_n\}$ is a basis for $\mathcal{H}_K$, each function $f \in \mathcal{H}_K$ has the form

$$f = \sum_{j=1}^{n} d_j\phi_j,\; d_j \in \mathbb{C},\; j \in \mathbb{N}_n.$$

Clearly, $\|f\| := (\sum_{j=1}^{n} |d_j|^2)^{1/2}$ is a norm on $\mathcal{H}_K$. It is equivalent to the original one on $\mathcal{H}_K$ as $\dim \mathcal{H}_K < \infty$. It is implied that there exists some $C > 0$ such that

$$\sum_{j=1}^{n} |c_j(\xi,x)|^2 \le C\|K(x,\cdot)\xi\|^2_{\mathcal{H}_K} = C(K(x,x)\xi,\xi)_\Lambda \le C\|\xi\|^2_\Lambda \|K(x,x)\|. \tag{7}$$

Obviously, for each $x \in X$ and $j \in \mathbb{N}_n$, $c_j(\cdot,x)$ is a linear functional on Λ. This together with (7) implies that $c_j(\cdot,x)$ are bounded linear functionals on Λ. By the Riesz representation theorem, there exists $\psi_j : X \to \Lambda$, $j \in \mathbb{N}_n$ such that

$$c_j(\xi,x) = (\xi,\psi_j(x))_\Lambda.$$

We conclude that K has the form

$$K(x,y)\xi = \sum_{j=1}^{n} (\xi,\psi_j(x))_\Lambda \phi_j(y), \;\; x,y \in X,\; \xi \in \Lambda. \tag{8}$$

Since $\{\phi_j : j \in \mathbb{N}_n\}$ is an orthogonal basis for $\mathcal{H}_K$, by (3),

$$(\xi,\psi_j(x))_\Lambda = (K(x,\cdot)\xi,\phi_j)_{\mathcal{H}_K} = (\xi,\phi_j(x))_\Lambda,\; \xi \in \Lambda, x \in X.$$

It follows that $\psi_j = \phi_j,\; j \in \mathbb{N}_n$. Substituting this into (8) yields that

$$K(x,y)\xi = \sum_{j=1}^{n} (\xi,\phi_j(x))_\Lambda \phi_j(y), \;\; x,y \in X,\; \xi \in \Lambda,$$

which indeed is a special form of (6).

Conversely, assume that K has the form (6). We set $\mathcal{W}_A := I^2_A(\mathbb{N}_n) := \{c = (c_j : j \in \mathbb{N}_n) \in \mathbb{C}^n\}$ with inner product

$$(c,d)_{I^2_A(\mathbb{N}_n)} := \sum_{j=1}^{n}\sum_{k=1}^{n} c_j\bar{d}_k A_{jk}.$$

Introduce $\Phi : X \to \mathcal{L}(\Lambda,\mathcal{W}_A)$ by setting $\Phi(x)\xi := ((\xi,\phi_j(x))_\Lambda : j \in \mathbb{N}_n)$. Direct computations show that

$$\Phi^*(x)u = \sum_{j=1}^{n}\sum_{k=1}^{n} \phi_j(x)u_k A_{jk},\; u = (u_j : j \in \mathbb{N}_n) \in \mathcal{W}_A.$$

Thus, we see that $K(x,y) = \Phi(y)^*\Phi(x)$, $x, y \in X$, implying that K is an $\mathcal{L}(\Lambda)$-valued reproducing kernel. By the linear independence of ϕ_j, $j \in \mathbb{N}_n$, $\text{span}\{\Phi(x)\xi : x \in X, \xi \in \Lambda\} = \mathcal{W}_A$. We hence apply Lemma 2 to get that

$$\mathcal{H}_K = \{\Phi(\cdot)^* u : u \in \mathcal{W}_A'\} = \text{span}\{\phi_j : j \in \mathbb{N}_n\},$$

which is of dimension n. ■

By the above proposition, we let ϕ_j, $j \in \mathbb{N}_m$ be linearly independent functions from X to Λ, where $m \geq n$ are fixed positive integers. Let A and B be $n \times n$ and $m \times m$ hermitian and strictly positive-definite matrices, respectively. We define K by (6) in terms of matrix A and G by

$$G(x,y)\xi := \sum_{j=1}^{m} \sum_{k=1}^{m} B_{jk}(\xi, \phi_j(x))_\Lambda \phi_k(y),\ x, y \in X \tag{9}$$

and shall investigate conditions for G to be a refinement of K.

Proposition 5 *Let K, G be defined by (6) and (9), respectively. Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if B^{-1} is an augmentation of A^{-1}, namely, $B_{jk}^{-1} = A_{jk}^{-1}$, $j, k \in \mathbb{N}_n$. In particular, if K, G have the form*

$$K(x,y)\xi = \sum_{j \in \mathbb{N}_n} a_j(\xi, \phi_j(x))_\Lambda \phi_j(y),\ G(x,y)\xi = \sum_{k \in \mathbb{N}_m} b_k(\xi, \phi_k(x))_\Lambda \phi_k(y)$$

for some positive constants a_j, b_k, then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $a_j = b_j$, $j \in \mathbb{N}_n$. In both cases if $\mathcal{H}_K \preceq \mathcal{H}_G$ then G is a nontrivial refinement of K if and only if $m > n$.

Proof It suffices to prove the first claim. We observe that K, G have the feature spaces $\mathcal{W} = I_A^2(\mathbb{N}_n)$ and $\mathcal{W}' = I_B^2(\mathbb{N}_m)$, respectively, with feature maps

$$\Phi(x)\xi := ((\xi, \phi_j(x))_\Lambda : j \in \mathbb{N}_n),\ \Phi'(x)\xi := ((\xi, \phi_k(x))_\Lambda : k \in \mathbb{N}_m),\ x \in X, \xi \in \Lambda.$$

Suppose that $\mathcal{H}_K \preceq \mathcal{H}_G$, then by Theorem 3, there exists a bounded linear operator $T : \mathcal{W}' \to \mathcal{W}$ with properties as described there. It can be represented by an $n \times m$ matrix D as

$$(T\Phi'(x)\xi)_j = \sum_{k=1}^{m} D_{jk}(\xi, \phi_k(x))_\Lambda = (\xi, \phi_j(x))_\Lambda,\ \ x \in X, \xi \in \Lambda, \tag{10}$$

which implies that $D = [I_n, 0]$, where I_n denotes the $n \times n$ identity matrix. The adjoint operator T^* of T is then represented by

$$T^* u = B^{-1} \begin{bmatrix} A \\ 0 \end{bmatrix} u,\ u \in \mathbb{C}^n.$$

Since T^* is isometric, we get that

$$(T^*u, T^*v)_{\mathcal{W}'} = (u, v)_{\mathcal{W}},$$

which has the form

$$v^*[A, 0]B^{-1}BB^{-1} \begin{bmatrix} A \\ 0 \end{bmatrix} u = v^* A u,\ u, v \in \mathbb{C}^n.$$

We derive from the above equation that

$$[A,0]B^{-1}\left[\begin{array}{c} A \\ 0 \end{array}\right] = A.$$

Therefore, B^{-1} is an augmentation of A^{-1}. Conversely, if this is true then $T : \mathcal{W}' \to \mathcal{W}$ defined by

$$Tu' := [I_n, 0]u',\ u' \in \mathbb{C}^m$$

satisfies the two properties in Theorem 3. As a result, $\mathcal{H}_K \preceq \mathcal{H}_G$. ■

It is worthwhile to point out that the above characterization is independent of the Hilbert space Λ.

Unlike the previous two characterizations, the third one comes as a surprise, telling us that theoretically we are able to reduce our consideration to the scalar-valued case.

Introduce for each $\mathcal{L}(\Lambda)$-valued reproducing kernel K on X a scalar-valued reproducing kernel $\tilde{K}$ on the *extended input space* $\tilde{X} := X \times \Lambda$ by setting

$$\tilde{K}((x,\xi),(y,\eta)) := (K(x,y)\xi,\eta)_\Lambda,\ \ x,y \in X,\ \xi,\eta \in \Lambda. \tag{11}$$

By (1), $\tilde{K}$ is indeed positive-definite.

Proposition 6 *There holds $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$. Furthermore, G is a nontrivial refinement of K on X if and only if $\tilde{G}$ is a nontrivial refinement of $\tilde{K}$ on $\tilde{X}$.*

Proof We first explore the close relationship between $\mathcal{H}_K$ and $\mathcal{H}_{\tilde{K}}$. By (3),

$$\tilde{K}((x,\xi),(y,\eta)) = (K(x,y)\xi,\eta)_\Lambda = (K(x,\cdot)\xi, K(y,\cdot)\eta)_{\mathcal{H}_K},$$

which provides a natural feature map $\Phi : \tilde{X} \to \mathcal{H}_K$ of $\tilde{K}$

$$\Phi((x,\xi)) := K(x,\cdot)\xi,\ \ x \in X,\ \xi \in \Lambda.$$

The density condition $\mathcal{W}_\Phi = \mathcal{H}_K$ is clearly satisfied by (3). We hence obtain by (2) that every function $\tilde{f}$ in $\mathcal{H}_{\tilde{K}}$ is of the form

$$\tilde{f}(x,\xi) := (f(x),\xi)_\Lambda \text{ for some } f \in \mathcal{H}_K$$

with

$$\|\tilde{f}\|_{\mathcal{H}_{\tilde{K}}} = \|f\|_{\mathcal{H}_K}.$$

Similar observations can be made about $\mathcal{H}_{\tilde{G}}$.

It follows immediately that $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$ if $\mathcal{H}_K \preceq \mathcal{H}_G$. On the other hand, suppose that $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$. Then for each $f \in \mathcal{H}_K$ there exists some $g \in \mathcal{H}_G$ such that

$$(f(x),\xi)_\Lambda = \tilde{f}(x,\xi) = \tilde{g}(x,\xi) = (g(x),\xi)_\Lambda \text{ for all } x \in X,\ \xi \in \Lambda \tag{12}$$

and

$$\|f\|_{\mathcal{H}_K} = \|\tilde{f}\|_{\mathcal{H}_{\tilde{K}}} = \|\tilde{g}\|_{\mathcal{H}_{\tilde{G}}} = \|g\|_{\mathcal{H}_G}.$$

Equation (12) implies that $f = g$. Therefore, $\mathcal{H}_K \preceq \mathcal{H}_G$. ■

It appears by Proposition 6 that we do not have to bother studying refinement of operator-valued reproducing kernels. Although the strategy sometimes does simplify the problem, the difficulty is generally not reduced significantly. Instead, the result might be viewed as transferring the complexity to the input space. Moreover, desirable properties such as translation invariance of the original kernels might be lost in the process. As a result, an independent study of the operator-valued case remains necessary and challenging.

4. Integral Representations

We shall characterize in this section the refinement of operator-valued kernels defined by two kinds of integral representations: the integral of operator-valued kernels with respect to a scalar-valued measure, and the integral of scalar-valued kernels with respect to an operator-valued measure. The characterizations to be established are crucial to the study of this paper as many useful operator-valued kernels are of an integral representation. Typical examples include the important translation-invariant operator-valued kernels and hessian kernels to be considered in the next section. We also point out in advance the difference in the refinement for the two kinds of integral representations. Firstly, the first refinement corresponds to the same feature map and different measures, while the other when the Radon-Nikodym property is engaged has different feature maps and the same measure. The arguments of the proofs and the obtained results are essential different. The characterization of the first kind of refinement can be viewed as a straightforward generalization of that obtained in Xu and Zhang (2009), while the other one is mathematically nontrivial.

This section will be built on the theory of vector-valued measures and integrals (Berberian, 1966; Diestel and Uhl, 1977). Necessary preliminaries on the subjects will be explained in sufficient details.

4.1 Operator-valued Kernels With Respect to Scalar-valued Measures

Let us first introduce integration of a vector-valued function with respect to a scalar-valued measure. Let $\mathcal{F}$ be a σ-algebra of subsets of a fixed set Ω, μ a finite nonnegative measure on $\mathcal{F}$, and $\mathcal{B}$ a Banach space. We are concerned with $\mathcal{B}$-valued functions on Ω. A function $f : \Omega \to \mathcal{B}$ is said to be *simple* if

$$f = \sum_{j=1}^{n} a_j \chi_{E_j} \tag{13}$$

for some finitely many $a_j \in \mathcal{B}$ and pairwise disjoint subsets $E_j \in \mathcal{F}$, $j \in \mathbb{N}_n$. A function $f : \Omega \to \mathcal{B}$ is called *μ-measurable* if there exists a sequence of $\mathcal{B}$-valued simple functions f_n on Ω such that

$$\lim_{n\to\infty} \|f_n(t) - f(t)\|_{\mathcal{B}} = 0 \text{ for } \mu - \text{a.e. } t \in \Omega,$$

where $\mu -$ a.e. stands for "everywhere except for a set of zero μ measure". Finally, a $\mathcal{B}$-valued function f on Ω is called *μ-Bochner integrable* if there exists a sequence of simple functions $f_n : \Omega \to \mathcal{B}$ such that

$$\lim_{n\to\infty} \int_{\Omega} \|f_n(t) - f(t)\|_{\mathcal{B}} \, d\mu(t) = 0. \tag{14}$$

The integral of a simple function f of the form (13) on any $E \in \mathcal{F}$ with respect to μ is defined by

$$\int_E f d\mu := \sum_{j=1}^{n} a_j \mu(E_j \cap E).$$

In general, suppose that f is a μ-Bochner integrable function from Ω to $\mathcal{B}$, that is, (14) holds true. Then it is obvious that for each $E \in \mathcal{F}$, $\int_E f_n d\mu$, $n \in \mathbb{N}$ form a Cauchy sequence in $\mathcal{B}$. Therefore,

$$\int_E f d\mu := \lim_{n\to\infty} \int_E f_n d\mu.$$

The resulting integral $\int_E f d\mu$ is an element in $\mathcal{B}$.

It is known that a μ-measurable function $f : \Omega \to \mathcal{B}$ is Bochner integrable if and only if

$$\int_\Omega \|f(t)\|_{\mathcal{B}} d\mu(t) < +\infty.$$

This provides a way for us to comprehend the integral $\int_E f d\mu$ in the most needed case when f is $\mathcal{L}(\Lambda)$-valued. If $\mathcal{B} = \mathcal{L}(\Lambda)$ then we have for each $E \in \mathcal{F}$ that

$$\left(\int_E f d\mu \xi, \eta \right)_\Lambda = \int_E (f(t)\xi, \eta)_\Lambda d\mu(t), \ \ \xi, \eta \in \Lambda. \tag{15}$$

Clearly, the right hand side above defines a sesquilinear form on $\Lambda \times \Lambda$ which is bounded as

$$\left| \int_E (f(t)\xi, \eta)_\Lambda d\mu(t) \right| \le \int_E \|f(t)\|_{\mathcal{L}(\Lambda)} d\mu(t) \ \|\xi\|_\Lambda \|\eta\|_\Lambda,$$

where $\|\cdot\|_{\mathcal{L}(\Lambda)}$ is the operator norm on $\mathcal{L}(\Lambda)$. As a result, (15) gives an equivalent way of defining the integral $\int_E f d\mu$ as a bounded linear operator on Λ (Conway, 1990).

We introduce another notation before returning to reproducing kernels. Denote by $L^2(\Omega, \mathcal{B}, d\mu)$ the Banach space of all the μ-measurable functions $f : \Omega \to \mathcal{B}$ such that

$$\|f\|_{L^2(\Omega,\mathcal{B},d\mu)} := \left(\int_\Omega \|f(t)\|_{\mathcal{B}}^2 d\mu(t) \right)^{1/2} < +\infty. \tag{16}$$

When $\mathcal{B} = \mathbb{C}$, $L^2(\Omega, \mathbb{C}, d\mu)$ will be abbreviated as $L^2(\Omega, d\mu)$. When $\mathcal{B}$ is a Hilbert space, $L^2(\Omega, \mathcal{B}, d\mu)$ is also a Hilbert space with the inner product

$$(f, g)_{L^2(\Omega,\mathcal{B},d\mu)} := \int_\Omega (f(t), g(t))_{\mathcal{B}} d\mu(t), \ \ f, g \in L^2(\Omega, \mathcal{B}, d\mu).$$

The discussion in this section by far can be found in Diestel and Uhl (1977).

Let μ, ν be two finite nonnegative measures on a σ-algebra $\mathcal{F}$ of subsets of Ω. To introduce our $\mathcal{L}(\Lambda)$-valued reproducing kernels, we also let $\mathcal{W}$ be a Hilbert space and ϕ a mapping from $X \times \Omega$ to $\mathcal{L}(\Lambda, \mathcal{W})$ such that for each $x \in X$, $\phi(x, \cdot)$ belongs to both $L^2(\Omega, \mathcal{L}(\Lambda, \mathcal{W}), d\mu)$ and $L^2(\Omega, \mathcal{L}(\Lambda, \mathcal{W}), d\nu)$. We shall investigate conditions that ensure $\mathcal{H}_K \preceq \mathcal{H}_G$ where

$$K(x, y) = \int_\Omega \phi(y, t)^* \phi(x, t) d\mu(t), \ \ x, y \in X \tag{17}$$

and

$$G(x,y) = \int_\Omega \phi(y,t)^* \phi(x,t) d\nu(t), \quad x,y \in X, \tag{18}$$

where $\phi(y,t)^*$ is the adjoint operator of $\phi(y,t)$. Note that K, G are well-defined as the integrand is Bochner integrable with respect to both μ and ν. For instance, we observe by the Cauchy-Schwartz inequality for all $x,y \in X$ that

$$\begin{aligned}\int_\Omega \|\phi(y,t)^*\phi(x,t)\|_{\mathcal{L}(\Lambda)} d\mu(t) &\leq \int_\Omega \|\phi(y,t)^*\|_{\mathcal{L}(\mathcal{W},\Lambda)} \|\phi(x,t)\|_{\mathcal{L}(\Lambda,\mathcal{W})} d\mu(t) \\ &= \int_\Omega \|\phi(y,t)\|_{\mathcal{L}(\Lambda,\mathcal{W})} \|\phi(x,t)\|_{\mathcal{L}(\Lambda,\mathcal{W})} d\mu(t) \\ &\leq \|\phi(y,\cdot)\|_{L^2(\Omega,\mathcal{L}(\Lambda,\mathcal{W}),d\mu)} \|\phi(x,\cdot)\|_{L^2(\Omega,\mathcal{L}(\Lambda,\mathcal{W}),d\mu)}.\end{aligned}$$

An alternative of expressing K, G is for all $x,y \in X$, $\xi, \eta \in \Lambda$ that

$$\tilde{K}((x,\xi),(y,\eta)) = (K(x,y)\xi,\eta)_\Lambda = \int_\Omega (\phi(x,t)\xi, \phi(y,t)\eta)_\mathcal{W} d\mu(t)$$

and

$$\tilde{G}((x,\xi),(y,\eta)) = (G(x,y)\xi,\eta)_\Lambda = \int_\Omega (\phi(x,t)\xi, \phi(y,t)\eta)_\mathcal{W} d\nu(t).$$

When $\Lambda = \mathcal{W} = \mathbb{C}$, a characterization of $\mathcal{H}_K \preceq \mathcal{H}_G$ in terms of μ, ν has been established in Xu and Zhang (2009). The relation, between the two measures, which we shall need is absolute continuity. We say that μ is *absolutely continuous* with respect to ν if for all $E \in \mathcal{F}$, $\nu(E) = 0$ implies $\mu(E) = 0$. In this case, by the Radon-Nikodym theorem (see, Rudin, 1987, page 121) for scalar-valued measures, there exists a nonnegative ν-integrable function, denoted by $d\mu/d\nu$, such that

$$\mu(E) = \int_E \frac{d\mu}{d\nu}(t) d\nu(t) \text{ for all } E \in \mathcal{F}.$$

We write $\mu \preceq \nu$ if μ is absolutely continuous with respect to ν and $d\mu/d\nu \in \{0,1\}$ $\nu-$a.e. Thus, $\mu \preceq \nu$ if and only if μ is the restriction of ν on some measurable set in $\mathcal{F}$.

When $\Lambda = \mathcal{W} = \mathbb{C}$, it was proved in Theorem 8 of Xu and Zhang (2009) that if $\text{span}\{\phi(x,\cdot) : x \in X\}$ is dense in both $L^2(\Omega, d\mu)$ and $L^2(\Omega, d\nu)$ then G is a refinement of K if and only if $\mu \preceq \nu$. If $\mu \preceq \nu$ then G is a nontrivial refinement of K if and only if $\nu(\Omega) > \mu(\Omega)$.

Theorem 7 *Let K, G be given by (17) and (18). If $\text{span}\{\phi(x,\cdot)\xi : x \in X,\ \xi \in \Lambda\}$ is dense in both $L^2(\Omega, \mathcal{W}, d\mu)$ and $L^2(\Omega, \mathcal{W}, d\nu)$ then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mu \preceq \nu$. In this case, the refinement G of K is nontrivial if and only if $\nu(\Omega) - \mu(\Omega) > 0$.*

Proof When $\mathcal{W} = \mathbb{C}$, as a direct consequence of Theorem 8 in Xu and Zhang (2009), $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$ if and only if $\mu \preceq \nu$. The result hence follows from Proposition 6. When $\mathcal{W}$ is a general Hilbert space, it can be proved by arguments similar to those in Xu and Zhang (2009). ■

4.2 Scalar-valued Kernels with Respect to Operator-valued Measures

Again, $\mathcal{B}$ is a Banach space and $\mathcal{F}$ denotes a σ-algebra consisting of subsets of a fixed set Ω. A $\mathcal{B}$-valued measure on $\mathcal{F}$ is a function from $\mathcal{F}$ to $\mathcal{B}$ that is countably additive in the sense that for every sequence of pairwise disjoint sets $E_j \in \mathcal{F}$, $j \in \mathbb{N}$

$$\mu\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \mu(E_j),$$

where the series converges in the norm of $\mathcal{B}$. Every $\mathcal{B}$-valued measure μ on $\mathcal{F}$ comes with a scalar-valued measure $|\mu|$ on $\mathcal{F}$ defined by

$$|\mu|(E) := \sup_{\mathcal{P}} \sum_{F\in\mathcal{P}} \|\mu(F)\|_{\mathcal{B}}, \ \ E \in \mathcal{F}, \tag{19}$$

where the supremum is taken over all partitions $\mathcal{P}$ of E into countably many pairwise disjoint members of $\mathcal{F}$. We call $|\mu|$ the *variation* of μ and shall only work with these vector-valued measures μ that are of *bounded variation*, that is, $|\mu|(\Omega) < +\infty$. We note that μ vanishes on sets of zero $|\mu|$ measure. It implies that μ is absolutely continuous with respect to $|\mu|$ in the sense that

$$\lim_{|\mu(E)|\to 0} \mu(E) = 0.$$

The only type of integration that we shall need is to integrate a bounded $\mathcal{F}$-measurable scalar-valued function with respect to a $\mathcal{B}$-valued measure of bounded variation. Denote by $L^\infty(\Omega, d|\mu|)$ the Banach space of essentially bounded $\mathcal{F}$-measurable functions on Ω with the norm

$$\|f\|_{L^\infty(\Omega,d|\mu|)} := \inf\{a \geq 0 : |\mu|(\{|f| > a\}) = 0\}.$$

For a simple function $f : \Omega \to \mathbb{C}$ of the form

$$f = \sum_{j=1}^{n} \alpha_j \chi_{E_j},$$

where $\alpha_j \in \mathbb{C}$ and E_j are pairwise disjoint members in $\mathcal{F}$, we define

$$\int_E f d\mu := \sum_{j=1}^{n} \alpha_j \mu(E_j \cap E), \ \ E \in \mathcal{F}.$$

Clearly,

$$\left\| \int_E f d\mu \right\|_{\mathcal{B}} \leq \|f\|_{L^\infty(\Omega,d|\mu|)} |\mu|(E).$$

Therefore, the map sending a simple function f to $\int_E f d\mu$ can be uniquely extended to a bounded linear operator from $L^\infty(\Omega, d|\mu|)$ to $\mathcal{B}$. The outcome of the application of the resulting operator on a general $f \in L^\infty(\Omega, d|\mu|)$ is still denoted by $\int_E f d\mu$. This is how the $\mathcal{B}$-valued integral is defined.

It is time to present the second type of reproducing kernels defined by integration:

$$K(x,y) := \int_\Omega \Psi(x,y,t) d\mu(t), \ \ x,y \in X, \tag{20}$$

where μ is an $\mathcal{L}_+(\Lambda)$-valued measure on $\mathcal{F}$ of bounded variation, and Ψ is a scalar-valued function such that $\Psi(\cdot,\cdot,t)$ is a scalar-valued reproducing kernel on X for all $t \in \Omega$ and for all $x,y \in X$, $\Psi(x,y,\cdot)$ is bounded and $\mathcal{F}$-measurable. We verify that (20) indeed defines an $\mathcal{L}(\Lambda)$-valued reproducing kernel.

Proposition 8 *With the above assumptions on Ψ and μ, the function K defined by (20) is an $\mathcal{L}(\Lambda)$-valued reproducing kernel on X.*

Proof Fix finite $x_j \in X$ and $\xi_j \in \Lambda$, $j \in \mathbb{N}_n$. For any $\varepsilon > 0$, there exist simple functions

$$f_{j,k} := \sum_{l=1}^{m} \alpha_{j,k,l} \chi_{E_l}, \quad j,k \in \mathbb{N}_n$$

such that

$$\|\Psi(x_j, x_k, \cdot) - f_{j,k}\|_{L^\infty(\Omega, d|\mu|)} < \varepsilon, \quad j,k \in \mathbb{N}_n. \tag{21}$$

Here, $\alpha_{j,k,l} \in \mathbb{C}$ and E_l are pairwise disjoint sets in $\mathcal{F}$ with $|\mu|(E_l) > 0$, $l \in \mathbb{N}_m$. By (21) and the definition of integration in this section,

$$\left| \sum_{j=1}^{n} \sum_{k=1}^{n} (K(x_j, x_k)\xi_j, \xi_k)_\Lambda - \sum_{j=1}^{n} \sum_{k=1}^{n} \left(\left(\int_\Omega f_{j,k} d\mu \right) \xi_j, \xi_k \right)_\Lambda \right| \leq \varepsilon |\mu|(\Omega) \left(\sum_{j=1}^{n} \|\xi_j\|_\Lambda \right)^2. \tag{22}$$

We may choose by (21) for each $l \in \mathbb{N}_m$ some $t_l \in E_l$ such that

$$\left| \Psi(x_j, x_k, t_l) - \alpha_{j,k,l} \right| \leq \varepsilon.$$

Letting

$$S := \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{m} \Psi(x_j, x_k, t_l)(\mu(E_l)\xi_j, \xi_k)_\Lambda,$$

we get by the above equation that

$$\begin{aligned} \left| \sum_{j=1}^{n} \sum_{k=1}^{n} \left(\left(\int_\Omega f_{j,k} d\mu \right) \xi_j, \xi_k \right)_\Lambda - S \right| &\leq \left| \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{m} |\alpha_{j,k,l} - \Psi(x_j, x_k, t_l)| (\mu(E_l)\xi_j, \xi_k)_\Lambda \right| \\ &\leq \varepsilon \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{m} \|\mu(E_l)\|_{\mathcal{L}(\Lambda)} \|\xi_j\|_\Lambda \|\xi_k\|_\Lambda \leq \varepsilon |\mu|(\Omega) \left(\sum_{j=1}^{n} \|\xi_j\|_\Lambda \right)^2. \end{aligned} \tag{23}$$

Combining (22) and (23) yields that

$$\left| \sum_{j=1}^{n} \sum_{k=1}^{n} (K(x_j, x_k)\xi_j, \xi_k)_\Lambda - S \right| \leq 2\varepsilon |\mu|(\Omega) \left(\sum_{j=1}^{n} \|\xi_j\|_\Lambda \right)^2. \tag{24}$$

Since $\Psi(\cdot, \cdot, t_l)$ is a scalar-valued reproducing kernel on X, $[\Psi(x_j, x_k, t_l) : j,k \in \mathbb{N}_n]$ is a positive semi-definite matrix for each $l \in \mathbb{N}_m$. So are $[(\mu(E_l)\xi_j, \xi_k)_\Lambda : j,k \in \mathbb{N}_n]$, $l \in \mathbb{N}_m$ as $\mu(E_l) \in \mathcal{L}_+(\Lambda)$. By the Schur product theorem (see, for example, Horn and Johnson, 1991, page 309), the Hadamard product of two positive semi-definite matrices remains positive semi-definite. We obtain by this fact that $S > 0$, which together with (24), and the fact that ε can be arbitrarily small, proves (1). ■

To investigate the refinement relationship, we shall consider a simplified version of (20) that covers a large class of operator-valued reproducing kernels. Let $\phi : X \times \Omega \to \mathbb{C}$ be such that $\phi(x, \cdot)$ is a bounded $\mathcal{F}$-measurable function for every $x \in X$ and such that

$$\overline{\text{span}}\{\phi(x, \cdot) : x \in X\} = L^2(\Omega, d\gamma) \text{ for any finite nonnegative measure } \gamma \text{ on } \mathcal{F}. \tag{25}$$

We shall see by the concrete examples in the next section that the denseness requirement (25) is not too restricted in applications. The kernels we shall consider are

$$K(x,y) := \int_\Omega \phi(x,t)\overline{\phi(y,t)}d\mu(t), \quad x,y \in X \tag{26}$$

and

$$G(x,y) := \int_\Omega \phi(x,t)\overline{\phi(y,t)}d\nu(t), \quad x,y \in X, \tag{27}$$

where μ,ν are two $\mathcal{L}_+(\Lambda)$-valued measures on $\mathcal{F}$ of bounded variation. By Proposition 8, K,G are $\mathcal{L}(\Lambda)$-valued reproducing kernels on X. Our idea is to use the Radon-Nikodym property of vector-valued measures to study the refinement property.

Let $\mathcal{B}$ be a Banach space and γ a finite nonnegative measure on $\mathcal{F}$. We say that a $\mathcal{B}$-valued measure ρ on $\mathcal{F}$ of bounded variation has the *Radon-Nikodym property* with respect to γ if there is a γ-Bochner integrable function $\Gamma : \Omega \to \mathcal{L}_+(\Lambda)$ such that for all $E \in \mathcal{F}$

$$\rho(E) = \int_E \Gamma d\gamma.$$

Apparently, this could only be true when ρ is absolutely continuous with respect to γ. For this reason, we also say that the space $\mathcal{B}$ has the Radon-Nikodym property with respect to γ if every $\mathcal{B}$-valued measure of bounded variation that is absolutely continuous with respect to γ has the Radon-Nikodym property with respect to γ. Moreover, $\mathcal{B}$ is said to have the Radon-Nikodym property if it has it with respect to any finite nonnegative measure on any measure space $\mathcal{F}$.

Strikingly different from the scalar-valued case, a Banach space $\mathcal{B}$ may not have the Radon-Nikodym property. For instance, the Banach space c_0 of all sequences $\alpha := (\alpha_j \in \mathbb{C} : j \in \mathbb{N})$ with

$$\lim_{j\to\infty} |\alpha_j| = 0$$

under the norm $\|\alpha\|_{c_0} := \sup\{|\alpha_j| : j \in \mathbb{N}\}$ does not have the property with respect to the Lebesgue measure (see, Diestel and Uhl, 1977, page 60). Consequently, the space $\mathcal{L}(\Lambda)$ does not have the Radon-Nikodym property when Λ is infinite-dimensional. To see this, since Λ is separable we let $\{e_j : j \in \mathbb{N}\}$ be an orthonormal basis for Λ. Denote by $\mathcal{L}_0(\Lambda)$ the set of all the operators $T \in \mathcal{L}(\Lambda)$ such that

$$Te_j = \alpha_j e_j, \quad j \in \mathbb{N}$$

for some $\alpha \in c_0$. One sees that $\|T\|_{\mathcal{L}(\Lambda)} = \|\alpha\|_{c_0}$ (Conway, 1990). As a result, $\mathcal{L}_0(\Lambda)$ is a closed subspace of $\mathcal{L}(\Lambda)$ that is isometrically isomorphic to c_0. Since c_0 does not have the Radon-Nikodym property, neither does $\mathcal{L}_0(\Lambda)$. A Banach space has the Radon-Nikodym property if and only if each of its closed linear subspaces does (Diestel and Uhl, 1977). By this fact, $\mathcal{L}(\Lambda)$ does not have Radon-Nikodym property.

We shall focus on the situation where this desired property holds. For example, reflexive Banach spaces have the Radon-Nikodym property (Diestel and Uhl, 1977). In applications, Λ is usually finite-dimensional. In this case, $\mathcal{L}(\Lambda)$ is of finite dimension as well. Any two norms on a finite-dimensional Banach space are equivalent and a finite-dimensional $\mathcal{L}(\Lambda)$ can be endowed with a norm that makes it a Hilbert space. It yields that $\mathcal{L}(\Lambda)$ is reflexive. The conclusion is that when Λ is finite-dimensional, $\mathcal{L}(\Lambda)$ does have the Radon-Nikodym property. Another way of overcoming the difficulty is to confine to a subclass of $\mathcal{L}(\Lambda)$, for example, to the Schatten class (Birman and

Solomjak, 1987). Denote for each compact operator $T \in \mathcal{L}(\Lambda)$ by $s_j(T)$, $j \in \mathbb{N}$, the nonnegative square root of the j-th largest eigenvalue of T^*T. It is called the j-th *singular number* of T. For $p \in (1,+\infty)$, the p-th Schatten class $\mathcal{S}_p(\Lambda)$ consists of all the compact linear operators $T \in \mathcal{L}(\Lambda)$ with the norm

$$\|T\|_{\mathcal{S}_p(\Lambda)} := \left(\sum_{j=1}^{\infty} (s_j(T))^p\right)^{1/p} < +\infty.$$

The p-th Schatten class $S_p(\Lambda)$ is a reflexive Banach space and hence has the Radon-Nikodym property. When $p = 2$, $S_2(\Lambda)$ is the class of Hilbert-Schmidt operators and

$$\|T\|_{\mathcal{S}_2(\Lambda)} = \left(\sum_{j=1}^{\infty} \|Te_j\|_{\Lambda}\right)^{1/2}.$$

We shall not go into further details about the Radon-Nikodym property. Interested readers are referred to Chapter III of Diestel and Uhl (1977) and the references therein.

The assumption we shall need is that there exists a finite nonnegative measure γ on $\mathcal{F}$ such that both μ and ν have the Radon-Nikodym property with respect to γ. In other words, there exist γ-Bochner integrable functions $\Gamma_\mu, \Gamma_\nu : \Omega \to \mathcal{L}_+(\Lambda)$ such that

$$\mu(E) = \int_E \Gamma_\mu d\gamma \quad \text{and} \quad \nu(E) = \int_E \Gamma_\nu d\gamma \quad \text{for all } E \in \mathcal{F}. \tag{28}$$

Such two functions exist if $\gamma := |\mu| + |\nu|$ and μ, ν take values in the p-th Schatten class of $\mathcal{L}(\Lambda)$, $1 < p < +\infty$.

Suppose that K, G are given by (26) and (27), where ϕ, μ, ν satisfy (25) and (28). Our purpose is to investigate $\mathcal{H}_K \preceq \mathcal{H}_G$. To this end, let us first identify $\mathcal{H}_{\tilde{K}}$ and $\mathcal{H}_{\tilde{G}}$. We shall only present results for $\mathcal{H}_{\tilde{K}}$ as those for $\mathcal{H}_{\tilde{G}}$ have a similar form.

Lemma 9 *The RKHS $\mathcal{H}_{\tilde{K}}$ consists of functions F_f of the form*

$$F_f(x,\xi) := \int_\Omega (\Gamma_\mu(t) f(t), \xi)_\Lambda \overline{\phi(x,t)} d\gamma(t), \quad x \in X,\ \xi \in \Lambda,$$

where f can be an arbitrary element from the Hilbert space $\mathcal{W}_\mu$ of γ-measurable functions from Ω to Λ such that

$$\|f\|_{\mathcal{W}_\mu} := \left(\int_\Omega (\Gamma_\mu(t) f(t), f(t))_\Lambda d\gamma(t)\right)^{1/2} < +\infty.$$

Moreover, $\|F_f\|_{\mathcal{H}_{\tilde{K}}} = \|f\|_{\mathcal{W}_\mu}$ for all $f \in \mathcal{W}_\mu$.

Proof We observe for all $x, y \in X$ and $\xi, \eta \in \Lambda$ that

$$\tilde{K}((x,\xi),(y,\eta)) = \int_\Omega \phi(x,t)\overline{\phi(y,t)} (\Gamma_\mu(t)\xi, \eta)_\Lambda d\gamma(t).$$

Thus, we may choose $\mathcal{W}_\mu$ as a feature space for $\tilde{K}$. The associated feature map $\Phi_\mu : X \times \Lambda \to \mathcal{W}_\mu$ is then selected as

$$\Phi_\mu(x,\xi)(t) := \phi(x,t)\xi, \quad t \in \Omega.$$

We next verify the denseness condition that $\overline{\text{span}}\{\Phi_\mu(x,\xi) : x \in X,\ \xi \in \Lambda\} = \mathcal{W}_\mu'$. Suppose that $f \in \mathcal{W}_\mu'$ is orthogonal to $\Phi_\mu(x,\xi)$ for all $x \in X$ and $\xi \in \Lambda$, that is,

$$\int_\Omega (\Gamma_\mu f(t),\xi)_\Lambda \overline{\phi(x,t)} d\gamma(t) = 0 \text{ for all } x \in X,\ \xi \in \Lambda.$$

By (25),

$$(\Gamma_\mu(t) f(t),\xi)_\Lambda = 0\ \gamma-\text{a.e.}$$

As this holds for an arbitrary $\xi \in \Lambda$, $\Gamma_\mu(t) f(t) = 0\ \gamma-$a.e. It implies that $\|f\|_{\mathcal{W}_\mu'} = 0$. The result now follows immediately from Lemma 2. ■

For two operators $A, B \in \mathcal{L}_+(\Lambda)$, we write $A \preceq B$ if for all $\xi \in \Lambda$ there exists some $\eta \in \Lambda$ such that

$$A\xi = B\eta \text{ and } (A\xi,\xi)_\Lambda = (B\eta,\eta)_\Lambda. \tag{29}$$

We make a simple observation about this special relationship between two linear operators.

Let $\ker(A)$ and $\text{ran}\,(A)$ be the kernel and range of A, respectively. If $\text{ran}\,(A)$ is closed then as A is self-adjoint, there holds the direct sum decomposition

$$\Lambda = \ker(A) \oplus \text{ran}\,(A). \tag{30}$$

Thus, A is bijective and bounded from $\text{ran}\,(A)$ to $\text{ran}\,(A)$. By the open mapping theorem, it has a bounded inverse on $\text{ran}\,(A)$, which we denote by A^{-1}.

Proposition 10 *Suppose that $A, B \in \mathcal{L}_+(\Lambda)$ have closed range. Then $A \preceq B$ if and only if*

$$\text{ran}\,(A) \subseteq \text{ran}\,(B) \tag{31}$$

and

$$P_{B,A} B^{-1} = A^{-1} \text{ on } \text{ran}\,(A), \tag{32}$$

where $P_{B,A}$ denotes the orthogonal projection from $\text{ran}\,(B)$ *to* $\text{ran}\,(A)$. *Particularly, if A is onto then $A \preceq B$ if and only if $A = B$.*

Proof Let A, B have closed range. Suppose first that $A \preceq B$. Then (31) clearly holds true. Set for each $\xi \in \text{ran}\,(A)$

$$\eta_\xi := B^{-1} A\xi.$$

Clearly, the mapping $\xi \to \eta_\xi$ is linear from $\text{ran}\,(A)$ to $\text{ran}\,(B)$. Thus, we have for arbitrary $\xi, \xi' \in \Lambda$ that

$$(A\xi' + A\xi, \xi' + \xi)_\Lambda = (B\eta_{\xi'+\xi}, \eta_{\xi'+\xi})_\Lambda = (B\eta_{\xi'} + B\eta_\xi, \eta_{\xi'} + \eta_\xi)_\Lambda,$$

which implies that

$$\text{Re}\,(A\xi',\xi)_\Lambda = \text{Re}\,(B\eta_{\xi'}, \eta_\xi)_\Lambda.$$

A textbook trick yields that for all $\xi, \xi' \in \text{ran}\,(A)$,

$$(A\xi',\xi)_\Lambda = (B\eta_{\xi'}, \eta_\xi)_\Lambda = (A\xi', \eta_\xi)_\Lambda.$$

We hence obtain that $\xi - \eta_\xi \in \ker(A)$ for all $\xi \in \mathrm{ran}\,(A)$. Consequently,

$$A\xi - AB^{-1}A\xi = A\xi - A\eta_\xi = 0 \text{ for all } \xi \in \mathrm{ran}\,(A),$$

from which (32) follows.

On the other hand, suppose that (31) and (32) hold true. Then we choose for each $\xi \in \Lambda$

$$\eta := B^{-1}A\xi$$

and verify that $B\eta = A\xi$ and

$$(B\eta,\eta)_\Lambda = (A\xi, B^{-1}A\xi)_\Lambda = (A\xi, P_{B,A}B^{-1}A\xi)_\Lambda = (A\xi, A^{-1}A\xi)_\Lambda = (A\xi,\xi)_\Lambda.$$

Finally, if A is onto then by (31), $\mathrm{ran}\,(A) = \mathrm{ran}\,(B) = \Lambda$. According to (30), both A and B are injective. Therefore, they possess a bounded inverse on Λ. It implies that $P_{B,A}$ is the identity operator on Λ. By Equation (32), $A = B$. The proof is complete. ■

We are ready to present the main result of this section.

Theorem 11 *Let K, G be given by (26) and (27), where ϕ, μ, ν satisfy (25) and (28). Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\Gamma_\mu \preceq \Gamma_\nu$ $\gamma-$a.e.*

Proof By Proposition 6 and Lemma 9, $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if for all $f \in \mathcal{W}_\mu$, there exists some $g \in \mathcal{W}_\nu$ such that

$$\int_\Omega (\Gamma_\mu(t) f(t),\xi)_\Lambda \overline{\phi(x,t)} d\gamma(t) = \int_\Omega (\Gamma_\nu(t) g(t),\xi)_\Lambda \overline{\phi(x,t)} d\gamma(t) \text{ for all } x \in X,\ \xi \in \Lambda \tag{33}$$

and

$$\int_\Omega (\Gamma_\mu(t) f(t), f(t))_\Lambda d\gamma(t) = \int_\Omega (\Gamma_\nu(t) g(t), g(t))_\Lambda d\gamma(t). \tag{34}$$

By the denseness condition (25), (33) holds true if and only if

$$(\Gamma_\mu(t) f(t),\xi)_\Lambda = (\Gamma_\nu(t) g(t),\xi)_\Lambda \text{ for } \gamma-\text{a.e. } t \in \Omega \text{ and all } \xi \in \Lambda,$$

which is equivalent to

$$\Gamma_\mu(t) f(t) = \Gamma_\nu(t) g(t) \text{ for } \gamma-\text{a.e. } t \in \Omega. \tag{35}$$

We conclude that $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if for every $f \in \mathcal{W}_\mu$, there exists some $g \in \mathcal{W}_\nu$ such that Equations (34) and (35) hold true.

Suppose that $\Gamma_\mu \preceq \Gamma_\nu$ $\gamma-$a.e. Then clearly, for each $f \in \mathcal{W}_\mu$, we can find a function $g : \Omega \to \Lambda$ which is defined γ-almost everywhere and satisfies (35) and

$$(\Gamma_\mu(t) f(t), f(t))_\Lambda = (\Gamma_\nu(t) g(t), g(t))_\Lambda \text{ for } \gamma-\text{a.e. } t \in \Omega.$$

The above equation implies (34). Therefore, $\mathcal{H}_K \preceq \mathcal{H}_G$.

On the other hand, suppose that we can find for every $f \in \mathcal{W}_\mu$ some $g_f \in \mathcal{W}_\nu$ satisfying (34) and (35). The function g_f can be chosen so that $f \to g_f$ is linear from $\mathcal{W}_\mu$ to $\mathcal{W}_\nu$. A trick similar to that used in Lemma 9 enables us to obtain from (34) and (35) that

$$\int_\Omega (\Gamma_\mu(t) f'(t), f(t) - g_f(t))_\Lambda d\gamma(t) = 0 \text{ for all } f' \in \mathcal{W}_\mu.$$

Letting $f' := \phi(x,\cdot)\xi$ for arbitrary $x \in X$ and $\xi \in \Lambda$ in the above equation and invoking (25), we have that

$$\Gamma_\mu(t)(f(t) - g_f(t)) = 0 \text{ for } \gamma - \text{a.e. } t \in \Omega.$$

By the above equation and (35), we get for γ-almost every $t \in \Omega$ that

$$(\Gamma_\nu(t)g_f(t), g_f(t))_\Lambda = (\Gamma_\mu(t)f(t), g_f(t))_\Lambda = (f(t), \Gamma_\mu(t)g_f(t))_\Lambda = (f(t), \Gamma_\mu(t)f(t))_\Lambda = (\Gamma_\mu(t)f(t), f(t))_\Lambda.$$

Since (35) and the above equation are true for an arbitrary $f \in \mathcal{W}_\mu$, $\Gamma_\mu \preceq \Gamma_\nu$ $\gamma -$ a.e. ■

5. Examples

We present in this section several concrete examples of refinement of operator-valued reproducing kernels. They are built on the general characterizations established in the last two sections.

5.1 Translation Invariant Reproducing Kernels

Let $d \in \mathbb{N}$ and K be an $\mathcal{L}(\Lambda)$-valued reproducing kernel on $\mathbb{R}^d$. We say that K is *translation invariant* if for all $x, y, a \in \mathbb{R}^d$

$$K(x-a, y-a) = K(x,y).$$

A celebrated characterization due to Bochner (1959) states that every continuous scalar-valued translation invariant reproducing kernel on $\mathbb{R}^d$ must be the Fourier transform of a finite nonnegative Borel measure on $\mathbb{R}^d$, and vice versa. This result has been generalized to the operator-valued case (Berberian, 1966; Carmeli et al., 2010; Fillmore, 1970). Specifically, a continuous function K from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathcal{L}(\Lambda)$ is a translation invariant reproducing kernel if and only if it has the form

$$K(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\mu(t), \quad x, y \in \mathbb{R}^d, \tag{36}$$

for some $\mu \in \mathcal{B}(\mathbb{R}^d, \Lambda)$, the set of all the $\mathcal{L}_+(\Lambda)$-valued measures of bounded variation on the σ-algebra of Borel subsets in $\mathbb{R}^d$. Let G be the kernel given by

$$G(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\nu(t), \quad x, y \in \mathbb{R}^d, \tag{37}$$

where $\nu \in \mathcal{B}(\mathbb{R}^d, \Lambda)$. The purpose of this subsection is to characterize $\mathcal{H}_K \preceq \mathcal{H}_G$ in terms of μ, ν. To this end, we first investigate the structure of the RKHS of a translation invariant $\mathcal{L}(\Lambda)$-valued reproducing kernel.

Let γ be an arbitrary measure in $\mathcal{B}(\mathbb{R}^d, \Lambda)$ and L the associated translation invariant reproducing kernel defined by

$$L(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\gamma(t), \quad x, y \in \mathbb{R}^d.$$

There exists a decomposition of γ with respect to the Lebesgue measure dx on $\mathbb{R}^d$ (Diestel and Uhl, 1977) as follows:

$$\gamma = \gamma_c + \gamma_s, \tag{38}$$

where γ_c, γ_s are the unique measures in $\mathcal{B}(\mathbb{R}^d, \Lambda)$ such that γ_c is absolutely continuous with respect to dx, and for each continuous linear functional λ on $\mathcal{L}(\Lambda)$, the scalar-valued measure $\lambda\gamma_s$ and dx are mutually singular. It follows from this decomposition of measures a decomposition of L:

$$L = L_c + L_s,$$

where

$$L_c(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\gamma_c(t), \;\; L_s(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\gamma_s(t), \quad x,y \in \mathbb{R}^d. \tag{39}$$

Our first observation is that $\mathcal{H}_L$ is the orthogonal direct sum of $\mathcal{H}_{L_c}$ and $\mathcal{H}_{L_s}$. Two lemmas are needed to prove this useful fact.

Lemma 12 *Let L_c, L_s be given by (39). Then for all $\xi \in \Lambda$ and $x,y \in \mathbb{R}^d$*

$$(L_a(x,y)\xi,\xi)_\Lambda = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\gamma_{a,\xi}(t), \;\; a = c \text{ or } s, \tag{40}$$

where $\gamma_{a,\xi}$ is a scalar-valued Borel measure on $\mathbb{R}^d$ defined for each Borel set $E \subseteq \mathbb{R}^d$ by

$$\gamma_{a,\xi}(E) := (\gamma_a(E)\xi,\xi)_\Lambda, \;\; a = c \text{ or } s.$$

Proof Let $a \in \{c,s\}$, $\xi \in \Lambda$, $x,y \in \mathbb{R}^d$, and s_n be a sequence of simple functions on $\mathbb{R}^d$ that converges to $e^{i(x-y)\cdot t}$ in $L^\infty(\mathbb{R}^d, dx)$. Then

$$\lim_{n\to\infty} \left(\left(\int_{\mathbb{R}^d} s_n d\gamma_a \right) \xi, \xi \right)_\Lambda = (L_a(x,y)\xi,\xi)_\Lambda.$$

By definition, we have for each $n \in \mathbb{N}$ that

$$\lim_{n\to\infty} \left(\left(\int_{\mathbb{R}^d} s_n d\gamma_a \right) \xi, \xi \right)_\Lambda = \int_{\mathbb{R}^d} s_n d\gamma_{a,\xi}.$$

As

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} s_n d\gamma_{a,\xi} = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\gamma_{a,\xi}(t),$$

we conclude from the previous two equations that (40) holds true. ■

Lemma 13 *There holds $\mathcal{H}_{L_c} \cap \mathcal{H}_{L_s} = \{0\}$.*

Proof We introduce for each $\xi \in \Lambda$ two scalar-valued translation invariant reproducing kernels on $\mathbb{R}^d$ by setting

$$A_a(x,y) := (L_a(x,y)\xi,\xi)_\Lambda, \;\; x,y \in \mathbb{R}^d, \;\; a \in \{c,s\}.$$

By Lemma 12, we have the alternative representations for A_c and A_s

$$A_a(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\gamma_{a,\xi}(t), \;\; x,y \in \mathbb{R}^d, \;\; a = c \text{ or } s.$$

By the Lebesgue decomposition of γ, $\gamma_{c,\xi}$ is absolutely continuous with respect to dx while $\gamma_{s,\xi}$ and dx are mutually singular. As a consequence, $\mathcal{H}_{A_c} \cap \mathcal{H}_{A_s} = \{0\}$ by Lemma 17 in Xu and Zhang (2009).

Let $a \in \{c,s\}$. By (3),

$$A_a(x,y) = (L_a(x,\cdot)\xi, L_a(y,\cdot)\xi)_{\mathcal{H}_{L_a}}, \;\; x,y \in \mathbb{R}^d.$$

A feature map for A_a may hence be chosen as

$$\Phi_a(x) := L_a(x,\cdot)\xi, \;\; x \in \mathbb{R}^d$$

with the feature space being $\mathcal{H}_{L_a}$. We identify by Lemma 2 that

$$\mathcal{H}_{A_a} = \{(\tilde{f}(\cdot),\xi)_\Lambda : \tilde{f} \in \mathcal{H}_{L_a}\}. \tag{41}$$

Assume that $\mathcal{H}_{L_c} \cap \mathcal{H}_{L_s} \neq \{0\}$. Then there exist nontrivial functions $\tilde{f} \in \mathcal{H}_{L_c}$ and $\tilde{g} \in \mathcal{H}_{L_s}$ such that $\tilde{f} = \tilde{g}$. As a result, there exists some $\xi \in \Lambda$ such that $(\tilde{f}(\cdot),\xi)_\Lambda$ is not the trivial function. By equation (41)

$$(\tilde{f}(\cdot),\xi)_\Lambda = (\tilde{g}(\cdot),\xi)_\Lambda \in \mathcal{H}_{A_c} \cap \mathcal{H}_{A_s},$$

contradicting the fact that $\mathcal{H}_{A_c} \cap \mathcal{H}_{A_s} = \{0\}$. ■

Theorem 14 *The space $\mathcal{H}_L$ is the orthogonal direct sum of $\mathcal{H}_{L_c}$ and $\mathcal{H}_{L_s}$, namely, $\mathcal{H}_L = \mathcal{H}_{L_c} \bigoplus \mathcal{H}_{L_s}$.*

Proof The result follows directly from Lemma 13 and Proposition 1. ■

We are now in a position to study the refinement relationship $\mathcal{H}_K \preceq \mathcal{H}_G$, where K, G are defined by (36) and (37). Firstly, the task can be separated into two related ones according to the Lebesgue decomposition of measures μ, ν.

Proposition 15 *There holds $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ and $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$.*

Proof By Theorem 14, $\mathcal{H}_K = \mathcal{H}_{K_c} \bigoplus \mathcal{H}_{K_s}$ and $\mathcal{H}_G = \mathcal{H}_{G_c} \bigoplus \mathcal{H}_{G_s}$. Therefore, if $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ and $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$ then $\mathcal{H}_K \preceq \mathcal{H}_G$.

On the other hand, suppose that $\mathcal{H}_K \preceq \mathcal{H}_G$. Let $f \in \mathcal{H}_{K_c}$. Then $f \in \mathcal{H}_K$ and $\|f\|_{\mathcal{H}_{K_c}} = \|f\|_{\mathcal{H}_K}$. Since $\mathcal{H}_K \preceq \mathcal{H}_G$, there exists $g \in \mathcal{H}_{G_c}$ and $h \in \mathcal{H}_{G_s}$ such that

$$f = g + h$$

and

$$\|f\|^2_{\mathcal{H}_{K_c}} = \|f\|^2_{\mathcal{H}_K} = \|g+h\|^2_{\mathcal{H}_G} = \|g\|^2_{\mathcal{H}_{G_c}} + \|h\|^2_{\mathcal{H}_{G_s}}.$$

Therefore, to show that $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ it suffices to show that $h = 0$. Assume that $h \neq 0$. Note that $f - g \in \mathcal{H}_{K_c+G_c}$ (Pedrick, 1957), we get that

$$\mathcal{H}_{K_c+G_c} \cap \mathcal{H}_{G_s} \neq \{0\}. \tag{42}$$

However,

$$(K_c+G_c)(x,y)=\int_{\mathbb{R}^d}e^{i(x-y)\cdot t}d(\mu_c+\nu_c)(t),\ \ x,y\in\mathbb{R}^d$$

and $\mu_c+\nu_c$ is absolutely continuous with respect to dx. Thus, Equation (42) contradicts Lemma 13. The contradiction proves that $\mathcal{H}_{K_c}\preceq\mathcal{H}_{G_c}$. Likewise, one can prove that $\mathcal{H}_{K_s}\preceq\mathcal{H}_{G_s}$. ■

By Proposition 15, we shall study $\mathcal{H}_{K_c}\preceq\mathcal{H}_{G_c}$ and $\mathcal{H}_{K_s}\preceq\mathcal{H}_{G_s}$ separately. We shall restrict ourselves to the case when the measures corresponding to K_c and G_c have the Radon-Nikodym property with respect to the Lebesgue measure and the measures corresponding to K_s and G_s are discrete. Specifically, the kernels to be considered are of the following special forms:

$$K_c(x,y):=\int_{\mathbb{R}^d}e^{i(x-y)\cdot t}\varphi_1(t)dt,\ \ G_c(x,y):=\int_{\mathbb{R}^d}e^{i(x-y)\cdot t}\varphi_2(t)dt,\ \ x,y\in\mathbb{R}^d \tag{43}$$

and

$$K_s(x,y):=\sum_{j\in\mathbb{J}_1}e^{i(x-y)\cdot t_j}A_j,\ \ G_s(x,y):=\sum_{k\in\mathbb{J}_2}e^{i(x-y)\cdot t_k}B_k,\ \ x,y\in\mathbb{R}^d.$$

Here, φ_1,φ_2 are two dx-Bochner integrable functions from $\mathbb{R}^d$ to $\mathcal{L}_+(\Lambda)$, $\{t_j:j\in\mathbb{J}_1\}$ and $\{t_k:k\in\mathbb{J}_2\}$ are countable sets of pairwise distinct points in $\mathbb{R}^d$, and A_j,B_j are nonzero operators in $\mathcal{L}_+(\Lambda)$ such that

$$\sum_{j\in\mathbb{J}_1}\|A_j\|_{\mathcal{L}(\Lambda)}<+\infty,\ \ \sum_{k\in\mathbb{J}_2}\|B_k\|_{\mathcal{L}(\Lambda)}<+\infty.$$

The following characterization is a direct consequence of Theorem 11.

Proposition 16 *Let K_c,G_c be given by (43). Then $\mathcal{H}_{K_c}\preceq\mathcal{H}_{G_c}$ if and only if $\varphi_1(t)\preceq\varphi_2(t)$ for almost every $t\in\mathbb{R}^d$ except for a subset in $\mathbb{R}^d$ of zero Lebesgue measure.*

Proof As φ_1,φ_2 are dx-Bochner integrable,

$$\int_{\mathbb{R}^d}\|\varphi_j(t)\|_{\mathcal{L}(\Lambda)}dt<+\infty,\ \ j=1,2.$$

Define a finite nonnegative Borel measure γ on $\mathbb{R}^d$ by setting for each Borel subset E in $\mathbb{R}^d$

$$\gamma(E):=\int_E\|\varphi_1(t)\|_{\mathcal{L}(\Lambda)}+\|\varphi_2(t)\|_{\mathcal{L}(\Lambda)}dt.$$

Evidently, K_c,G_c have the form

$$K_c(x,y)=\int_{\mathbb{R}^d}e^{i(x-y)\cdot t}\Gamma_1(t)d\gamma(t),\ \ G_c(x,y)=\int_{\mathbb{R}^d}e^{i(x-y)\cdot t}\Gamma_2(t)d\gamma(t),\ \ \ x,y\in\mathbb{R}^d,$$

where for $j=1,2$,

$$\Gamma_j(t):=\begin{cases}\dfrac{\varphi_j(t)}{\|\varphi_1(t)\|_{\mathcal{L}(\Lambda)}+\|\varphi_2(t)\|_{\mathcal{L}(\Lambda)}}, & \text{if } \|\varphi_1(t)\|_{\mathcal{L}(\Lambda)}+\|\varphi_2(t)\|_{\mathcal{L}(\Lambda)}>0,\\ 0, & \text{otherwise.}\end{cases}$$

It is also clear that $\operatorname{span}\{e^{ix\cdot t}:x\in\mathbb{R}^d\}$ is dense in $L^2(\mathbb{R}^d,d\gamma)$. By Theorem 11, $\mathcal{H}_{K_c}\preceq\mathcal{H}_{G_c}$ if and only if $\Gamma_1\preceq\Gamma_2\ \gamma-$a.e. Note that $\Gamma_1(t)\preceq\Gamma_2(t)$ if and only if $\varphi_1(t)\preceq\varphi_2(t)$. If $\varphi_1\preceq\varphi_2\ dx-$a.e. then

$\Gamma_1 \preceq \Gamma_2$ $\gamma-$a.e. as γ is absolutely continuous with respect to the Lebesgue measure. On the other hand, suppose that $\Gamma_1 \preceq \Gamma_2$ $\gamma-$a.e. Set

$$E := \{t \in \mathbb{R}^d : \|\varphi_1(t)\|_{\mathcal{L}(\Lambda)} + \|\varphi_2(t)\|_{\mathcal{L}(\Lambda)} > 0\}.$$

For $t \in E^c$, $\varphi_1(t) = \varphi_2(t) = 0$, and thus, $\varphi_1(t) \preceq \varphi_2(t)$. Assume that there exists a Borel subset $F \subseteq \mathbb{R}^d$ with a positive Lebesgue measure on which $\varphi_1(t) \not\preceq \varphi_2(t)$. Then $F \subseteq E$. We reach that $\gamma(F) > 0$ and $\Gamma_1(t) \not\preceq \Gamma_2(t)$ for $t \in F$, contradicting the fact that $\Gamma_1 \preceq \Gamma_2$ $\gamma-$a.e. ■

For K_s, G_s, we have the following result.

Proposition 17 *There holds $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$ if and only if*

(1) $\{t_j : j \in \mathbb{J}_1\} \subseteq \{t_k : k \in \mathbb{J}_2\}$;

(2) *for each $j \in \mathbb{J}_1$, $A_j \preceq B_j$. Here, re-indexing by condition (1) if necessary, we may assume that $\mathbb{J}_1 \subseteq \mathbb{J}_2$.*

Proof Introduce a discrete scalar-valued Borel measure γ that is supported on $\{t_j : j \in \mathbb{J}_1\} \cup \{t_k : k \in \mathbb{J}_2\}$ by setting

$$\gamma(\{t_k\}) := \begin{cases} \|A_k\|_{\mathcal{L}(\Lambda)} + \|B_k\|_{\mathcal{L}(\Lambda)}, & k \in \mathbb{J}_1 \cap \mathbb{J}_2, \\ \|B_k\|_{\mathcal{L}(\Lambda)}, & k \in \mathbb{J}_2 \setminus \mathbb{J}_1, \\ \|A_k\|_{\mathcal{L}(\Lambda)}, & k \in \mathbb{J}_1 \setminus \mathbb{J}_2. \end{cases}$$

We also let

$$\Gamma_A(t_j) := \frac{A_j}{\gamma(\{t_j\})}, \ \ j \in \mathbb{J}_1 \text{ and } \Gamma_A(t_k) := \frac{B_k}{\gamma(\{t_k\})}, \ \ k \in \mathbb{J}_2.$$

They are discrete $\mathcal{L}(\Lambda)$-valued functions supported on $\{t_j : j \in \mathbb{J}_1\}$ and $\{t_k : k \in \mathbb{J}_2\}$, respectively. We reach the following integral representation:

$$K_s(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} \Gamma_A(t) d\gamma(t) \text{ and } G_s(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} \Gamma_B(t) d\gamma(t), \ \ x,y \in \mathbb{R}^d.$$

By Theorem 11, $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$ if and only if $\Gamma_A \preceq \Gamma_B$ $\gamma-$a.e. It is straightforward to verify that the latter is equivalent to conditions (1)-(2). ■

5.2 Hessian of Scalar-valued Reproducing Kernels

Propositions 16 and 17 were established based on Theorem 11. In this subsection, we shall consider special translation invariant reproducing kernels and establish the characterization of refinement using Theorem 7.

Let k be a continuously differentiable translation invariant reproducing kernel on $\mathbb{R}^d$. We consider the following matrix-valued functions

$$K(x,y) := \nabla^2_{xy} k(x,y) := \left[\frac{\partial^2 k}{\partial_{x_j} \partial_{y_k}}(x,y) : j,k \in \mathbb{N}_d\right], \ \ x,y \in \mathbb{R}^d. \tag{44}$$

To ensure that K is an $\mathcal{L}(\mathbb{C}^d)$-valued reproducing kernels on $\mathbb{R}^d$, we make use of the Bochner theorem to get some finite nonnegative Borel measure μ on $\mathbb{R}^d$ such that

$$k(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\mu(t), \;\; x,y \in \mathbb{R}^d \tag{45}$$

and impose the requirement that

$$\int_{\mathbb{R}^d} tt^T d\mu(t) < +\infty. \tag{46}$$

One sees by the Lebesgue dominated convergence theorem that

$$K(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} tt^T d\mu(t), \;\; x,y \in \mathbb{R}^d, \tag{47}$$

where we view $t \in \mathbb{R}^d$ as a $d \times 1$ vector and t^T denotes its transpose $[t_1, t_2, \ldots, t_d]$. By the general integral representation (17) of operator-valued reproducing kernels, K defined by (44) is an $\mathcal{L}(\mathbb{C}^d)$-valued reproducing kernel on $\mathbb{R}^d$. Matrix-valued translation invariant reproducing kernels of the form (44) are useful for the development of divergence-free kernel methods for solving some special partial differential equations (see, for example, Lowitzsh, 2003; Wendland, 2009, and the references therein). Another class of kernels constructed from the Hessian of a scalar-valued translation invariant reproducing kernel is widely applied to the learning of a multivariate function together with its gradient simultaneously (Mukherjee and Wu, 2006; Mukherjee and Zhou, 2006; Ying and Campbell, 2008). Such applications make use of kernels of the form

$$\overline{K}(x,y) := \begin{bmatrix} k(x,y) & (\nabla_y k(x,y))^* \\ \nabla_x k(x,y) & \nabla^2_{xy} k(x,y) \end{bmatrix}. \tag{48}$$

One sees that under condition (46)

$$\overline{K}(x,y) = \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} \rho(t)\rho(t)^* d\mu(t), \;\; x,y \in \mathbb{R}^d,$$

where

$$\rho(t) = [1, it_1, it_2, \ldots, it_d]^T, \;\; t \in \mathbb{R}^d.$$

We aim at refining matrix-valued reproducing kernels of the forms (44) and (48) in this subsection. Specifically, we let ν be another finite nonnegative Borel measure on $\mathbb{R}^d$ satisfying

$$\int_{\mathbb{R}^d} tt^T d\nu(t) < +\infty \tag{49}$$

and define for $x, y \in \mathbb{R}^d$

$$g(x,y) := \int_{\mathbb{R}^d} e^{i(x-y)\cdot t} d\nu(t), \; G(x,y) := \nabla^2_{xy} g(x,y), \; \overline{G}(x,y) := \begin{bmatrix} g(x,y) & (\nabla_y g(x,y))^* \\ \nabla_x g(x,y) & \nabla^2_{xy} g(x,y) \end{bmatrix}. \tag{50}$$

Our purpose is to characterize $\mathcal{H}_K \preceq \mathcal{H}_G$ and $\mathcal{H}_{\overline{K}} \preceq \mathcal{H}_{\overline{G}}$ in terms of k, g and μ, ν.

Theorem 18 *Let μ, ν be finite nonnegative Borel measures on $\mathbb{R}^d$ satisfying (46) and (49), and k, g defined by (45) and (50). Then $K, G, \overline{K}, \overline{G}$ are matrix-valued translation invariant reproducing kernels on $\mathbb{R}^d$. The four relationships $\mathcal{H}_K \preceq \mathcal{H}_G$, $\mathcal{H}_{\overline{K}} \preceq \mathcal{H}_{\overline{G}}$, $\mathcal{H}_k \preceq \mathcal{H}_g$, and $\mu \preceq \nu$ are equivalent.*

Proof By Theorem 7 or a result in Xu and Zhang (2009), $\mathcal{H}_k \preceq \mathcal{H}_g$ if and only if $\mu \preceq \nu$. We shall show by Theorem 7 that $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mu \preceq \nu$. The equivalence of $\mathcal{H}_{\overline{K}} \preceq \mathcal{H}_{\overline{G}}$ and $\mu \preceq \nu$ can be proved similarly. Set

$$\phi(x,t) := e^{ix\cdot t}t^T, \ \ x,t \in \mathbb{R}^d.$$

Then for each $x,t \in \mathbb{R}^d$, $\phi(x,t)$ is a linear functional from $\mathbb{C}^d$ to $\mathbb{C}$. We observe by (47) that (17) holds true. So does (18). To apply Theorem 7, it remains to verify that $\text{span}\{\phi(x,\cdot)\xi : x \in \mathbb{R}^d, \ \xi \in \mathbb{C}^d\}$ is dense in the Hilbert space $L^2(\mathbb{R}^d, d\mu)$, which is straightforward. The claim follows immediately from Theorem 7. ■

5.3 Transformation Reproducing Kernels

Let us consider a particular class of matrix-valued reproducing kernels whose universality was studied in Caponnetto et al. (2008). The kernels we shall construct are from an input space X to output space $\Lambda = \mathbb{C}^n$, where $n \in \mathbb{N}$. To this end, we let k, g be two scalar-valued reproducing kernels on another input space Y and T_p be mappings from X to Y, $p \in \mathbb{N}_n$. Set

$$K(x,y) := [k(T_p x, T_q y) : p,q \in \mathbb{N}_n], \ \ G(x,y) := [g(T_p x, T_q y) : p,q \in \mathbb{N}_n], \ \ x,y \in X. \tag{51}$$

It is known that K, G defined above are indeed $\mathcal{L}(\mathbb{C}^n)$-valued reproducing kernels (Caponnetto et al., 2008). This also becomes clear in the proof below. We are interested in the conditions for $\mathcal{H}_K \preceq \mathcal{H}_G$ to hold.

Proposition 19 *Let K, G be defined by (51). Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mathcal{H}_{\overline{k}} \preceq \mathcal{H}_{\overline{g}}$, where $\overline{k}, \overline{g}$ are the restriction of k, g on $\cup_{p=1}^n T_p(X)$. In particular, if*

$$\bigcup_{p=1}^n T_p(X) = Y \tag{52}$$

then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mathcal{H}_k \preceq \mathcal{H}_g$.

Proof It is legitimate to assume that (52) holds true as otherwise, we may replace Y by $\cup_{p=1}^n T_p(X)$, and k, g by $\overline{k}, \overline{g}$, respectively.

Choose arbitrary feature maps and feature spaces $\Phi_1 : Y \to \mathcal{W}_1$ for k and $\Phi_2 : Y \to \mathcal{W}_2$ for g such that

$$\overline{\text{span}}\, \Phi_j(Y) = \mathcal{W}_j, \ \ j = 1,2. \tag{53}$$

By Proposition 6, $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$. We observe for all $x,y \in X$ and $\xi, \eta \in \mathbb{C}^n$ that

$$\begin{aligned} \tilde{K}((x,\xi),(y,\eta)) &= (K(x,y)\xi,\eta)_{\mathbb{C}^n} = \sum_{p=1}^n \sum_{q=1}^n \xi_p \overline{\eta_q} k(T_p x, T_q y) \\ &= \sum_{p=1}^n \sum_{q=1}^n \xi_p \overline{\eta_q} (\Phi_1(T_p x), \Phi_1(T_q y))_{\mathcal{W}_1} \\ &= \left(\sum_{p=1}^n \xi_p \Phi_1(T_p x), \sum_{q=1}^n \eta_q \Phi_1(T_q y) \right)_{\mathcal{W}_1}. \end{aligned}$$

Thus, $\tilde{\Phi}_1 : X \times \mathbb{C}^n \to \mathcal{W}_1$ defined by

$$\tilde{\Phi}_1(x,\xi) := \sum_{p=1}^{n} \xi_p \Phi_1(T_p x), \;\; x \in X,\, \xi \in \mathbb{C}^n$$

is a feature map for $\tilde{K}$. We next verify that $\text{span}\,\{\tilde{\Phi}_1(x,\xi) : x \in X,\, \xi \in \mathbb{C}^n\}$ is dense in $\mathcal{W}_1$. Assume that $u \in \mathcal{W}_1$ is orthogonal to this linear span, that is,

$$\left(u, \sum_{p=1}^{n} \xi_p \Phi_1(T_p x) \right)_{\mathcal{W}_1} = 0 \text{ for all } x \in X,\, \xi \in \mathbb{C}^n.$$

Then we have $(u, \Phi_1(T_p x))_{\mathcal{W}_1} = 0$ for all $x \in X$ and $p \in \mathbb{N}_n$. It follows from (52) and (53) that $u = 0$. Similar facts hold for G.

By Lemma 2, $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$ if and only if for every $u \in \mathcal{W}_1$, there exists $v \in \mathcal{W}_2$ such that

$$\left(u, \sum_{p=1}^{n} \xi_p \Phi_1(T_p x) \right)_{\mathcal{W}_1} = \left(v, \sum_{p=1}^{n} \xi_p \Phi_2(T_p x) \right)_{\mathcal{W}_2} \text{ for all } x \in X \tag{54}$$

and

$$\|u\|_{\mathcal{W}_1} = \|v\|_{\mathcal{W}_2}. \tag{55}$$

Recall also that $\mathcal{H}_k \preceq \mathcal{H}_g$ if and only if for all $u \in \mathcal{W}_1$ there exists some $v \in \mathcal{W}_2$ satisfying (55) and

$$(u, \Phi_1(y))_{\mathcal{W}_1} = (v, \Phi_2(y))_{\mathcal{W}_2} \text{ for all } y \in Y. \tag{56}$$

Clearly, (56) implies (54). Conversely, if (54) holds true then we get that

$$(u, \Phi_1(T_p x))_{\mathcal{W}_1} = (v, \Phi_2(T_p x))_{\mathcal{W}_2} \text{ for all } x \in X \text{ and } p \in \mathbb{N}_n,$$

which together with (52) implies (56). We conclude that $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$ if and only if $\mathcal{H}_k \preceq \mathcal{H}_g$. ■

A more general case of refinement of transformation reproducing kernels is discussed below. It can be proved by arguments similar to those for the previous proposition.

Proposition 20 *Let T_p, S_p be mappings from X to Y and k, g be scalar-valued reproducing kernels on Y. Define*

$$K(x,y) := [k(T_p x, T_q y) : p,q \in \mathbb{N}_n], \;\; G(x,y) := [g(S_p x, S_q y) : p,q \in \mathbb{N}_n], \;\; x,y \in X.$$

Suppose that for all $p \in \mathbb{N}_n$, $\text{span}\,\{k(T_p x, \cdot) : x \in X\}$ and $\text{span}\,\{g(S_p x, \cdot) : x \in X\}$ are dense in $\mathcal{H}_k$ and $\mathcal{H}_g$, respectively. Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mathcal{H}_{k_p} \preceq \mathcal{H}_{g_p}$ for all $p \in \mathbb{N}_n$, where

$$k_p(x,y) := k(T_p x, T_p y), \;\; g_p(x,y) := g(S_p x, S_p y), \quad x,y \in X.$$

5.4 Finite Hilbert-Schmidt Reproducing Kernels

We consider refinement of finite Hilbert-Schmidt reproducing kernels in this subsection. Let B_j, C_j be invertible operators in $\mathcal{L}_+(\Lambda)$, $n \le m \in \mathbb{N}$, and Ψ_j, $j \in \mathbb{N}_m$, be scalar-valued reproducing kernels on the input space X. Define

$$K(x,y) := \sum_{j=1}^{n} B_j \Psi_j(x,y), \;\; G(x,y) = \sum_{j=1}^{m} C_j \Psi_j(x,y), \quad x,y \in X. \tag{57}$$

By the general integral representation (20) and Proposition 8, K, G above are $\mathcal{L}(\Lambda)$-valued reproducing kernels on X. To ensure that representation (57) can not be further simplified, we shall work under the assumption that

$$\mathcal{H}_{\Psi_j} \cap \mathcal{H}_{\overline{\Psi}_j} = \{0\} \text{ for all } j \in \mathbb{N}_m, \tag{58}$$

where

$$\overline{\Psi}_j := \sum_{k \in \mathbb{N}_m \setminus \{j\}} \Psi_k.$$

Theorem 21 *Let K, G be defined by (57), where $B_j, C_j \in \mathcal{L}_+(\Lambda)$ are invertible and Ψ_j, $j \in \mathbb{N}_m$, are scalar-valued reproducing kernels on X satisfying (58). Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $B_j = C_j$, $j \in \mathbb{N}_n$.*

Proof We first find a feature map for $\tilde{K}$ and $\tilde{G}$. Let $\phi_j : X \to \mathcal{W}_j$ be an arbitrary feature map for Ψ_j such that $\operatorname{span} \phi_j(X)$ is dense in $\mathcal{W}_j$, and denote by $\Lambda \otimes \mathcal{W}_j$ the tensor product of Hilbert spaces Λ and $\mathcal{W}_j$, $j \in \mathbb{N}_m$. The space $\Lambda \otimes \mathcal{W}_j$ is a Hilbert space with the inner product

$$(\xi \otimes u, \eta \otimes v)_{\Lambda \otimes \mathcal{W}_j} := (\xi, \eta)_\Lambda (u,v)_{\mathcal{W}_j}, \;\; \xi, \eta \in \Lambda, \; u, v \in \mathcal{W}_j.$$

Set $\mathcal{W}$ the orthogonal direct sum of $\Lambda \otimes \mathcal{W}_j$, $j \in \mathbb{N}_n$, whose inner product is defined by

$$((\xi_j \otimes u_j : j \in \mathbb{N}_n), (\eta_j \otimes v_j : j \in \mathbb{N}_n))_{\mathcal{W}} := \sum_{j=1}^{n} (\xi_j, \eta_j)_\Lambda (u_j, v_j)_{\mathcal{W}_j}, \;\; \xi_j, \eta_j \in \Lambda, \; u_j, v_j \in \mathcal{W}_j, \;\; j \in \mathbb{N}_n.$$

We claim that $\Phi : X \times \Lambda \to \mathcal{W}$ defined by

$$\Phi(x,\xi) := (\sqrt{B_j}\xi \otimes \phi_j(x) : j \in \mathbb{N}_n), \;\; x \in X, \; \xi \in \Lambda$$

is a feature map for $\tilde{K}$. Here, $\sqrt{B_j}$, the square root of B_j, is the the unique operator A in $\mathcal{L}_+(\Lambda)$ such that $A^2 = B_j$. We verify for all $x, y \in X$ and $\xi, \eta \in \Lambda$ that

$$\begin{aligned} (\Phi(x,\xi), \Phi(y,\eta))_{\mathcal{W}} &= \sum_{j=1}^{n} (\sqrt{B_j}\xi, \sqrt{B_j}\eta)_\Lambda (\phi_j(x), \phi_j(y))_{\mathcal{W}_j} = \sum_{j=1}^{n} (B_j \xi, \eta)_\Lambda \Psi_j(x,y) \\ &= (K(x,y)\xi, \eta) = \tilde{K}((x,\xi),(y,\eta)). \end{aligned}$$

We next show that the denseness condition

$$\overline{\operatorname{span}}\{\Phi(x,\xi) : \; x \in X, \; \xi \in \Lambda\} = \mathcal{W} \tag{59}$$

is satisfied. To this end, suppose that we have $w_j \in \Lambda \otimes \mathcal{W}_j$, $j \in \mathbb{N}_n$ such that

$$((w_j : j \in \mathbb{N}_n), \Phi(x,\xi))_{\mathcal{W}} = \sum_{j=1}^{n} (w_j, \sqrt{B_j}\xi \otimes \phi_j(x))_{\Lambda\otimes\mathcal{W}_j} = 0 \text{ for all } x \in X \text{ and } \xi \in \Lambda. \tag{60}$$

Let $\{e_i : i \in \mathbb{I}\}$ and $\{f_k : k \in \mathbb{J}_j\}$ be an orthonormal basis for Λ and $\mathcal{W}_j$, respectively. Then $\{e_i \otimes f_k : i \in \mathbb{I}, k \in \mathbb{J}_j\}$ is an orthonormal basis for $\Lambda \otimes \mathcal{W}_j$. Note that although $\mathbb{I}$ or $\mathbb{J}_j$ might be uncountable, for each $\xi \in \Lambda$, $u \in \mathcal{W}_j$ and $w \in \Lambda \otimes \mathcal{W}_j$, the sets $\{i \in \mathbb{I} : (\xi, e_i)_\lambda \neq 0\}$, $\{k \in \mathbb{I}_j : (u, f_k)_{\mathcal{W}_j} \neq 0\}$ and $\{(i,j) \in \mathbb{I} \times \mathbb{J}_j : (w, e_i \otimes f_k)_{\Lambda\otimes\mathcal{W}_j} \neq 0\}$ are all countable. By resorting to these orthonormal bases, we see that

$$(w_j, \sqrt{B_j}\xi \otimes \phi_j(x))_{\Lambda\otimes\mathcal{W}_j} = \sum_{k\in\mathbb{J}_j} \sum_{i\in\mathbb{I}} (w_j, e_i \otimes f_k)_{\Lambda\otimes\mathcal{W}_j} (e_i, \sqrt{B_j}\xi)_\Lambda (f_k, \phi_j(x))_{\mathcal{W}_j}.$$

One verifies by the Cauchy-Schwartz inequality that

$$\sum_{k\in\mathbb{J}_j} \sum_{i\in\mathbb{I}} (w_j, e_i \otimes f_k)_{\Lambda\otimes\mathcal{W}_j} (e_i, \sqrt{B_j}\xi)_\Lambda f_k$$

converges in $\mathcal{W}_j$. As a consequence, $(w_j, \sqrt{B_j}\xi \otimes \phi_j(\cdot))_{\Lambda\otimes\mathcal{W}_j} \in \mathcal{H}_{\Psi_j}$. This together with (60) implies by the assumption (58) that

$$(w_j, \sqrt{B_j}\xi \otimes \phi_j(x))_{\Lambda\otimes\mathcal{W}_j} = 0 \text{ for all } j \in \mathbb{N}_n,\ x \in X \text{ and } \xi \in \Lambda.$$

The above equation can be equivalently formulated as

$$\left(\sum_{k\in\mathbb{J}_j} \sum_{i\in\mathbb{I}} (w_j, e_i \otimes f_k)_{\Lambda\otimes\mathcal{W}_j} (e_i, \sqrt{B_j}\xi)_\Lambda f_k, \phi_j(x) \right)_{\mathcal{W}_j} = 0$$

By the denseness of $\phi_j(X)$ in $\mathcal{W}_j$,

$$\sum_{i\in\mathbb{I}} (w_j, e_i \otimes f_k)_{\Lambda\otimes\mathcal{W}_j} (e_i, \sqrt{B_j}\xi)_\Lambda = 0 \text{ for all } j \in \mathbb{N}_n,\ k \in \mathbb{J}_j \text{ and } \xi \in \Lambda.$$

We thus have for all $j \in \mathbb{N}_n$ and $k \in \mathbb{J}_j$ that $\sum_{i\in\mathbb{I}} (w_j, e_i \otimes f_k)_{\Lambda\otimes\mathcal{W}_j} e_i = 0$, which implies

$$(w_j, e_i \otimes f_k)_{\Lambda\otimes\mathcal{W}_j} = 0 \text{ for all } j \in \mathbb{N}_n,\ k \in \mathbb{J}_j,\ i \in \mathbb{I}.$$

Therefore, $w_j = 0$ for all $j \in \mathbb{N}_n$. Equation (59) hence holds true. Similar facts hold for $\tilde{G}$.

By Proposition 6, $\mathcal{H}_K \preceq \mathcal{H}_G$ is equivalent to $\mathcal{H}_{\tilde{K}} \preceq \mathcal{H}_{\tilde{G}}$, which by the above discussion and Lemma 2 holds true if and only if for all $w_j \in \Lambda \otimes \mathcal{W}_j$, $j \in \mathbb{N}_n$ there exist unique $\tilde{w}_j \in \Lambda \otimes \mathcal{W}_j$, $j \in \mathbb{N}_m$ such that

$$\sum_{j=1}^{n} (w_j, \sqrt{B_j}\xi \otimes \phi_j(x))_{\Lambda\otimes\mathcal{W}_j} = \sum_{j=1}^{m} (\tilde{w}_j, \sqrt{C_j}\xi \otimes \phi_j(x))_{\Lambda\otimes\mathcal{W}_j} \text{ for all } \xi \in \Lambda \text{ and } x \in X \tag{61}$$

and

$$\sum_{j=1}^{n} (w_j, w_j)_{\Lambda\otimes\mathcal{W}_j} = \sum_{j=1}^{m} (\tilde{w}_j, \tilde{w}_j)_{\Lambda\otimes\mathcal{W}_j}. \tag{62}$$

Let $w_j \in \Lambda \otimes \mathcal{W}_j$, $j \in \mathbb{N}_n$ be given. If $B_j = C_j$ for $j \in \mathbb{N}_n$ then we set $\tilde{w}_j := w_j$ for $j \in \mathbb{N}_n$, and $\tilde{w}_j = 0$ for $n+1 \le j \le m$. Clearly, such a choice satisfies Equations (61) and (62). Therefore, $\mathcal{H}_K \preceq \mathcal{H}_G$. Conversely, suppose that $\mathcal{H}_K \preceq \mathcal{H}_G$. Then for the special choice $w_j := \xi_j \otimes u_j$, $\xi_j \in \Lambda$, $u_j \in \mathcal{W}_j$, $j \in \mathbb{N}_n$, there exists $\tilde{w}_j \in \Lambda \otimes \mathcal{W}_j$, $j \in \mathbb{N}_m$ satisfying (61) and (62). As $\tilde{w}_j$ is unique by the denseness of the feature map for $\tilde{G}$, we must have $w_j = (\sqrt{C_j}^{-1}\sqrt{B_j}\xi_j) \otimes u_j$ for $j \in \mathbb{N}_n$, and $\tilde{w}_j = 0$ for $n+1 \le j \le m$. This together with (62) yields that

$$\sum_{j=1}^{n} (\xi_j, \xi_j)_\Lambda (u_j, u_j)_{\mathcal{W}_j} = \sum_{j=1}^{n} (\sqrt{B_j} C_j^{-1} \sqrt{B_j} \xi_j, \xi_j)_\Lambda (u_j, u_j)_{\mathcal{W}_j}.$$

By successively making $\xi_j \otimes u_j \neq 0$ and $\xi_k \otimes u_k = 0$ for $k \in \mathbb{N}_n \setminus \{j\}$, for $j \in \mathbb{N}_n$, we reach that

$$(\xi_j, \xi_j)_\Lambda = (\sqrt{B_j} C_j^{-1} \sqrt{B_j} \xi_j, \xi_j)_\Lambda \text{ for all } \xi_j \in \Lambda \text{ and } j \in \mathbb{N}_n.$$

As $\sqrt{B_j} C_j^{-1} \sqrt{B_j}$ is hermitian, it equals the identity operator on Λ. It follows that $B_j = C_j$ for all $j \in \mathbb{N}_n$. The proof is complete. ■

As a corollary of Theorem 21, we obtain an orthogonal decomposition of $\mathcal{H}_K$.

Corollary 22 *Let K be defined by (57), where B_j are invertible and Ψ_j, $j \in \mathbb{N}_n$ satisfy (58). Then*

$$\mathcal{H}_K = \bigoplus_{j=1}^{n} \mathcal{H}_{B_j \Psi_j}$$

and

$$\mathcal{H}_{\sum_{j=1}^{k} B_j \Psi_j} \preceq \mathcal{H}_{\sum_{j=1}^{k+1} B_j \Psi_j} \text{ for } k \in \mathbb{N}_{n-1}.$$

A simplest case of (57) occurs when $\mathcal{H}_{\Psi_j}$ is of dimension 1 for $j \in \mathbb{N}_m$, which is covered below.

Corollary 23 *Let $B_j, C_k \in \mathcal{L}_+(\Lambda)$ be invertible for $j \in \mathbb{N}_n$ and $k \in \mathbb{N}_m$, and $\psi_k : X \to \mathbb{C}$, $k \in \mathbb{N}_m$, be linearly independent. Set*

$$K(x,y) := \sum_{j=1}^{n} B_j \psi_j(x) \overline{\psi_j(y)}, \;\; G(x,y) := \sum_{k=1}^{m} C_k \psi_k(x) \overline{\psi_k(y)}, \;\; x, y \in X.$$

Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $B_j = C_j$ for all $j \in \mathbb{N}_n$.

More generally, we might consider K, G defined by two distinct classes of linearly independent functions from X to $\mathbb{C}$. The result below can be proved using arguments similar to those for Theorem 21.

Proposition 24 *Let $n \le m \in \mathbb{N}_n$, $B_j, C_k \in \mathcal{L}_+(\Lambda)$ be invertible for $j \in \mathbb{N}_n$ and $k \in \mathbb{N}_m$, and $\{\psi_j : j \in \mathbb{N}_n\}$ and $\{\varphi_k : k \in \mathbb{N}_m\}$ be two classes of linearly independent functions from X to $\mathbb{C}$. Set*

$$K(x,y) := \sum_{j=1}^{n} B_j \psi_j(x) \overline{\psi_j(y)}, \;\; G(x,y) := \sum_{k=1}^{m} C_k \varphi_k(x) \overline{\varphi_k(y)}, \;\; x, y \in X.$$

Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if

(1) $\psi_j \in \operatorname{span}\{\varphi_k : k \in \mathbb{N}_m\}$ *for all* $j \in \mathbb{N}_n$*;*

(2) *the coefficients* $\lambda_{jl} \in \mathbb{C}$ *in the linear span*

$$\psi_j = \sum_{l=1}^{m} \lambda_{jl}\varphi_l, \quad j \in \mathbb{N}_n$$

satisfy

$$\sum_{l=1}^{m} \lambda_{jl}\lambda_{kl}C_l^{-1} = \delta_{j,k}B_j^{-1} \text{ for all } j,k \in \mathbb{N}_n.$$

We close this section with several concrete examples of finite Hilbert-Schmidt reproducing kernels of the form described in Corollary 23 and Proposition 24:

- polynomial kernels:

$$K(x,y) := \sum_{j=1}^{n} x^{\alpha_j} \cdot y^{\alpha_j} B_j, \quad x,y \in \mathbb{R}^d$$

  where α_j are multi-indices and B_j are invertible operators in $\mathcal{L}_+(\Lambda)$, or

$$K(x,y) := \sum_{j=1}^{n} (x \cdot y)^{\beta_j} B_j, \quad x,y \in \mathbb{R}^d$$

  where β_j are nonnegative integers.

- exponential kernels:

$$K(x,y) := \sum_{j=1}^{n} e^{i(x-y)\cdot t_j} B_j, \quad x,y \in \mathbb{R}^d$$

  where $t_j \in \mathbb{R}^d$.

6. Existence

This section is devoted to the existence of nontrivial refinement of operator-valued reproducing kernels. Most of the results to be presented here are straightforward extensions of those in the scalar-valued case (Xu and Zhang, 2009).

Let X be the input space and Λ be a Hilbert space. The reproducing kernels under consideration are $\mathcal{L}(\Lambda)$-valued.

Proposition 25 *There does not exist a nontrivial refinement of an* $\mathcal{L}(\Lambda)$*-valued reproducing kernel* K *on* X *if and only if* $\mathcal{H}_K = \Lambda^X$*, the set of all the functions from* X *to* Λ*. If the cardinality of* X *is infinite then every* $\mathcal{L}(\Lambda)$*-valued reproducing kernel on* X *has a nontrivial refinement.*

Surprisingly, nontrivial results about the existence appear when X is of finite cardinality.

Proposition 26 *Let* X *consist of finitely many points* x_j*,* $j \in \mathbb{N}_n$ *for some* $n \in \mathbb{N}_n$*. A necessary condition for an* $\mathcal{L}(\Lambda)$*-valued reproducing kernel on* X *to have no nontrivial refinements is that*

$$\sum_{j=1}^{n}\sum_{k=1}^{n} (K(x_j,x_k)\xi_j,\xi_k)_\Lambda > 0 \text{ for all } \xi_j \in \Lambda, j \in \mathbb{N}_n \text{ with } \sum_{j=1}^{n} \|\xi_j\|_\Lambda > 0. \tag{63}$$

A sufficient condition for K to have no nontrivial refinements is that

$$\sum_{j=1}^{n}\sum_{k=1}^{n}(K(x_j,x_k)\xi_j,\xi_k)_\Lambda \geq \lambda\sum_{j=1}^{n}\|\xi_j\|_\Lambda^2 \text{ for all } \xi_j\in\Lambda, j\in\mathbb{N}_n \tag{64}$$

for some constant $\lambda>0$. Consequently, if Λ is finite-dimensional then K does not have a nontrivial refinement if and only if (63) holds true.

Proof Suppose that there exist $\xi_j\in\Lambda$, $j\in\mathbb{N}_n$, at least one of which is nonzero, such that

$$\sum_{j=1}^{n}\sum_{k=1}^{n}(K(x_j,x_k)\xi_j,\xi_k)_\Lambda = 0.$$

This implies that

$$\sum_{j=1}^{n}K(x_j,\cdot)\xi_j = 0.$$

We get by (3) that for all $f\in\mathcal{H}_K$

$$\sum_{j=1}^{n}(f(x_j),\xi_j)_\Lambda = \left(f,\sum_{j=1}^{n}K(x_j,\cdot)\xi_j\right)_{\mathcal{H}_K} = 0.$$

As a consequence, $\mathcal{H}_K$ does not contain the function $f:X\to\Lambda$ taking values $f(x_j)=\xi_j$ for $j\in\mathbb{N}_n$. By Proposition 25, there exist nontrivial refinements for K on X.

Suppose that (64) holds true for some positive constant λ. Assume that $\mathcal{H}_K$ is a proper subset of Λ^X. Then there exists some nonzero vector $(\xi_k:k\in\mathbb{N}_n)\in\Lambda^n$ orthogonal to $(f(x_k):k\in\mathbb{N}_n)$ in Λ^n for all $f\in\mathcal{H}_K$. Letting $f=\sum_{j=1}^{n}K(x_j,\cdot)\xi_j$ yields that

$$\sum_{j=1}^{n}\sum_{k=1}^{n}(K(x_j,x_k)\xi_j,\xi_k)_\Lambda = \sum_{k=1}^{n}(f(x_k),\xi_k)_\Lambda = 0,$$

contradicting (64).

We complete the proof by pointing out that when Λ is finite-dimensional, (63) and (64) are equivalent. ■

It is worthwhile to note that when Λ is infinite-dimensional, condition (63) might not be sufficient for K to not have a nontrivial refinement. We give a concrete example to illustrate this.

Let X be a singleton $\{x\}$, $\Lambda:=\ell^2(\mathbb{N})$ consisting of square-summable sequences indexed by $\mathbb{N}$, and $K(x_1,x_1)$ be the operator T on $\ell^2(\mathbb{N})$ defined by

$$Ta := \left(\frac{a_j}{j}:j\in\mathbb{N}\right),\ \ a\in\ell^2(\mathbb{N}).$$

Apparently, $T\in\mathcal{L}_+(\ell^2(\mathbb{N}))$ and condition (63) is satisfied. Let $f\in\mathcal{H}_K$. Then there exist $a_n\in\ell^2(\mathbb{N})$, $n\in\mathbb{N}$ such that $K(x,\cdot)a_n$ converges to f in $\mathcal{H}_K$. Being a Cauchy sequence in $\mathcal{H}_K$, $\{K(x,\cdot)a_n:n\in\mathbb{N}\}$ satisfies

$$\lim_{n,m\to\infty}\|K(x,\cdot)a_n-K(x,\cdot)a_m\|_{\mathcal{H}_K}^2 = 0.$$

By (3),

$$\begin{aligned}\|K(x,\cdot)a_n - K(x,\cdot)a_m\|^2_{\mathcal{H}_K} &= (K(x,\cdot)(a_n-a_m), K(x,\cdot)(a_n-a_m))_{\mathcal{H}_K} \\ &= (K(x,x)(a_n-a_m), a_n-a_m)_{\ell^2(\mathbb{N})} = (T(a_n-a_m), a_n-a_m)_{\ell^2(\mathbb{N})} \\ &= \|\sqrt{T}a_n - \sqrt{T}a_m\|^2_{\ell^2(\mathbb{N})}.\end{aligned}$$

Combining the above two equations yields $\sqrt{T}a_n$ converges to some $b \in \ell^2(\mathbb{N}_n)$. We now have for each $c \in \ell^2(\mathbb{N})$ that

$$\begin{aligned}(f(x),c)_{\ell^2(\mathbb{N})} &= (f, K(x,\cdot)c)_{\mathcal{H}_K} = \lim_{n\to\infty}(K(x,\cdot)a_n, K(x,\cdot)c)_{\mathcal{H}_K} \\ &= \lim_{n\to\infty}(K(x,x)a_n, c)_{\ell^2(\mathbb{N})} = \lim_{n\to\infty}(Ta_n, c)_{\ell^2(\mathbb{N})} \\ &= \lim_{n\to\infty}(\sqrt{T}a_n, \sqrt{T}c)_{\ell^2(\mathbb{N})} = (b, \sqrt{T}c)_{\ell^2(\mathbb{N})} \\ &= (\sqrt{T}b, c)_{\ell^2(\mathbb{N})},\end{aligned}$$

which implies that $f(x) = \sqrt{T}b$. Since this is true for an arbitrary function $f \in \mathcal{H}_K$, the function $g : X \to \Lambda$ defined by

$$g(x) := \left(\frac{1}{j} : j \in \mathbb{N}\right)$$

is not in $\mathcal{H}_K$. Thus, K has a nontrivial refinement on X.

In the process of refining an operator-valued reproducing kernel, it is usually desirable to preserve favorable properties of the original kernel. We shall show that this is feasible as far as continuity and universality of operator-valued reproducing kernels are concerned. Let X be a metric space and K an $\mathcal{L}(\Lambda)$-valued reproducing kernel that is continuous from $X \times X$ to $\mathcal{L}(\Lambda)$ when the latter is equipped with the operator norm. Then one sees that $\mathcal{H}_K$ consists of continuous functions from X to Λ. For each compact subset $\mathcal{Z} \subseteq X$, denote by $\mathcal{C}(\mathcal{Z},\Lambda)$ the Banach space of all the continuous functions from $\mathcal{Z}$ to Λ with the norm

$$\|f\|_{\mathcal{C}(\mathcal{Z},\Lambda)} := \max_{x\in\mathcal{Z}} \|f(x)\|_\Lambda, \quad f \in \mathcal{C}(\mathcal{Z},\Lambda).$$

Following Micchelli et al. (2006) and Caponnetto et al. (2008), we call K a *universal kernel* on X if for all compact sets $\mathcal{Z} \subseteq X$ and all continuous functions $f : X \to \Lambda$ there exist

$$f_n \in \operatorname{span}\{K(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\}, \quad n \in \mathbb{N},$$

such that

$$\lim_{n\to\infty} \|f_n - f\|_{\mathcal{C}(\mathcal{Z},\Lambda)} = 0.$$

In other words, K is universal if for all compact subsets $\mathcal{Z} \subseteq X$, the closure of $\operatorname{span}\{K(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\}$ in $\mathcal{C}(\mathcal{Z},\Lambda)$ equals the whose space $\mathcal{C}(\mathcal{Z},\Lambda)$.

For the preservation of continuity, we have the following affirmative result, whose proof is similar to the scalar-valued case (Xu and Zhang, 2009).

Proposition 27 *Let X be a metric space with infinite cardinality. Then every continuous $\mathcal{L}(\Lambda)$-valued reproducing kernel on X has a nontrivial continuous refinement.*

The following lemma about universality has been proved in Caponnetto et al. (2008), and in Micchelli et al. (2006) in the scalar-valued case. We provide a simplified proof here.

Lemma 28 *Let K be a continuous $\mathcal{L}(\Lambda)$-valued reproducing kernel on X with the feature map representation (5), where $\Phi : X \to \mathcal{L}(\Lambda, \mathcal{W})$ is continuous. Then for each compact subset $\mathcal{Z} \subseteq X$,*

$$\overline{\text{span}}\{K(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\} = \overline{\{\Phi(\cdot)^* u : u \in \mathcal{W}\}},$$

where the closures are relative to the norm in $\mathcal{C}(\mathcal{Z},\Lambda)$.

Proof All the closures to appear in the proof are relative to the norm in $\mathcal{C}(\mathcal{Z},\Lambda)$. Let $K_{\mathcal{Z}}$ be the restriction of K on $\mathcal{Z}$. Then the restriction of Φ on $\mathcal{Z}$ remains a feature map for $K_{\mathcal{Z}}$. By Lemma 2,

$$\mathcal{H}_{K_{\mathcal{Z}}} = \{\Phi(\cdot)^* u : u \in \mathcal{W}\}. \tag{65}$$

It hence suffices to show that

$$\overline{\text{span}}\{K(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\} = \overline{\text{span}}\{K_{\mathcal{Z}}(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\} = \overline{\mathcal{H}_{K_{\mathcal{Z}}}}.$$

As $\text{span}\{K_{\mathcal{Z}}(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\} \subseteq \mathcal{H}_{K_{\mathcal{Z}}}$,

$$\overline{\text{span}}\{K_{\mathcal{Z}}(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\} \subseteq \overline{\mathcal{H}_{K_{\mathcal{Z}}}}. \tag{66}$$

On the other hand, for each $f \in \mathcal{H}_{K_{\mathcal{Z}}}$ there exist $f_n \in \text{span}\{K_{\mathcal{Z}}(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\}$, $n \in \mathbb{N}$ that converges to f in the norm of $\mathcal{H}_{K_{\mathcal{Z}}}$. It follows that f_n converges to f in the norm of $\mathcal{C}(\mathcal{Z},\Lambda)$. Therefore, $f \in \overline{\text{span}}\{K_{\mathcal{Z}}(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\}$, implying that

$$\overline{\mathcal{H}_{K_{\mathcal{Z}}}} \subseteq \overline{\text{span}}\{K_{\mathcal{Z}}(x,\cdot)\xi : x \in \mathcal{Z},\ \xi \in \Lambda\}. \tag{67}$$

Combining Equations (65), (66), and (67) proves the result. ■

The following positive result about universality can be proved by Lemma 28 and arguments similar to those used in Proposition 14 of Xu and Zhang (2009).

Proposition 29 *Let X be a metric space and K a continuous $\mathcal{L}(\Lambda)$-valued reproducing kernel on X. Then every continuous refinement of K on X remains universal.*

7. Numerical Experiments

We present in this final section three numerical experiments on the application of refinement of operator-valued reproducing kernels to multi-task learning. Suppose that f_0 is a function from the input space X to the output space Λ that we desire to learn from its finite sample data $\{(x_j,\xi_j) : j \in \mathbb{N}_m\} \subseteq X \times \Lambda$. Here m is the number of sampling points and

$$\xi_j = f_0(x_j) + \delta_j, \quad j \in \mathbb{N}_m$$

where $\delta_j \in \Lambda$ is the noise dominated by some unknown probability measure. To deal with the noise and have an acceptable generalization error, we use the following regularization network

$$\min_{f \in \mathcal{H}_K} \frac{1}{m} \sum_{j=1}^{m} \| f(x_j) - \xi_j \|_\Lambda^2 + \sigma \| f \|_{\mathcal{H}_K}^2, \tag{68}$$

where K is a chosen Λ-valued reproducing kernel on X. Our experiments will be designed so that underfitting and overfitting both have the chance to occur. To echo with the motivations in Section 2, when underfitting happens in the first experiment, we shall find a refinement G of K aiming at improving the performance of the minimizer of (68) in prediction. On the other hand, when overfitting appears in the second experiment, we shall then find a Λ-valued reproducing kernel L on X such that $\mathcal{H}_L \preceq \mathcal{H}_K$ with the same purpose.

Before moving on to the experiments, we make a remark on how (68) can be solved. The issue has been understood in the work by Micchelli and Pontil (2005). We say that K is *strictly positive-definite* if for all finite $y_j \in X$, $j \in \mathbb{N}_p$, and for all $\eta_j \in \Lambda$, $j \in \mathbb{N}_p$ all of which are not zero

$$\sum_{j=1}^{p}\sum_{k=1}^{p}(K(y_j,y_k)\eta_j,\eta_k)_\Lambda > 0.$$

If K is strictly positive-definite then the minimizer f_K of (68) has the form

$$f_K = \sum_{j=1}^{m} K(x_j,\cdot)\eta_j \tag{69}$$

where η_j's satisfy

$$\sum_{k=1}^{m} K(x_k,x_j)\eta_k + m\sigma\eta_j = \xi_j, \quad j \in \mathbb{N}_m. \tag{70}$$

7.1 Experiment 1: Underfitting

The vector-valued function to be learned from finite examples is from the input space $X = [-1,1]$ to output space $\Lambda = \mathbb{R}^n$, where $n \in \mathbb{N}$. Specifically, it has the form

$$f_0(x) := \left[a_k|x-b_k| + c_k e^{-d_k x} : k \in \mathbb{N}_n\right], \quad x \in [-1,1], \tag{71}$$

where a,b,c,d are constant vectors to be randomly generated. The $\mathcal{L}_+(\mathbb{R}^n)$-valued reproducing kernel that we shall use in the regularization network (68) is a Gaussian kernel

$$K(x,y) := S\exp\left(-\frac{(x-y)^2}{2}\right), \quad x,y \in \mathbb{R},$$

where $S \in \mathcal{L}_+(\mathbb{R}^n)$ is strictly positive-definite. It can be identified by Lemma 2 that functions in $\mathcal{H}_K$ are of the form $\sqrt{S}v$, where v is an $\mathbb{R}^n$-valued function whose components come from the RKHS $\mathcal{H}_\mathcal{G}$ of the scalar-valued Gaussian kernel

$$\mathcal{G}(x,y) := \exp\left(-\frac{(x-y)^2}{2}\right), \quad x,y \in \mathbb{R}. \tag{72}$$

Thus, each component of $\sqrt{S}v$ is from $\mathcal{H}_\mathcal{G}$. The function f_0 to be approximated is defined by (71). As $|x-b_k|$ is not even continuously differentiable, functions from the RKHS of the Gaussian kernel (72) with a fixed variance may not well approximate f_0. Underfitting is hence expected. If this is indeed observed then a remedy is to use the refinement of K given by

$$G(x,y) := S\exp\left(-\frac{(x-y)^2}{2}\right) + T(1+xy)^3, \quad x,y \in \mathbb{R},$$

where $T \in \mathcal{L}_+(\mathbb{R}^n)$ is also strictly positive-definite. The RKHS of the scalar-valued polynomial kernel $(1+xy)^3$ clearly does not have a nontrivial intersection with the RKHS of the scalar-valued Gaussian kernel. Thus, by Corollary 22, $\mathcal{H}_K \preceq \mathcal{H}_G$, namely, G is a nontrivial refinement of K. Furthermore, as low order polynomials are added, the ability for functions in $\mathcal{H}_G$ to approximate the function $|x-b_k|$ is expected to be superior to those in $\mathcal{H}_K$. We perform extensive numerical simulations to confirm these conjectures.

The dimension n will be chosen from $\{2,4,8,16\}$. The number m of sampling points will be set to be 30. The sampling points x_j, $j \in \mathbb{N}_m$ will be randomly sampled from $[-1,1]$ by the uniform distribution and the outputs ξ_j are generated by

$$\xi_j = f_0(x_j) + \delta_j, \quad j \in \mathbb{N}_m, \tag{73}$$

where δ_j are vectors whose components will be randomly generated by the uniform distribution on $[-\delta,\delta]$ with δ being the noise level selected from $\{0.1,0.3,0.5\}$. For each dimension $n \in \{2,4,8,16\}$ and noise level $\delta \in \{0.1,0.3,0.5\}$, we run 50 simulations. In each of the simulations, we do the followings:

1. the components of the coefficient vectors a,b,c,d in the function f_0 given by (71) are randomly generated by the uniform distribution on $[1,3]$, $[-1,1]$, $[-2,2]$, and $[0,3]$, respectively;

2. the sampling points are randomly sampled from $[-1,1]$ by the uniform distribution and the outputs ξ_j are then generated by (73);

3. the matrices S and T are given by $S = A'A$ and $T = B'B$ where A,B are $n \times n$ real matrices whose components are randomly sampled from $[1,3]$ by the uniform distribution;

4. we then solve the minimizer f_K of (68) by (69) and (70);

5. for the refinement kernel G, we also obtain f_G as the minimizer of

$$\min_{f \in \mathcal{H}_G} \frac{1}{m} \sum_{j=1}^{m} \| f(x_j) - \xi_j \|_\Lambda^2 + \sigma \| f \|_{\mathcal{H}_G}^2, \tag{74}$$

6. the regularization parameters in (68) and (74) are optimally chosen so that the relative square approximation errors

$$\mathcal{E}_K := \frac{\int_{-1}^1 \| f_K(t) - f_0(t) \|^2 dt}{\int_{-1}^1 \| f_0(t) \|^2 dt}, \quad \mathcal{E}_G := \frac{\int_{-1}^1 \| f_G(t) - f_0(t) \|^2 dt}{\int_{-1}^1 \| f_0(t) \|^2 dt}. \tag{75}$$

   are minimized, respectively.

We call $(\mathcal{E}_K, \mathcal{E}_G)$ obtained in each simulation an instance of approximation errors. Hence, we have 50 instances for each pair of (n,δ). They are said to form a group. There are 12 groups of instances of approximation errors. For each (n,δ), we shall calculate the mean and standard deviation of the difference $\mathcal{E}_K - \mathcal{E}_G$ in the corresponding group as a measurement of the difference in the performance of learning schemes (68) and (74). Before that, outliers of instances should be excluded. Although we do not know the distributions of $\mathcal{E}_K$ and $\mathcal{E}_G$, we shall use the three-sigma rule in statistics. In other words, we regard an instance $(\mathcal{E}_K, \mathcal{E}_G)$ as an outlier if the deviation of $\mathcal{E}_K$

	n=2	n=4	n=8	n=16
$\delta = 0.1$	(0.1024,0.0084) (0.0091,0.0081) (0.4128,0.0006) (0.6783,0.0025)	(0.0215,0.0182) (0.4095,0.0034)	(0.0230,0.0070) (0.0513,0.0091) (0.1554,0.0011) (0.1464,0.0026)	(0.0712,0.0015) (0.0364,0.0124)
$\delta = 0.3$	(0.0286,0.0228) (0.4811,0.0020)	(0.0663,0.0321) (0.1892,0.0041) (0.1674,0.0095)	(0.0407,0.0194) (0.1809,0.0023)	(0.1592,0.0018) (0.0309,0.0127) (0.0229,0.0099)
$\delta = 0.5$	(0.2053,0.0020) (0.1267,0.0034) (0.0669,0.0465)	(0.0377,0.0376) (0.3547,0.0033)	(0.2445,0.0028) (0.2762,0.0020) (0.0119,0.0264)	(0.1612,0.0043) (0.0541,0.0081)

Table 1: Outliers of instances of approximation errors $(\mathcal{E}_K, \mathcal{E}_G)$. An instance $(\mathcal{E}_K, \mathcal{E}_L)$ is considered to be an outlier if the deviation of one of its components to the respective mean in the group is more than three times the standard deviation of the group. Outliers are listed in an independent table because they should be excluded from the calculation of the mean and standard deviation of the approximation errors. Another reason is that adding them will make the plot of the approximation errors highly disproportional.

or $\mathcal{E}_G$ to their respective mean in the group exceeds three times their respective standard deviation. There are 32 outliers among the entire 600 instances, which are listed below in Table 1.

We make a few observations from Table 1. Firstly, $\mathcal{E}_G$ is smaller than $\mathcal{E}_K$ except for only one instance. For a large portion of the outliers, the approximation error $\mathcal{E}_K$ is considerably large (larger than 10%), a sign of underfitting of the kernel K. Those instances are of the greatest interest to us as we desire to see if the refinement kernel G can make a remedy when underfitting does happen. We see from Table 1 that for all of those outliers, the refinement kernel G always brings down the relative approximation error to be less than 1%. The improvement brought by G for other instances is also significant. The observations indicate that (74) performs significantly better in learning the function (71) from finite examples than (68). For further comparison, we compute the mean and standard deviation of the difference $\mathcal{E}_K - \mathcal{E}_G$ of the approximation errors after excluding the above outliers. The results are tabulated in Table 2 below. Note that a positive value of the mean implies that (74) performs better than (68). It is worthwhile to point out that among all the rest 568 instances excluding the outliers, there are only 33 where $\mathcal{E}_G$ is larger than $\mathcal{E}_K$. The largest value of $\mathcal{E}_G - \mathcal{E}_K$ is 0.0020. Therefore, we conclude that for all the (n, δ), (74) is superior to (68), and the larger the standard deviation in Table 2 is, the greater improvement the refinement kernel G brings.

We shall also plot the 12 groups of approximation errors $\mathcal{E}_K, \mathcal{E}_G$ for a visual comparison. To this end, we take out the instances for which $\mathcal{E}_K$ is too large to have an appropriate range in the vertical axes in the figures. Therefore, Figures 1 and 2 are not full embodiment of the improvement of (74) over (68). Nevertheless, one sees that the improvement brought by the refinement kernel G in these relatively well-behaved instances is still dramatic.

	n=2	n=4	n=8	n=16
$\delta = 0.1$	0.0098 (0.0182)	0.0139 (0.0335)	0.0160 (0.0241)	0.0108 (0.0135)
$\delta = 0.3$	0.0076 (0.0144)	0.0141 (0.0245)	0.0143 (0.0208)	0.0188 (0.0259)
$\delta = 0.5$	0.0054 (0.0121)	0.0127 (0.0307)	0.0103 (0.0186)	0.0091 (0.0102)

Table 2: The mean and standard deviation (in parentheses) of $\mathcal{E}_K - \mathcal{E}_G$. The outliers of instances listed in Table 1 are not counted toward these calculations. If they were added, the improvement brought by the refinement kernel G would have been more dramatic.

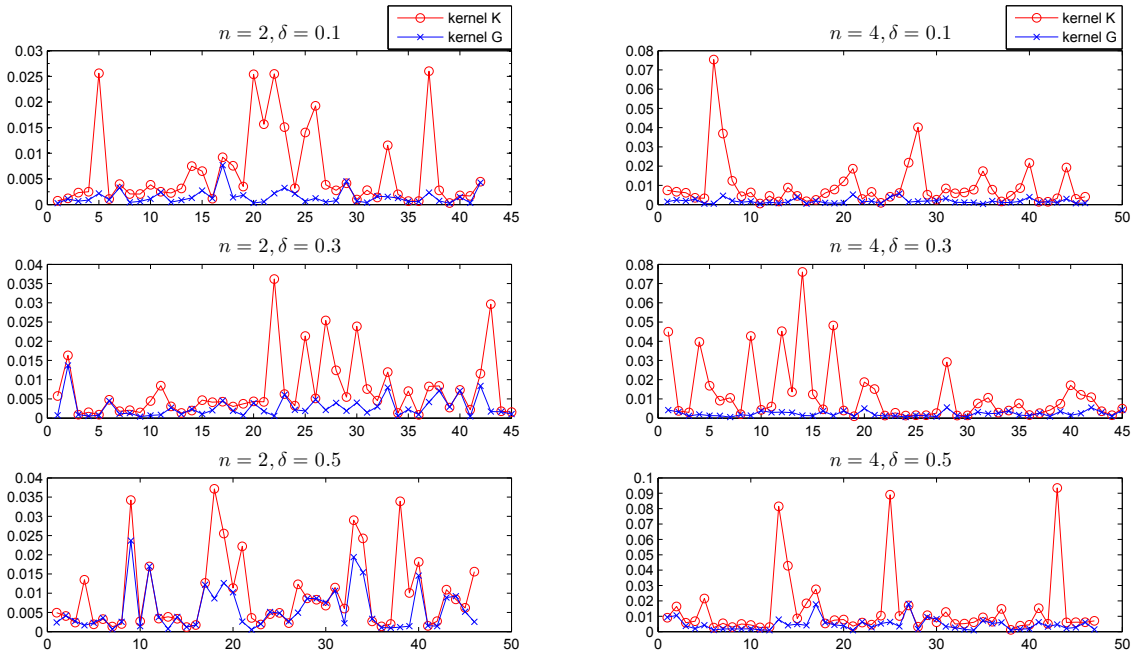

Figure 1: Relative approximation errors $\mathcal{E}_K, \mathcal{E}_G$ for $n = 2, 4$ and $\delta = 0.1, 0.3, 0.5$. The outliers listed in Table 1 are not plotted here as they would make the figure highly disproportional.

7.2 Experiment 2: Overfitting

The target function we consider in the second experiment is

$$f_0(x) = \left[\frac{a_k}{1 + 25(x - b_k)^2} + c_k e^{-d_k x} : k \in \mathbb{N}_n \right], \quad x \in [-1, 1], \tag{76}$$

where the components of the vectors $a, b, c, d \in \mathbb{R}^n$ will be randomly sampled by the uniform distribution from $[1, 4]$, $[0, \frac{1}{2}]$, $[-2, 2]$, and $[0, 2]$ respectively in the numerical simulations. The dimension

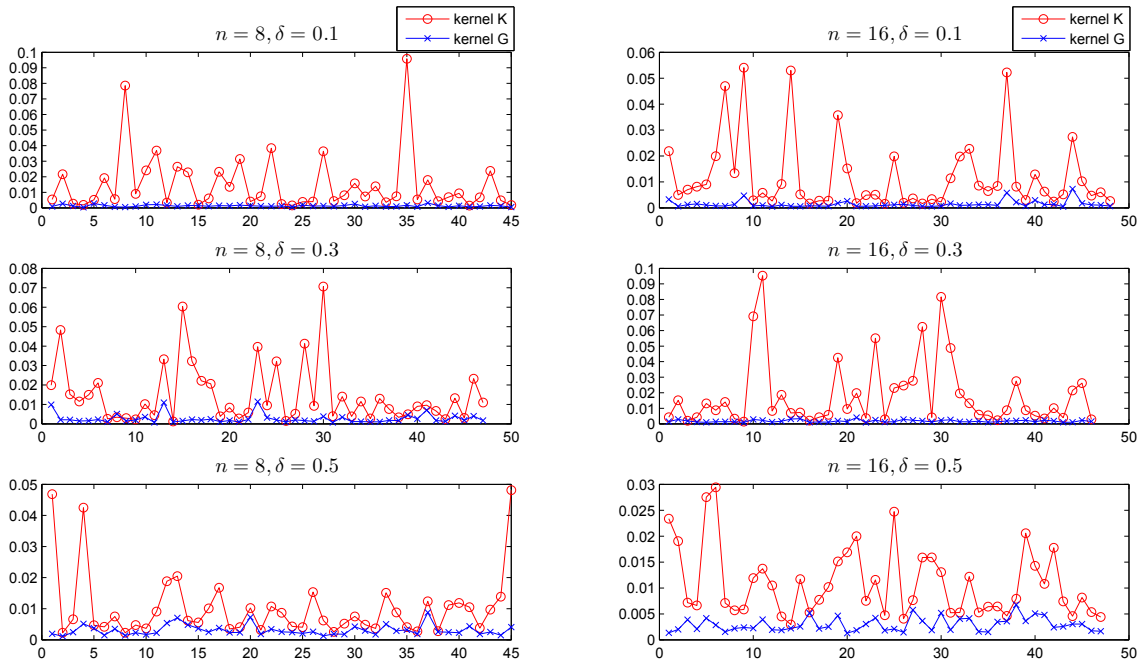

Figure 2: Relative approximation errors $\mathcal{E}_K, \mathcal{E}_G$ for $n = 8, 16$ and $\delta = 0.1, 0.3, 0.5$. The outliers listed in Table 1 are not plotted in the figure here.

n will be chosen from $\{2,4,8,16\}$. We fix $m := 20$ and shall sample the inputs x_j, $j \in \mathbb{N}_m$ randomly by the uniform distribution from $[-1,1]$. Similarly, the outputs $\xi_j \in \mathbb{R}^n$, $j \in \mathbb{N}_m$ will be generated by (73) where the noise level is to be selected from $\{0.1, 0.3, 0.5\}$.

In the first step, we substitute the sample data $\{(x_j, \xi_j) : j \in \mathbb{N}_m\}$ into the regularization network (68) with the following kernel

$$K(x,y) := S \exp\left(-\frac{(x-y)^2}{2}\right) + T(1+xy)^{18}, \quad x,y \in [-1,1], \tag{77}$$

where $S = A'A$ and $T = B'B$ with A, B being $n \times n$ real-matrices whose components will be randomly sampled by the uniform distribution from $[1,2]$. The target function (76) contains translations of the Runge function

$$\frac{1}{1+25x^2}, \quad x \in [-1,1].$$

It is well-known that approximating the Runge function by high order polynomial interpolations leads to overfitting. One sees by (70) that the regulation network (68) might be regarded as a regularized interpolation. Note also that the order of the polynomial kernel in (77) is 18, which is close to the number $m = 20$ of sampling points. Overfitting is hence expected. When this occurs, we propose to reduce the order of the polynomial kernel by considering

$$L(x,y) := S \exp\left(-\frac{(x-y)^2}{2}\right) + T \sum_{k=0}^{10} \binom{18}{k} (xy)^k, \quad x,y \in [-1,1].$$

	$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.5$
n=2	(0.9000, 0.7843)	(2.9906,1.3509)	(1.8065, 0.8044),(1.1332, 0.3213) (19.6416, 7.6578)
n=4	(8.2450,5.8717) (1.6654,2.0466) (18.9615,12.0513) (0.9536, 1.0998)	(1.1760,0.1354) (0.4591, 0.7845)	(4.6316,7.0497),(2.0850,1.3204) (2.4657,1.1386) (5.7967,0.6122) (5.1196,2.6692)
n=8	(0.9102,1.3862) (1.2233,0.9489) (0.6711,0.2249)	(1.3517, 1.8339) (0.8450,0.2605) (0.3571, 0.7221) (2.2403, 2.0108) (5.6153,5.0954) (2.0763,1.3718) (2.2567,1.4024)	(0.6369, 0.3698),(0.6945,0.2878) (2.2371, 2.4008) (1.0738,0.4172) (1.0561,0.3067) (0.6791,1.0980) (3.6689,3.9566) (1.1238,0.2467)
n=16	(4.4905, 5.8886) (7.9187, 4.3445) (2.1619, 0.5061) (17.5145, 13.7894)	(26.0758,7.6125) (1.2255, 0.3181) (0.5140, 0.1817) (2.4289, 1.9022)	(73.0854,42.6904),(1.6070, 1.4224) (3.2674, 2.2622),(2.1632, 1.7059) (2.8067, 0.5791),(9.0120, 3.5443) (0.6064, 0.3365),(4.0484 , 0.4220) (1.0064, 0.8287)

Table 3: Outliers of instances of relative approximation errors $(\mathcal{E}_K, \mathcal{E}_L)$.

By Corollary 22, $\mathcal{H}_L \preceq \mathcal{H}_K$, namely, K is a refinement of L. We shall demonstrate by numerical simulations that

$$\min_{f\in\mathcal{H}_L} \frac{1}{m}\sum_{j=1}^{m} \|f(x_j)-\xi_j\|^2 + \sigma\|f\|^2_{\mathcal{H}_L} \tag{78}$$

outperforms (68) with the kernel (77). To this end, we shall conduct numerical experiments similar to those in the last subsection. Let f_K and f_L be the minimizer of (68) and (78), respectively. We shall measure the performance by the relative square approximation errors $\mathcal{E}_K$ and $\mathcal{E}_L$, which are defined in the same way as (75). For each pair of (n,δ), where $n \in \{2,4,8,16\}$ and $\delta \in \{0.1,0.3,0.5\}$, we run 20 numerical simulations where the regularization parameters σ are to be chosen so that $\mathcal{E}_K$ and $\mathcal{E}_L$ are minimized, respectively. As in the first experiment, we shall calculate the mean and standard deviation of $\mathcal{E}_K$ and $\mathcal{E}_L$ in each group after taking out some outliers. We shall also plot the relative errors for comparison. The results are shown below in the form of tables and figures.

We have more outliers compared to the first experiment. Using fewer sampling points and approximating the Runge function by polynomials both contributes to this. We observe that for the majority of these outliers, $\mathcal{E}_L$ is significantly smaller than $\mathcal{E}_K$, showing improvement of learning scheme (78) over (68). For further comparison, we shall compute the mean and variances of $\mathcal{E}_K - \mathcal{E}_L$ and plot the relative approximation errors $\mathcal{E}_K$ and $\mathcal{E}_L$ for the rest of instances.

A positive value of the mean in Table 4 implies that (78) performs better than (68). It is observed that kernel L brings improvement for all the choices of $n \in \{2,4,8,16\}$ and $\delta \in \{0.1,0.3,0.5\}$. We also remark that among all the 188 instances counted in Table 4, there are only 32 for which $\mathcal{E}_L > \mathcal{E}_K$. The mean and standard deviation of $\mathcal{E}_L - \mathcal{E}_K$ for these 32 instances are 0.0264 and

	n=2	n=4	n=8	n=16
$\delta = 0.1$	0.0289 (0.0846)	0.0511 (0.0587)	0.0173 (0.0779)	0.0157 (0.0146)
$\delta = 0.3$	0.0404 (0.0922)	0.0661 (0.0705)	0.0671 (0.0929)	0.0657 (0.0918)
$\delta = 0.5$	0.0629 (0.1098)	0.0130 (0.0233)	0.0484 (0.0758)	0.0625 (0.0821)

Table 4: The mean and standard deviation (in parentheses) of $\mathcal{E}_K - \mathcal{E}_L$. The outliers of instances listed in Table 3 are not counted toward these calculations. If they were added, the improvement brought by the refinement kernel G would have been more dramatic.

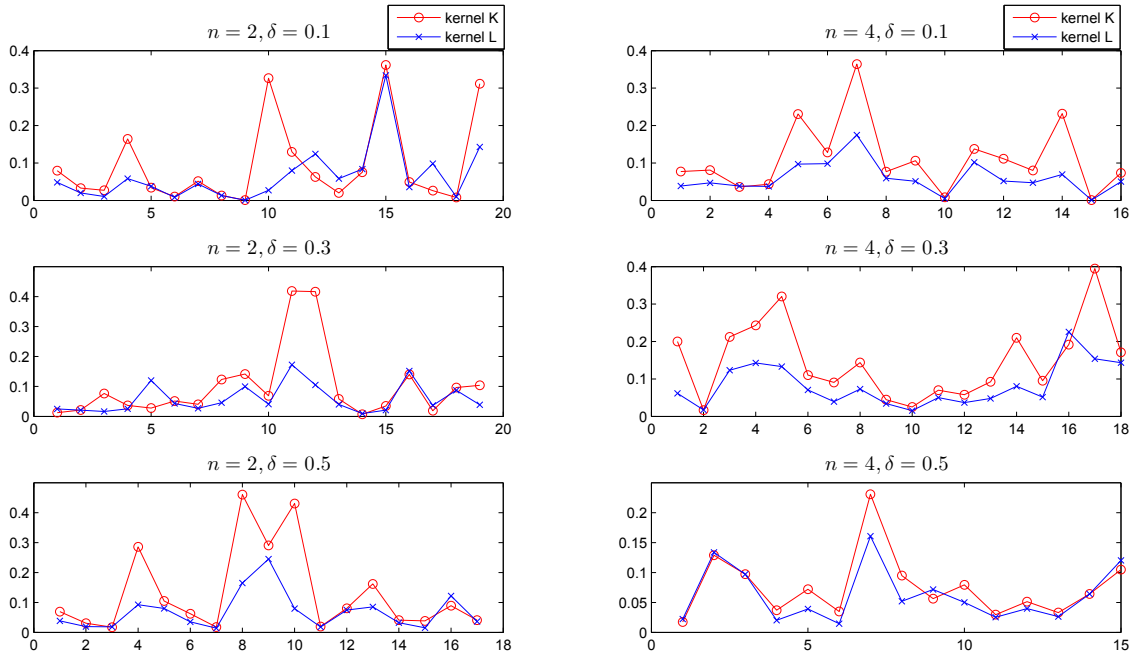

Figure 3: Relative approximation errors $\mathcal{E}_K, \mathcal{E}_L$ for $n = 2, 4$ and $\delta = 0.1, 0.3, 0.5$. The outliers listed in Table 3 are not plotted here as they will make the figure highly disproportional.

0.0306. We conclude that compared to (68), (78) improves the performance considerably in learning the function (76).

7.3 Experiment 3: Impact of Irrelevant Signals

Suggested by one of the anonymous reviewers, we shall examine the impact of irrelevant signals in the refinement kernel method. More specifically, we plan to apply the refinement kernel method

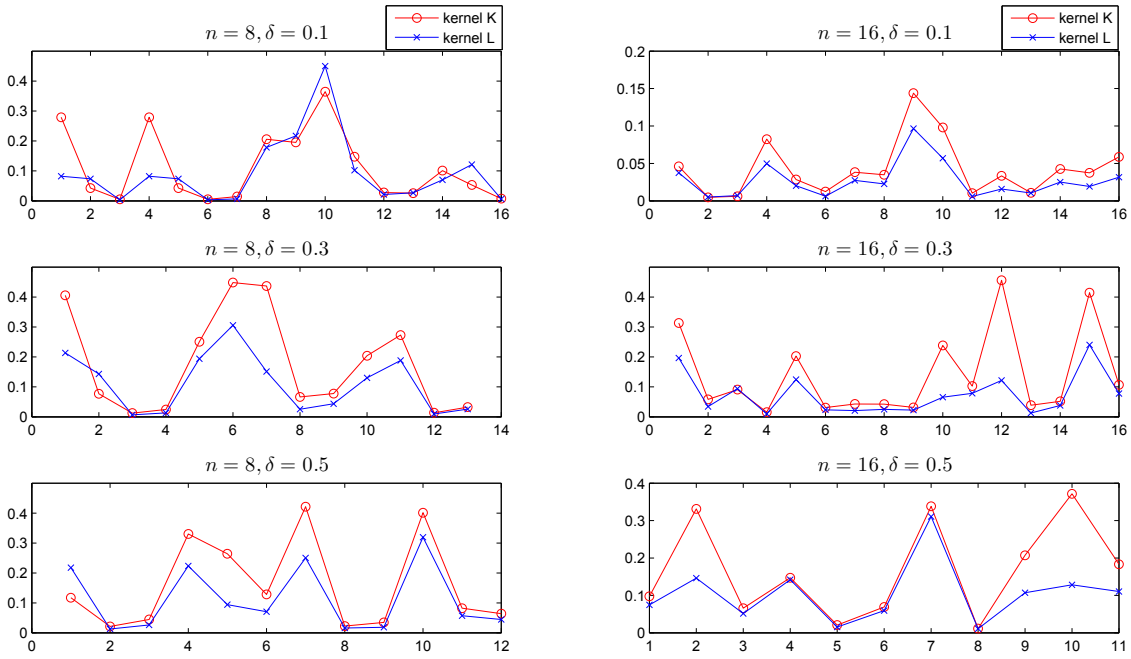

Figure 4: Relative approximation errors $\mathcal{E}_K, \mathcal{E}_L$ for $n = 8, 16$ and $\delta = 0.1, 0.3, 0.5$. The outliers listed in Table 3 are not plotted here.

	$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.5$
$n = 4$	(0.0164, 0.0083) (0.2930, 0.0044) (0.0074, 0.0076)	(0.1760,0.0074) (0.0415, 0.0189) (0.1464, 0.0254)	(0.1550, 0.0049) (0.0837, 0.0302)

Table 5: Outliers of instances of relative approximation errors $(\mathcal{E}_K, \mathcal{E}_G)$.

to the learning a vector-valued function whose components might be irrelevant. To avoid repetition and save space, we shall consider the underfitting case only and limit ourself to dimension $n = 4$. The instance investigated here is the function f_0 of the form (71), where we shall set $a_3 = a_4 = c_1 = c_2 = 0$. Thus, the first two components are irrelevant with the last two components of f_0. We then proceed with the same simulation procedures as those in experiment 1.

We obtain 3 groups of relative approximation error $(\mathcal{E}_K, \mathcal{E}_G)$ corresponding to the noise level $\delta = 0.1, 0.3, 0.5$. As in experiment 1, we first list all the outliers by the three-sigma rule in Table 5 below.

We observe from Table 5 that under the impact of irrelevant signals, among the above outliers, $\mathcal{E}_G$ is smaller than $\mathcal{E}_K$ except for only one instance $(0.0074, 0.0076)$. In 4 instances of the outliers, $\mathcal{E}_K$ is larger than 14%, while the refinement kernel G always brings down the relative approximation error to be less than 3%. In the overall 150 instances of relative approximation errors computed,

	$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.5$
$n = 4$	0.0077 (0.0131)	0.0114 (0.0257)	0.0117 (0.0205)

Table 6: The mean and standard deviation (in parentheses) of $\mathcal{E}_K - \mathcal{E}_G$. The outliers of instances listed in Table 5 are not counted toward these calculations. If they were added, the improvement brought by the refinement kernel G would have been more dramatic.

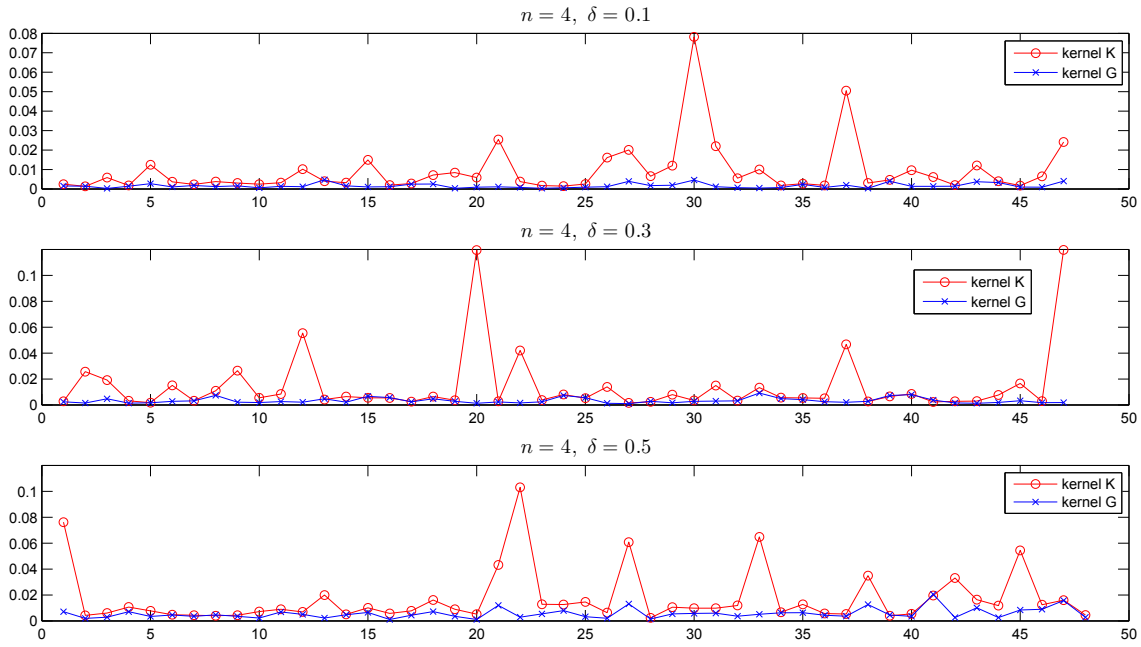

Figure 5: Relative approximation errors $\mathcal{E}_K, \mathcal{E}_G$ for $n = 4$ and $\delta = 0.1, 0.3, 0.5$. The outliers listed in Table 5 are not plotted here as they would make the figure highly disproportional.

there are only 13 instances where $\mathcal{E}_K$ is smaller than $\mathcal{E}_G$. For all these instances, $\mathcal{E}_G$ are of the same magnitude level with $\mathcal{E}_K$, showing competitive performance. For further comparison, we compute the mean and standard deviation of the difference $\mathcal{E}_K - \mathcal{E}_G$ after the above outliers are excluded. The results are shown in Table 6 below.

Finally, we plot the 3 groups of relative approximation errors $\mathcal{E}_K, \mathcal{E}_G$ for a visual comparison after the outliers in Table 5 are excluded.

We conclude from Tables 5, 6 and Figure 5 that for the learning problem considered in this subsection, the refinement kernel method works well under the impact of irrelevant signals.

8. Conclusion and Discussion

The refinement relationship between two operator-valued reproducing kernels provides a promising way of updating kernels for multi-task machine learning when overfitting or underfitting occurs. We establish several general characterizations of the refinement relationship. Particular attention has been paid to the case when the kernels under investigation have a vector-valued integral representation, the most general form of operator-valued reproducing kernels. By the characterizations, we present concrete examples of refining the translation invariant operator-valued reproducing kernels, Hessian of the scalar-valued Gaussian kernel, and finite Hilbert-Schmidt operator-valued reproducing kernels. Three numerical experiments confirm the potential usefulness of the proposed refinement method in updating kernels for multi-task learning. We plan to investigate the effect of the method by real application data in another occasion.

We discuss three issues that might deserve future research attention. The first one concerns about the computational saving brought by the refinement kernel method. Suppose a minimizer in an RKHS resulting from a particular learning algorithm is already computed but turns out to be unsatisfactory due to underfitting. When the kernel corresponding to the RKHS is refined, instead of running the algorithm from the scratch in the updated RKHS, we are wondering if the original minimizer can be made use of in order to reduce computational costs. In the scalar-valued case, it has been shown that this can be done for the classical regularization networks (Xu and Zhang, 2009). For the vector-valued case, one would need to carefully handle the complexity brought by the high dimension of the output space in order to establish a similar analysis. The second question is whether a multi-resolution analysis for vector-valued RKHS can be achieved by using the refinement kernel method. Our initial thinking and impression is that the approach in Xu and Zhang (2007) of using a bijective self-mapping of the input space can be carried over without much difficulty. Finally, we look at the requirement in the definition of refinement that the norm on the RKHS of the refinement kernel should coincide with that in the RKHS of the original kernel. As seen by the results in Section 5 and those in Xu and Zhang (2009), this strong condition poses a serious restriction in searching for refinement kernels. A remedy is to ask the two norms to be equivalent in the smaller space or to even just focus on the inclusion relation. Study along this direction has been done for scalar-valued kernels (Zhang and Zhao, 2011). It is shown there that this relaxation brings more freedom and choices in choosing kernels for refinement. Vector-valued counterparts are yet to be investigated. This approach also connects to a popular way of updating kernels pointed out by one of the reviewers, which is to tune a parameter (for example, the variance in the Gaussian kernel, the degree in a polynomial kernel, etc.) in the kernel. Although this practice seldom corresponds to a refinement, it does sometimes fall into the approach considered in Zhang and Zhao (2011). Examples include the exponential kernels, the inverse multiquadrics, the B-spline kernels, and the polynomial kernels (Zhang and Zhao, 2011).

Acknowledgments

The authors would like to express their appreciation to the anonymous reviewers for many useful comments and suggestions that help improve the paper. The project was supported by Guangdong Provincial Government of China through the “Computational Science Innovative Research Team” program. Haizhang Zhang was partially supported by Natural Science Foundation of China under grants 11101438 and 91130009, and by the One Hundred Distinguished Scholars Program of Sun

Yat-sen University. Yuesheng Xu was partially supported by the US National Science Foundation under grant DMS-0712827, by the Natural Science Foundation of China under grants 91130009 and 11071286, and by US Air Force Office of Scientific Research under grant FA9550-09-1-0511. Qinghui Zhang was partially supported by the Natural Science Foundation of China under grant 11001282, by Fundamental Research Funds for Central Universities of China, and by Guangdong Provincial Natural Science Foundation of China under grant S2011040003030.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

S. K. Berberian. *Notes on Spectral Theory*. Van Nostrand, New York, 1966.

M. S. Birman and M. Z. Solomjak. *Spectral Theory of Self-Adjoint Operators in Hilbert Space*. D. Reidel Publishing Company, Dordrecht, Holland, 1987.

S. Bochner. *Lectures on Fourier Integrals with an Author's Supplement on Monotonic Functions, Stieltjes Integrals, and Harmonic Analysis*. Annals of Mathematics Studies 42, Princeton University Press, New Jersey, 1959.

J. Burbea and P. Masani. *Banach and Hilbert Spaces of Vector-valued Functions*. Pitman Research Notes in Mathematics 90, Boston, MA, 1984.

A. Caponnetto, C. A. Micchelli, M. Pontil and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

C. Carmeli, E. De Vito and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Anal. Appl.*, 4:377–408, 2006.

C. Carmeli, E. De Vito, A. Toigo and V. Umanita. Vector valued reproducing kernel Hilbert spaces and universality. *Anal. Appl.*, 8:19–61, 2010.

J. B. Conway. *A Course in Functional Analysis*. 2nd Edition, Springer-Verlag, New York, 1990.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39:1–49, 2002.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, 2007.

I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

J. Diestel and J.J. Uhl, Jr. *Vector Measures*. American Mathematical Society, Providence, 1977.

T. Evgeniou, C. A. Micchelli and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

T. Evgeniou, M. Pontil and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13:1–50, 2000.

P. A. Fillmore. *Notes on Operator Theory*. Van Nostrand Company, New York, 1970.

R. A. Horn and C. B. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

S. Lowitzsh. Approximation and interpolation employing divergence-free radial basis functions with applications. Ph.D. Thesis, Texas A&M University, College Station, Texas, 2003.

S. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 315:69–87, 1989.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Comput.*, 17:177–204, 2005.

C. A. Micchelli, Y. Xu and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

S. Mukherjee and Q. Wu. Estimation of gradients and coordinate covariation in classification. *Journal of Machine Learning Research*, 7:2481-2514, 2006.

S. Mukherjee and D. X. Zhou. Learning coordinate covariances via gradients. *Journal of Machine Learning Research*, 7:519–549, 2006.

G. B. Pedrick. Theory of reproducing kernels for Hilbert spaces of vector valued functions. *Technical Report* **19**, University of Kansas, 1957.

W. Rudin. *Real and Complex Analysis*. 3rd Edition, McGraw-Hill, New York, 1987.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Mass, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

H. Wendland. Divergence-free kernel methods for approximating the Stokes problem. *SIAM J. Numer. Anal.*, 47:3158–3179, 2009.

Y. Xu and H. Zhang. Refinable kernels. *Journal of Machine Learning Research*, 8:2083–2120, 2007.

Y. Xu and H. Zhang. Refinement of reproducing kernels. *Journal of Machine Learning Research*, 10:107–140, 2009.

Y. Ying and C. Campbell. Learning coordinate gradients with multi-task kernels. In *COLT*, 2008.

H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

H. Zhang and L. Zhao. On the inclusion relation of reproducing kernel Hilbert spaces. *Anal. Appl.*, accetped subject to minor revision, *arXiv:1106.4075*, 2011.