Learning in RKHM: a C^*-Algebraic Twist for Kernel Machines

Yuka Hashimoto
NTT, Tokyo, Japan /
RIKEN AIP, Tokyo, Japan

Masahiro Ikeda
RIKEN AIP, Tokyo, Japan /
Keio University, Yokohama, Japan

Hachem Kadri
Aix-Marseille University, CNRS, LIS,
Marseille, France

Abstract

Supervised learning in reproducing kernel Hilbert space (RKHS) and vector-valued RKHS (vvRKHS) has been investigated for more than 30 years. In this paper, we provide a new twist to this rich literature by generalizing supervised learning in RKHS and vvRKHS to reproducing kernel Hilbert C^*-module (RKHM), and show how to construct effective positive-definite kernels by considering the perspective of C^*-algebra. Unlike the cases of RKHS and vvRKHS, we can use C^*-algebras to enlarge representation spaces. This enables us to construct RKHMs whose representation power goes beyond RKHSs, vvRKHSs, and existing methods such as convolutional neural networks. Our framework is suitable, for example, for effectively analyzing image data by allowing the interaction of Fourier components.

1 INTRODUCTION

Supervised learning in reproducing kernel Hilbert space (RKHS) has been actively investigated since the early 1990s (Murphy, 2012; Christmann & Steinwart, 2008; Shawe-Taylor & Cristianini, 2004; Schölkopf & Smola, 2002; Boser et al., 1992). The notion of reproducing kernels as dot products in Hilbert spaces was first brought to the field of machine learning by Aizerman et al. (1964), while the theoretical foundation of reproducing kernels and their Hilbert spaces dates back to at least Aronszajn (1950). By virtue of the representer theorem (Schölkopf et al., 2001), we can compute the solution of an infinite-dimensional minimization problem in RKHS with given finite samples. In addition to the standard RKHSs, applying vector-valued RKHSs (vvRKHSs) to supervised learning has also been proposed and used in analyzing vector-valued data (Micchelli & Pontil, 2005; Álvarez et al., 2012; Kadri et al., 2016; Minh et al., 2016; Brouard et al., 2016; Laforgue et al., 2020; Huusari & Kadri, 2021). Generalization bounds of the supervised problems in RKHS and vvRKHS are also derived (Mohri et al., 2018; Caponnetto & De Vito, 2007; Audiffren & Kadri, 2013; Huusari & Kadri, 2021).

Reproducing kernel Hilbert C^*-module (RKHM) is a generalization of RKHS and vvRKHS by means of C^*-algebra. C^*-algebra is a generalization of the space of complex values. It has a product and an involution structures. Important examples are the C^*-algebra of bounded linear operators on a Hilbert space and the C^*-algebra of continuous functions on a compact space. RKHMs have been originally studied for pure operator algebraic and mathematical physics problems (Manuilov & Troitsky, 2000; Heo, 2008; Moslehian, 2022). Recently, applying RKHMs to data analysis has been proposed by Hashimoto et al. (2021). They generalized the representer theorem in RKHS to RKHM, which allows us to analyze structured data such as functional data with C^*-algebras.

In this paper, we investigate supervised learning in RKHM. This provides a new twist to the state-of-the-art kernel-based learning algorithms and the development of a novel kind of reproducing kernels. An advantage of RKHM over RKHS and vvRKHS is that we can enlarge the C^*-algebra characterizing the RKHM to construct a representation space. This allows us to represent more functions than the case of RKHS and make use of the product structure in the C^*-algebra. Our main contributions are:

- We define positive definite kernels from the perspective of C^*-algebra, which are suitable for learning in RKHM and adapted to analyze image data.
- We derive a generalization bound of the supervised learning problem in RKHM, which generalizes existing results of RKHS and vvRKHS. We also show that the computational complexity of our method can be reduced if parameters in the C^*-algebra-valued positive definite kernels have specific structures.
- We show that our framework generalizes existing methods based on convolution operations.

Proceedings of the 26^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Important applications of the supervised learning in RKHM are tasks whose inputs and outputs are images. If the proposed kernels have specific parameters, then the product structure is the convolution, which corresponds to the pointwise product of Fourier components. By extending the C^*-algebra to a larger one, we can enjoy more general operations than the convolutions. This enables us to analyze image data effectively by making interactions between Fourier components. Regarding the generalization bound, we derive the same type of bound as those obtained for RKHS and vvRKHS via Rademacher complexity theory. This is to our knowledge, the first generalization bound for RKHM hypothesis classes. Concerning the connection with existing methods, we show that using our framework, we can reconstruct existing methods such as the convolutional neural network (LeCun et al., 1998) and the convolutional kernel (Mairal et al., 2014) and further generalize them. This fact implies that the representation power of our framework goes beyond the existing methods.

The remainder of this paper is organized as follows: In Section 2, we review mathematical notions related to this paper. We propose C^*-algebra-valued positive definite kernels in Section 3 and investigate supervised learning in RKHM in Section 4. Then, we show connections with existing convolution-based methods in Section 5. We confirm the advantage of our method numerically in Section 6 and conclude the paper in Section 7. All technical proofs are documented in the supplementary material (SM).

2 PRELIMINARIES

2.1 C^*-Algebra and Hilbert C^*-Module

C^*-algebra is a Banach space equipped with a product and an involution that satisfies the C^* identity.

Definition 2.1 (C^*-algebra) *A set $\mathcal{A}$ is called a C^*-*algebra *if it satisfies the following conditions:*

1. *$\mathcal{A}$ is an algebra over $\mathbb{C}$ and equipped with a bijection $(\cdot)^*: \mathcal{A} \to \mathcal{A}$ that satisfies the following conditions for $\alpha, \beta \in \mathbb{C}$ and $a, b \in \mathcal{A}$:*
   - $(\alpha a + \beta b)^* = \overline{\alpha} a^* + \overline{\beta} b^*$,
   - $(ab)^* = b^* a^*$,
   - $(a^*)^* = a$.
2. *$\mathcal{A}$ is a normed space endowed with $\|\cdot\|_{\mathcal{A}}$, and for $a, b \in \mathcal{A}$, $\|ab\|_{\mathcal{A}} \le \|a\|_{\mathcal{A}} \|b\|_{\mathcal{A}}$ holds. In addition, $\mathcal{A}$ is complete with respect to $\|\cdot\|_{\mathcal{A}}$.*
3. *For $a \in \mathcal{A}$, $\|a^* a\|_{\mathcal{A}} = \|a\|_{\mathcal{A}}^2$ holds.*

A C^*-algebra $\mathcal{A}$ is called unital if there exists $a \in \mathcal{A}$ such that $ab = b = ba$ for any $b \in \mathcal{A}$. We denote a by $1_{\mathcal{A}}$.

An example of C^*-algebra is group C^*-algebra (Kirillov, 1976). Let $p \in \mathbb{N}$, and let $\mathbb{Z}/p\mathbb{Z}$ be the set of integers modulo p.

Definition 2.2 (Group C^*-algebra on a finite cyclic group) *Let $\omega = \mathrm{e}^{2\pi\sqrt{-1}/p}$. The group C^*-algebra on $\mathbb{Z}/p\mathbb{Z}$, which is denoted as $C^*(\mathbb{Z}/p\mathbb{Z})$, is the set of maps from $\mathbb{Z}/p\mathbb{Z}$ to $\mathbb{C}$ equipped with the following product, involution, and norm:*

- $(x \cdot y)(z) = \sum_{w \in \mathbb{Z}/p\mathbb{Z}} x(z-w) y(w)$ *for* $z \in \mathbb{Z}/p\mathbb{Z}$,
- $x^*(z) = \overline{x(-z)}$,
- $\|x\| = \max_{n \in \{0,\ldots,p-1\}} |\sum_{z \in \mathbb{Z}/p\mathbb{Z}} x(z) \omega^{zn}|$.

Since the product is the convolution, group C^*-algebras offer a new way to define positive definite kernels, which are effective in analyzing image data as we will see in Section 3. Elements in $C^*(\mathbb{Z}/p\mathbb{Z})$ are described by circulant matrices (Gray, 2006). Let $Circ(p) = \{x \in \mathbb{C}^{p\times p} \mid x \text{ is a circulant matrix}\}$. Moreover, we denote the circulant matrix whose first row is v as $\mathrm{circ}(v)$. The discrete Fourier transform (DFT) matrix, whose (i,j)-entry is $\omega^{(i-1)(j-1)}/\sqrt{p}$, is denoted as F.

Lemma 2.3 *Any circulant matrix $x \in Circ(p)$ has an eigenvalue decomposition $x = F \Lambda_x F^*$, where*

$$\Lambda_x = \mathrm{diag}\left(\sum_{z \in \mathbb{Z}/p\mathbb{Z}} x(z) \omega^{z \cdot 0}, \ldots, \sum_{z \in \mathbb{Z}/p\mathbb{Z}} x(z) \omega^{z(p-1)} \right).$$

Lemma 2.4 *The group C^*-algebra $C^*(\mathbb{Z}/p\mathbb{Z})$ is C^*-isomorphic to $Circ(p)$.*

We now review important notions about C^*-algebra. We denote a C^*-algebra by $\mathcal{A}$.

Definition 2.5 (Positive) *An element a of $\mathcal{A}$ is called* positive *if there exists $b \in \mathcal{A}$ such that $a = b^* b$ holds. For $a, b \in \mathcal{A}$, we write $a \le_{\mathcal{A}} b$ if $b - a$ is positive, and $a \lneq_{\mathcal{A}} b$ if $b - a$ is positive and not zero. We denote by $\mathcal{A}_+$ the subset of $\mathcal{A}$ composed of all positive elements in $\mathcal{A}$. For any $a \in \mathcal{A}_+$, there exists a unique $b \in \mathcal{A}_+$ such that $a = b^2$. We denote b by $a^{1/2}$.*

Definition 2.6 (Minimum) *For a subset $\mathcal{S}$ of $\mathcal{A}$, $a \in \mathcal{A}$ is said to be a* lower bound *with respect to the order $\le_{\mathcal{A}}$, if $a \le_{\mathcal{A}} b$ for any $b \in \mathcal{S}$. Then, a lower bound $c \in \mathcal{A}$ is said to be an* infimum *of $\mathcal{S}$, if $a \le_{\mathcal{A}} c$ for any lower bound a of $\mathcal{S}$. If $c \in \mathcal{S}$, then c is said to be a* minimum *of $\mathcal{S}$.*

Hilbert C^*-module is a generalization of Hilbert space. We can define an $\mathcal{A}$-valued inner product and a (real nonnegative-valued) norm as a natural generalization of the complex-valued inner product. See Section A in SM for further details. Then, we define Hilbert C^*-module as follows.

Definition 2.7 (Hilbert C^*-module) *Let $\mathcal{M}$ be a C^*-module over $\mathcal{A}$ equipped with an $\mathcal{A}$-valued inner product. If $\mathcal{M}$ is complete with respect to the norm induced by the $\mathcal{A}$-valued inner product, it is called a* Hilbert C^*-module *over $\mathcal{A}$ or* Hilbert $\mathcal{A}$-module.

2.2 Reproducing Kernel Hilbert C^*-Module

RKHM is a generalization of RKHS by means of C^*-algebra. Let $\mathcal{X}$ be a non-empty set for data.

Definition 2.8 ($\mathcal{A}$-valued positive definite kernel) *An $\mathcal{A}$-valued map $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ is called a* positive definite kernel *if it satisfies the following conditions:*

- $k(x, y) = k(y, x)^*$ *for* $x, y \in \mathcal{X}$,
- $\sum_{i,j=1}^{n} c_i^* k(x_i, x_j) c_j \geq_{\mathcal{A}} 0$ *for* $n \in \mathbb{N}$, $c_i \in \mathcal{A}$, $x_i \in \mathcal{X}$.

Let $\phi : \mathcal{X} \to \mathcal{A}^{\mathcal{X}}$ be the *feature map* associated with k, which is defined as $\phi(x) = k(\cdot, x)$ for $x \in \mathcal{X}$. We construct the following C^*-module composed of $\mathcal{A}$-valued functions:

$$\mathcal{M}_{k,0} := \left\{ \sum_{i=1}^{n} \phi(x_i) c_i \,\middle|\, n \in \mathbb{N},\ c_i \in \mathcal{A},\ x_i \in \mathcal{X} \right\}.$$

Define an $\mathcal{A}$-valued map $\langle \cdot, \cdot \rangle_{\mathcal{M}_k} : \mathcal{M}_{k,0} \times \mathcal{M}_{k,0} \to \mathcal{A}$ as

$$\Big\langle \sum_{i=1}^{n} \phi(x_i) c_i, \sum_{j=1}^{l} \phi(y_j) b_j \Big\rangle_{\mathcal{M}_k} := \sum_{i=1}^{n} \sum_{j=1}^{l} c_i^* k(x_i, y_j) b_j.$$

By the properties of k in Definition 2.8, $\langle \cdot, \cdot \rangle_{\mathcal{M}_k}$ is well-defined and has the reproducing property

$$\langle \phi(x), v \rangle_{\mathcal{M}_k} = v(x),$$

for $v \in \mathcal{M}_{k,0}$ and $x \in \mathcal{X}$. Also, it is an $\mathcal{A}$-valued inner product. The *reproducing kernel Hilbert $\mathcal{A}$-module (RKHM)* associated with k is defined as the completion of $\mathcal{M}_{k,0}$. We denote by $\mathcal{M}_k$ the RKHM associated with k.

Hashimoto et al. (2021) showed the representer theorem in RKHM.

Proposition 2.9 (Representer theorem) *Let $\mathcal{A}$ be a unital C^*-algebra. Let $x_1, \ldots, x_n \in \mathcal{X}$ and $y_1, \ldots, y_n \in \mathcal{A}$. Let $h : \mathcal{X} \times \mathcal{A} \times \mathcal{A} \to \mathcal{A}_+$ be an error function and let $g : \mathcal{A}_+ \to \mathcal{A}_+$ satisfy $g(a) \lneq_{\mathcal{A}} g(b)$ for $a \lneq_{\mathcal{A}} b$. Assume the module (algebraically) generated by $\{\phi(x_i)\}_{i=1}^{n}$ is closed. Then, any $u \in \mathcal{M}_k$ minimizing $\sum_{i=1}^{n} h(x_i, y_i, u(x_i)) + g(|u|_{\mathcal{M}_k})$ admits a representation of the form $\sum_{i=1}^{n} \phi(x_i) c_i$ for some $c_1, \ldots, c_n \in \mathcal{A}$.*

Notation In the following, we denote the inner product, absolute value, and norm in $\mathcal{M}_k$ by $\langle \cdot, \cdot \rangle_k$, $|\cdot|_k$, and $\|\cdot\|_k$, respectively. See Table A in SM for more details.

3 C^*-ALGEBRA-VALUED POSITIVE DEFINITE KERNELS

To investigate the supervised learning problem in RKHM, we begin by constructing suitable C^*-algebra-valued positive definite kernels. The product structure used in these kernels will be shown to be effective in analyzing image data. However, the proposed kernels are general, and their application is not limited to image data.

Let $\mathcal{A}_1$ be a C^*-algebra. By the Gelfand–Naimark theorem (see, for example, Murphy 1990), there exists a Hilbert space $\mathcal{H}$ such that $\mathcal{A}_1$ is a subalgebra of the C^*-algebra $\mathcal{A}_2$ of bounded linear operators on $\mathcal{H}$. For image data we can set $\mathcal{A}_1$ and $\mathcal{A}_2$ as follows.

Example 3.1 (Image data analysis) *Let $p \in \mathbb{N}$, $\mathcal{A}_1 = C^*(\mathbb{Z}/p\mathbb{Z})$, and $\mathcal{A}_2 = \mathbb{C}^{p \times p}$. Then, $\mathcal{A}_1$ is a subalgebra of $\mathcal{A}_2$. Indeed, by Lemma 2.4, $\mathcal{A}_1 \simeq Circ(p)$. For example, in image processing, we represent filters by circulant matrices (Chanda & Majumder, 2011). If we regard $\mathbb{Z}/p\mathbb{Z}$ as the space of p pixels, then elements in $C^*(\mathbb{Z}/p\mathbb{Z})$ can be regarded as functions from pixels to intensities. Thus, we can also regard grayscale and color images with p pixels as elements in $C^*(\mathbb{Z}/p\mathbb{Z})$ and $C^*(\mathbb{Z}/p\mathbb{Z})^3$, respectively. Note that $\mathcal{A}_2$ is noncommutative, although $\mathcal{A}_1$ is commutative.*

We consider the case where the inputs are in $\mathcal{A}_1^d$ for $d \in \mathbb{N}$ and define linear, polynomial, and Gaussian C^*-algebra-valued positive definite kernels as follows. For example, we can consider the case where inputs are d images.

Definition 3.2 *Let $\mathcal{X} \subseteq \mathcal{A}_1^d$ and $x = [x_1, \ldots, x_d] \in \mathcal{X}$.*

1. *For $a_{i,1}, a_{i,2} \in \mathcal{A}_2$, the* linear kernel $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}_2$ *is defined as $k(x, y) = \sum_{i=1}^{d} a_{i,1}^* x_i^* a_{i,2}^* a_{i,2} y_i a_{i,1}$.*
2. *For $q \in \mathbb{N}$ and $a_{i,j} \in \mathcal{A}_2$ $(i = 1, \ldots d, j = 1, \ldots q+1)$, the* polynomial kernel $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}_2$ *is defined as*
$$k(x, y) = \sum_{i=1}^{d} \left(\prod_{j=1}^{q} a_{i,j}^* x_i^* \right) a_{i,q+1}^* a_{i,q+1} \left(\prod_{j=1}^{q} y_i a_{i,q+1-j} \right).$$
3. *Let Ω be a measurable space and μ is an $\mathcal{A}_2$-valued positive measure on Ω.[1] For $a_{i,1}, a_{i,2} : \Omega \to \mathcal{A}_2$, the* Gaussian kernel $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}_2$ *is defined as*
$$k(x, y) = \int_{\omega \in \Omega} \mathrm{e}^{-\sqrt{-1} \sum_{i=1}^{d} a_{i,1}(\omega)^* x_i^* a_{i,2}(\omega)^*} \mathrm{d}\mu(\omega) \times \mathrm{e}^{\sqrt{-1} \sum_{i=1}^{d} a_{i,2}(\omega) y_i a_{i,1}(\omega)}.$$
Here, we assume the integral does not diverge. In addition, the exponential is defined as the exponential of a bounded linear operator in $\mathcal{H}$.

Remark 3.3 *We can construct new kernels by the composition of functions to the kernels defined in Definition 3.2. For example, let $\psi_{i,j} : \mathcal{A}_1 \to \mathcal{A}_2$ for $i = 1, \ldots d$ and $j = 1, \ldots, q+1$. Then, the map defined by replacing x_i and y_i in the polynomial kernel by $\psi_{i,j}(x_i)$ and $\psi_{i,j}(y_i)$ is also an C^*-algebra-valued positive definite kernel.*

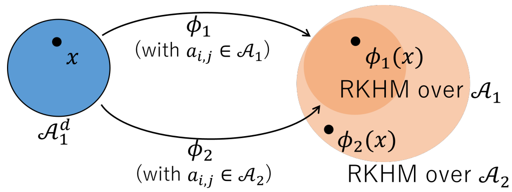


Figure 1: Representing samples in RKHM

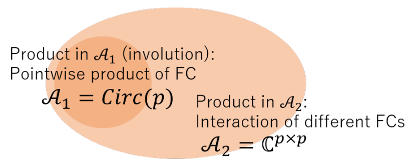


Figure 2: Product in $\mathcal{A}_1$ and $\mathcal{A}_2$ in Example 3.4

If $\mathcal{A}_1 = \mathcal{A}_2 = \mathbb{C}$, then the above kernels are reduced to the standard complex-valued positive definite kernels and the RKHMs associated with them are reduced to RKHSs. In this case, if $\mathcal{X} = \mathcal{A}^d$, the input space and the RKHS are both Hilbert spaces (Hilbert $\mathbb{C}$-modules). On the other hand, for RKHMs, if we choose $\mathcal{A}_1 \subsetneq \mathcal{A}_2$, then the input space $\mathcal{X}$ is a Hilbert $\mathcal{A}_1$-module, but the RKHM is a Hilbert $\mathcal{A}_2$-module, not $\mathcal{A}_1$-module. Applying RKHMs, we can construct higher dimensional spaces than input spaces but also enlarge the C^*-algebras characterizing the RKHMs, which allows us to represent more functions than RKHSs and make use of the product structure in $\mathcal{A}_2$. Figure 1 schematically shows the representation of samples in RKHM. We show an example related to image data below.

Example 3.4 (Image data analysis) *If $\mathcal{A}_1 = C^*(\mathbb{Z}/p\mathbb{Z})$, $\mathcal{A}_2 = \mathbb{C}^{p\times p}$ ($\mathcal{A}_1 \subsetneq \mathcal{A}_2$), and $a_{i,j} \in \mathcal{A}_1$, then $a_{i,j}$ in Definition 3.2 behaves as convolutional filters. In fact, by Definition 2.2, the multiplication of $a_{i,j}$ and x_i is represented by the convolution. The convolution of two functions corresponds to the multiplication of each Fourier component of them. Thus, each Fourier component of x_i does not interact with other Fourier components. Choosing $a_{i,j} \in \mathcal{A}_2$ outside $\mathcal{A}_1$ corresponds to the multiplication of different Fourier components of two functions. Indeed, let $x \in \mathcal{A}_1$. Then, by Lemma 2.4, x is represented as a circulant matrix and by Lemma 2.3, it is decomposed as $x = F\Lambda_x F^*$. In this case, Λ_x is the diagonal matrix whose ith diagonal is the ith Fourier component (FC) of x. Thus, if $a_{i,j} \in \mathcal{A}_1$, then we have $xa_{i,j} = F\Lambda_x\Lambda_{a_{i,j}}F^*$ and each Fourier component of x is multiplied by the same Fourier component of $a_{i,j}$. On the other hand, if $a_{i,j} \in \mathcal{A}_2 \setminus \mathcal{A}_1$, then $\Lambda_{a_{i,j}}$ is not a diagonal matrix, and the elements of $\Lambda_x\Lambda_{a_{i,j}}$ are composed of the weighted sum of different Fourier components of x. Figure 2 summarizes this example.*

[1] $\mathcal{A}_2$-valued measure which takes its values in $(\mathcal{A}_2)_+$. See Hashimoto et al. (2021, Appendix B) for a rigorous definition of C^*-algebra-valued measure.

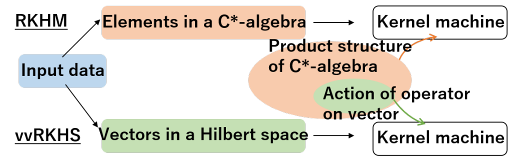


Figure 3: Comparison of RKHM with vvRKHS

Comparison with vvRKHS From the perspective of vvRKHS, defining kernels as in Definition 3.2 is difficult since for vvRKHS, the output space is a Hilbert space, and we do not have product structures in it. Indeed, the inner product in a vvRKHS is described by an action of an operator on a vector. We can regard the vector as a rank-one operator whose range is the one-dimensional space spanned by the vector. Thus, the action is regarded as the product of only two operators. On the other hand, from the perspective of C^*-algebra, we can multiply more than two elements in C^*-algebra, which allows us to define C^*-algebra-valued kernels naturally in the same manner as complex-valued kernels. See Figure 3 for a schematic explanation.

4 SUPERVISED LEARNING IN RKHM

We investigate supervised learning in RKHM. We first formulate the problem and derive a learning algorithm. Then, we characterize its generalization error and investigate its computational complexity.

We do not assume $\mathcal{X} \subseteq \mathcal{A}_1^d$ in Subsections 4.1 and 4.2. The input space $\mathcal{X}$ can be an arbitrary nonempty set in these sections. Thus, although we focus on the case of $\mathcal{X} \subseteq \mathcal{A}_1^d$ in this paper, the supervised learning in RKHM is applied to general problems whose output space is a C^*-algebra $\mathcal{A}$.

4.1 Problem Setting

Let $x_1, \ldots, x_n \in \mathcal{X}$ be input training samples and $y_1, \ldots, y_n \in \mathcal{A}$ be output training samples. Let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ be an $\mathcal{A}$-valued positive definite kernel, and let ϕ and $\mathcal{M}_k$ be the feature map and RKHM associated with k, respectively. We find a function $f : \mathcal{X} \to \mathcal{A}$ in $\mathcal{M}_k$ that maps input data to output data. For this purpose, we consider the following minimization problem:

$$\min_{f\in\mathcal{M}_k} \bigg(\sum_{i=1}^{n} |f(x_i) - y_i|_{\mathcal{A}}^2 + \lambda |f|_k^2 \bigg), \tag{1}$$

where $\lambda \geq 0$ is the regularization parameter. By the representer theorem (Proposition 2.9), we find a solution f in the submodule generated by $\{\phi(x_1), \ldots, \phi(x_n)\}$. As the case of RKHS (Schölkopf et al., 2001), representing f as

$\sum_{j=1}^n \phi(x_j)c_j$ ($c_j \in \mathcal{A}$), the problem is reduced to

$$\min_{c_j\in\mathcal{A}}\bigg(\sum_{i=1}^n\bigg|\sum_{j=1}^n k(x_i,x_j)c_j-y_i\bigg|_{\mathcal{A}}^2+\lambda\bigg|\sum_{j=1}^n\phi(x_j)c_j\bigg|_k^2\bigg)$$
$$=\min_{c_j\in\mathcal{A}}(\mathbf{c}^*\mathbf{G}^2\mathbf{c}-\mathbf{c}^*\mathbf{G}\mathbf{y}-\mathbf{y}^*\mathbf{G}\mathbf{c}+\lambda\mathbf{c}^*\mathbf{G}\mathbf{c}), \qquad (2)$$

where $\mathbf{G}$ is the $\mathcal{A}^{n\times n}$-valued Gram matrix whose (i,j)-entry is defined as $k(x_i,x_j)\in\mathcal{A}$, $\mathbf{c}=[c_1,\ldots,c_n]^T$, $\mathbf{y}=[y_1,\ldots,y_n]^T$, and $|a|_{\mathcal{A}}=(a^*a)^{1/2}$ for $a\in\mathcal{A}$. Note that $\mathbf{G}$ is a bounded linear operator on the Hilbert $\mathcal{A}$-module $\mathcal{A}^n$. If $\mathbf{G}+\lambda I$ is invertible, the solution of Problem (2) is $\mathbf{c}=(\mathbf{G}+\lambda I)^{-1}\mathbf{y}$.

4.2 Generalization Bound

We derive a generalization bound of the supervised problem in RKHM. We first define an $\mathcal{A}$-valued Rademacher complexity. Let (Ω,P) be a probability space. For a random variable (measurable map) $g:\Omega\to\mathcal{A}$, we denote by $\mathrm{E}[g]$ the Bochner integral of g, i.e., $\int_{\omega\in\Omega} g(\omega)\mathrm{d}P(\omega)$.

Definition 4.1 *Let $\sigma_1,\ldots,\sigma_n$ be i.i.d and mean zero $\mathcal{A}$-valued random variables and let $x_1,\ldots,x_n\in\mathcal{X}$ be given samples. Let $\boldsymbol{\sigma}=\{\sigma_i\}_{i=1}^n$ and $\mathbf{x}=\{x_i\}_{i=1}^n$. Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{A}$. The* $\mathcal{A}$-valued empirical Rademacher complexity $\hat{R}_{\mathcal{A}}(\mathcal{F},\boldsymbol{\sigma},\mathbf{x})$ *is defined as*

$$\hat{R}_{\mathcal{A}}(\mathcal{F},\boldsymbol{\sigma},\mathbf{x})=\mathrm{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(x_i)^*\sigma_i\bigg|_{\mathcal{A}}\bigg].$$

We derive an upper bound of the complexity of a function space related to the RKHM $\mathcal{M}_k$. We assume $\mathcal{A}$ is a unital C^*-algebra (See Definition 2.1).

Proposition 4.2 *Let $B>0$ and let $\mathcal{F}=\{f\in\mathcal{M}_k \mid \|f\|_k\le B\}$ and let $C=\|\,E[|\sigma_i|_{\mathcal{A}}^2]\,\|_{\mathcal{A}}$. Then, we have*

$$\hat{R}_{\mathcal{A}}(\mathcal{F},\boldsymbol{\sigma},\mathbf{x})\le_{\mathcal{A}}\frac{B\sqrt{C}}{n}\bigg(\sum_{i=1}^n\|k(x_i,x_i)\|_{\mathcal{A}}\bigg)^{1/2}1_{\mathcal{A}}.$$

Here, $1_{\mathcal{A}}$ is the unit in $\mathcal{A}$.

To prove Proposition 4.2, we first show the following $\mathcal{A}$-valued version of Jensen's inequality.

Lemma 4.3 *For a positive $\mathcal{A}$-valued random variable $c:\Omega\to\mathcal{A}_+$, we have $\mathrm{E}[c^{1/2}]\le_{\mathcal{A}}\mathrm{E}[c]^{1/2}$.*

4.2.1 Results for $\mathcal{A}\subseteq\mathbb{C}^{p\times p}$

In the following, we focus on that case and consider the trace of matrices. In Example 3.4, we focused on the case of $\mathcal{A}\subseteq\mathbb{C}^{p\times p}$, which is effective, for example, in analyzing image data. Thus, the results in this subsection are valid for these practical situations.

Deriving a bound for general C^*-algebras is technically difficult. One reason is the fact that the Radmecher complexity is C^*-algebra-valued and not scalar-valued. Thus, we focus on the the case of matrices and provide a standard way of transforming a matrix into a scalar value using the trace. The trace gives us more detailed information on eigenvalues compared to other measures, such as the operator norm. Indeed, for a positive definite matrix, the trace is the sum of the eigenvalues, while the operator norm is the largest eigenvalue. Moreover, the trace is linear and forms the Hilbert–Schmidt inner product.

Let $B>0$ and $E>0$. We put $\mathcal{F}=\{f\in\mathcal{M}_k \mid \|f\|_k\le B,\ f(x)\in\mathbb{R}^{p\times p}$ for any $x\in\mathcal{X}\}$, $\mathcal{G}(\mathcal{F})=\{\mathcal{X}\times\mathcal{Y}\ni(x,y)\mapsto|f(x)-y|_{\mathcal{A}}^2\in\mathcal{A}\mid f\in\mathcal{F}\}$, and $\mathcal{Y}=\{y\in\mathbb{R}^{p\times p}\mid\|y\|_{\mathcal{A}}\le E\}$. Let $x_1,\ldots,x_n\in\mathcal{X}$ and $y_1,\ldots,y_n\in\mathcal{Y}$. We assume there exists $D>0$ such that for any $x\in\mathcal{X}$, $\|k(x,x)\|_{\mathcal{A}}\le D$. In addition, let $L=2\sqrt{2}(B\sqrt{D}+E)$ and $M=DB^2+2\sqrt{D}BE+E^2$. Using the upper bound of the Rademacher complexity, we derive the following generalization bound.

Proposition 4.4 *Let $\mathrm{tr}(a)$ be the trace of $a\in\mathbb{C}^{p\times p}$. For any $g\in\mathcal{G}(\mathcal{F})$, any random variable $z:\Omega\to\mathcal{X}\times\mathcal{Y}$, and any $\delta\in(0,1)$, with probability $\ge 1-\delta$, we obtain*

$$\mathrm{tr}\bigg(\mathrm{E}[g(z)]-\frac{1}{n}\sum_{i=1}^n g(x_i,y_i)\bigg)\le 2\frac{LB\sqrt{D}p}{\sqrt{n}}+3\sqrt{2M}p\sqrt{\frac{\log(2/\delta)}{n}}.$$

Note that the same type of bounds is derived for RKHS (Mohri et al., 2018, Theorem 3.3) and for vvRKHS (Huusari & Kadri, 2021, Corollary 16, Sindhwani et al., 2013, Theorem 3.1, Sangnier et al., 2016, Theorem 4.1). Proposition 4.4 generalizes them to RKHM.

To show Proposition 4.4, we first evaluate the Rademacher complexity with respect to the squared loss function $(x,y)\mapsto|f(x)-y|_{\mathcal{A}}^2$. We use Theorem 3 of Maurer (2016) to obtain the following bound.

Lemma 4.5 *Let $s_1,\ldots,s_n$ be $\{-1,1\}$-valued Rademacher variables (i.e. independent uniform random variables taking values in $\{-1,1\}$) and let $\sigma_1,\ldots,\sigma_n$ be i.i.d. $\mathcal{A}$-valued random variables each of whose element is the Rademacher variable. Let $\mathbf{s}=\{s_i\}_{i=1}^n$, and $\mathbf{z}=\{(x_i,y_i)\}_{i=1}^n$. Then, we have*

$$\hat{R}_{\mathbb{C}}(\mathrm{tr}\mathcal{G}(\mathcal{F}),\mathbf{s},\mathbf{z})\le L\,\mathrm{tr}\,\hat{R}_{\mathcal{A}}(\mathcal{F},\boldsymbol{\sigma},\mathbf{x}),$$

where $\mathrm{tr}\,\mathcal{G}(\mathcal{F})=\{z\mapsto\mathrm{tr}(g(z))\mid g\in\mathcal{G}(\mathcal{F})\}$.

Next, we use Theorem 3.3 of Mohri et al. (2018) to derive an upper bound of the generalization error.

Lemma 4.6 *Let $z:\Omega\to\mathcal{X}\times\mathcal{Y}$ be a random variable and let $g\in\mathcal{G}(\mathcal{F})$. Under the same notations and assumptions* *as Lemma 4.5, for any $\delta \in (0,1)$, with probability $\geq 1-\delta$, we have*

$$\operatorname{tr}\left(\mathrm{E}[g(z)] - \frac{1}{n}\sum_{i=1}^n g(x_i,y_i)\right) \leq 2\hat{R}_{\mathbb{C}}(\operatorname{tr}\mathcal{G}(\mathcal{F}),\mathbf{s},\mathbf{z}) + 3\sqrt{2M}p\sqrt{\frac{\log(2/\delta)}{n}}.$$

4.3 Computational Complexity

As mentioned at the beginning of this section, we need to compute $(\mathbf{G}+\lambda I)^{-1}\mathbf{y}$ for a Gram matrix $\mathbf{G} \in \mathcal{A}^{n\times n}$ and a vector $y \in \mathcal{A}^n$ for solving the minimization problem (2). When $\mathcal{A} = \mathbb{C}^{p\times p}$, we have $\mathcal{A}^{n\times n} = \mathbb{C}^{np\times np}$, and $\mathbf{G}$ is huge if n, the number of samples, or p, the dimension of $\mathcal{A}$, is large. If we construct the np by np matrix explicitly and compute $(\mathbf{G}+\lambda I)^{-1}\mathbf{y}$ with a direct method such as Gaussian elimination and back substitution (for example, see Trefethen & Bau 1997), the computational complexity is $O(n^3p^3)$. However, if $\mathcal{X} = \mathcal{A}_1^d$, $\mathcal{A}_1 \subsetneq \mathcal{A}_2$, and parameters in the positive definite kernel have a specific structure, then we can reduce the computational complexity. For example, applying the fast Fourier transform, we can compute a multiplication of the DFT matrix F and a vector with $O(p\log p)$ (Van Loan, 1992). Let $\mathcal{A}_1 = C^*(\mathbb{Z}/p\mathbb{Z})$ and let $\mathcal{A}_2 = \mathbb{C}^{p\times p}$. Let k be an $\mathcal{A}_1$ or $\mathcal{A}_2$-valued positive definite kernel defined in Definition 3.2.

Proposition 4.7 *For $a_{i,j} \in \mathcal{A}_1$, the computational complexity for computing $(\mathbf{G}+\lambda I)^{-1}\mathbf{y}$ by direct methods for solving linear systems of equations is $O(np^2\log p + n^3p)$.*

We can use an iteration method for linear systems, such as the conjugate gradient (CG) method (Hestenes & Stiefel, 1952) to reduce the complexity with respect to n. Note that we need $O(np^2\log p)$ operations after all the iterations.

Proposition 4.8 *For $a_{i,j} \in \mathcal{A}_1$, the computational complexity for 1 iteration step of CG method is $O(n^2p)$.*

Proposition 4.9 *Let $a_{i,j} \in \mathcal{A}_2$ whose number of nonzero elements is $O(p\log p)$. Then, the computational complexity for 1 iteration step of CG method is $O(n^2p^2\log p)$.*

Remark 4.10 *If we do not use the structure of $\mathcal{A}_1$, then the computational complexities in Propositions 4.7, 4.8, and 4.9 are $O(n^3p^3)$, $O(n^2p^3)$, and $O(n^2p^3)$, respectively.*

In the case of RKHSs, techniques such as the random Fourier feature have been proposed to alleviate the computational cost of kernel methods (Rahimi & Recht, 2007). It could be interesting to inspect how to further reduce the computational complexity of learning in RKHM using random feature approximations for C^*-algebra-valued kernels; this is left for future work.

5 CONNECTION WITH EXISTING METHODS

5.1 Connection with Convolutional Neural Network

Convolutional neural network (CNN) has been one of the most successful methods for analyzing image data (LeCun et al., 1998; Li et al., 2021). We investigate the connection of the supervised learning problem in RKHM with CNN. In this subsection, we set $\mathcal{X} \subseteq \mathcal{A}_1 = C^*(\mathbb{Z}/p\mathbb{Z})$ and $\mathcal{A}_2 = \mathbb{C}^{p\times p}$. Since the product in $C^*(\mathbb{Z}/p\mathbb{Z})$ is characterized by the convolution, our framework with a specific $\mathcal{A}_1$-valued positive definite kernel enables us to reconstruct a similar model as the CNN.

We first provide an $\mathcal{A}_1$-valued positive definite kernel related to the CNN.

Proposition 5.1 *For $a_1,\ldots,a_L,b_1,\ldots,b_L \in \mathcal{A}_1$ and $\sigma_1,\ldots,\sigma_L : \mathcal{A}_1 \to \mathcal{A}_1$ each of which has a uniformly convergent series expansion $\sigma_j(x) = \sum_{l=1}^{\infty}\alpha_{j,l}x^l$ with $\alpha_{j,l} \geq 0$, let $\hat{k} : \mathcal{X}\times\mathcal{X} \to \mathcal{A}_1$ be defined as*

$$\hat{k}(x,y) = \sigma_L(b_L^*b_L + \sigma_{L-1}(b_{L-1}^*b_{L-1} + \cdots + \sigma_2(b_2^*b_2 + \sigma_1(b_1^*b_1 + x^*a_1^*a_1y)a_2^*a_2)\cdots \times a_{L-1}^*a_{L-1})a_L^*a_L). \tag{3}$$

Then, $\hat{k}$ is an $\mathcal{A}_1$-valued positive definite kernel.

Using the positive definite kernel (3), the solution f of the problem (2) is written as

$$f(x) = \sum_{i=1}^n \sigma_L(b_L^*b_L + \sigma_{L-1}(b_{L-1}^*b_{L-1} + \cdots + \sigma_2(b_2^*b_2 + \sigma_1(b_1^*b_1 + x^*a_1^*a_1x_i)a_2^*a_2)\cdots \times a_{L-1}^*a_{L-1})a_L^*a_L)c_i, \tag{4}$$

for some $c_i \in \mathcal{A}_1$. We regard $a_1^*a_1x_i$ and $a_j^*a_j$ for $j=2,\ldots,L$ as convolutional filters, $b_j^*b_j$ for $j=1,\ldots,L$ as biases, and σ_j for $j=1,\ldots,L$ as activation functions. Then, optimizing $a_1,\ldots,a_L,b_1,\ldots,b_L$ simultaneously with c_i corresponds to learning the CNN of the form (4).

The following proposition shows that the C^*-algebra-valued polynomial kernel defined in Definition 3.2 is general enough to represent the $\mathcal{A}_1$-valued positive definite kernel $\hat{k}$, related to the CNN. Therefore, by applying $\mathcal{A}_2$-valued polynomial kernel, not $\mathcal{A}_1$-valued polynomial kernel, we can go beyond the method with the convolution.

Proposition 5.2 *The $\mathcal{A}_1$-valued positive definite kernel $\hat{k}$ defined as Eq. (3) is composed of the sum of $\mathcal{A}_1$-valued polynomial kernels.*

5.2 Connection with Convolutional Kernel

For image data, a ($\mathbb{C}$-valued) positive definite kernel called convolutional kernel is proposed to bridge a gap between kernel methods and neural networks (Mairal et al., 2014; Mairal, 2016). In this subsection, we construct two C^*-algebra-valued positive definite kernels that generalize the convolutional kernel. Similar to the case of the CNN, we will first show that we can reconstruct the convolutional kernel using a C^*-algebra-valued positive definite kernel. Moreover, we will show that our framework gives another generalization of the convolutional kernel. A generalization of neural networks to C^*-algebra-valued networks is proposed (Hashimoto et al., 2022). This generalization allows us to generalize the analysis of the CNNs with kernel methods to that of C^*-algebra-valued CNNs.

Let Ω be a finite subset of $\mathbb{Z}^m$. For example, Ω is the space of m-dimensional grids. Let $\tilde{\mathcal{A}}_1$ be the space of $\mathbb{C}$-valued maps on Ω and $\mathcal{X} \subseteq \tilde{\mathcal{A}}_1$. The convolutional kernel is defined as follows (Mairal et al., 2014, Definition 2).

Definition 5.3 *Let $\beta, \sigma > 0$. The* convolutional kernel $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ *is defined as*

$$\tilde{k}(x,y) = \sum_{z,z' \in \Omega} |x(z)|\,|y(z')| \mathrm{e}^{-\frac{1}{2\beta^2}\|z-z'\|^2} \times \mathrm{e}^{-\frac{1}{2\sigma^2}|\tilde{x}(z)-\tilde{y}(z')|^2}. \quad (5)$$

Here, $\|\cdot\|$ is the standard norm in $\mathbb{C}^m$. In addition, for $x \in \mathcal{X}$, $\tilde{x}(z) = x(z)/|x(z)|$.

Let $\Omega = \{z_1, \ldots, z_p\}$, $\mathcal{A}_1 = C^*(\mathbb{Z}/p\mathbb{Z})$, and $\mathcal{A}_2 = \mathbb{C}^{p\times p}$. We first construct an $\mathcal{A}_1$-valued positive definite kernel, which reconstructs the convolutional kernel (5).

Proposition 5.4 *Define $\hat{k} : \mathcal{X} \times \mathcal{X} \to \mathcal{A}_1$ as*

$$\hat{k}(x,y) = \int_{\mathbb{R}} \int_{\mathbb{R}^m} c_x(\omega,\eta)^* c_y(\omega,\eta)\, \mathrm{d}\lambda_\beta(\omega)\mathrm{d}\lambda_\sigma(\eta), \quad (6)$$

where $\mathrm{d}\lambda_\beta(\omega) = \beta \mathrm{e}^{-\frac{\beta^2\omega^2}{2}}\mathrm{d}\omega$ for $\beta > 0$ and

$$c_x(\omega,\eta) = \mathrm{circ}\Big(|x(z_1)|\mathrm{e}^{\sqrt{-1}\omega\cdot z_1}\mathrm{e}^{\sqrt{-1}\eta\cdot\tilde{x}(z_1)}, \cdots, |x(z_p)|\mathrm{e}^{\sqrt{-1}\omega\cdot z_p}\mathrm{e}^{\sqrt{-1}\eta\cdot\tilde{x}(z_p)}\Big),$$

for $x \in \mathcal{X}$, $\omega \in \mathbb{R}^m$, and $\eta \in \mathbb{R}$. Then, $\hat{k}$ is an $\mathcal{A}_1$-valued positive definite kernel, and for any $l = 1, \ldots, p$, $\tilde{k}$ in Eq. (5) is written as

$$\tilde{k}(x,y) = \frac{1}{p}\sum_{i,j=1}^{p} \hat{k}(x,y)_{i,j} = \sum_{j=1}^{p} \hat{k}(x,y)_{l,j},$$

where $\hat{k}(x,y)_{i,j}$ is the (i,j)-entry of $\hat{k}(x,y)$.

Remark 5.5 *Similar to Subsection 5.1, we can generalize $\hat{k}$ by replacing $c_x(\cdot,\cdot)^* c_y(\cdot,\cdot)$ by an $\mathcal{A}_2$-valued polynomial kernel with respect to $c_x(\cdot,\cdot)$ and $c_y(\cdot,\cdot)$ in Eq. (6).*

Instead of $\mathcal{A}_1$-valued, we can also construct an $\tilde{\mathcal{A}}_1$-valued kernel, which reconstructs the convolutional kernel (5).

Definition 5.6 *Let $\beta, \sigma > 0$. Define $\check{k} : \mathcal{X} \times \mathcal{X} \to \tilde{\mathcal{A}}_1$ as*

$$\check{k}(x,y)(w) = \sum_{z,z' \in \Omega} |x(\psi(z,w))|\,|y(\psi(z',w))| \times \mathrm{e}^{\frac{-1}{2\beta^2}\|\psi(z,w)-\psi(z',w)\|^2}\mathrm{e}^{\frac{-1}{2\sigma^2}|\tilde{x}(\psi(z,w))-\tilde{y}(\psi(z',w))|^2} \quad (7)$$

for $w \in \Omega$. Here, $\psi : \Omega \times \Omega \to \Omega$ is a map satisfying $\psi(z,0) = z$ for any $z \in \Omega$.

The $\tilde{\mathcal{A}}_1$-valued map $\check{k}$ is a generalization of the ($\mathbb{C}$-valued) convolutional kernel $\tilde{k}$ in the following sense, which is directly derived from the definitions of $\check{k}$ and $\tilde{k}$.

Proposition 5.7 *For $\tilde{k}$ and $\check{k}$ defined as Eqs.* (5) *and* (7)*, respectively, we have $\check{k}(x,y)(0) = \tilde{k}(x,y)$.*

We further generalize the $\tilde{\mathcal{A}}_1$-valued kernel $\check{k}$ to an $\mathcal{A}_2$-valued positive definite kernel.

Definition 5.8 *Let $\beta, \sigma > 0$ and $a_i \in \mathcal{A}_2$ for $i = 1,2,3,4$. Let ψ be the same map as that in Eq.* (7)*. Let $k : \mathcal{X} \times \mathcal{X} \to \mathcal{A}_2$ be defined as*

$$k(x,y) = \int_{\mathbb{R}}\int_{\mathbb{R}^m} \sum_{z,z' \in \Omega} a_1^* \mathbf{x}(z) a_2^* b(z,\omega)^* a_3^* \tilde{\mathbf{x}}(z,\eta)^* \times a_4^* a_4 \tilde{\mathbf{y}}(z',\eta) a_3 b(z',\omega) a_2 \mathbf{y}(z') a_1 \,\mathrm{d}\lambda_\beta(\omega)\mathrm{d}\lambda_\sigma(\eta) \quad (8)$$

for $x, y \in \mathcal{X}$. Here, for $x \in \mathcal{X}$,

$$\mathbf{x}(z) = \mathrm{diag}(|x(\psi(z,z_1)|, \ldots, |x(\psi(z,z_p))|) \in \mathcal{A}_2,$$
$$\tilde{\mathbf{x}}(z,\omega) = \mathrm{diag}(\mathrm{e}^{-\sqrt{-1}\omega\cdot\tilde{x}(\psi(z,z_1))}, \ldots, \mathrm{e}^{-\sqrt{-1}\omega\cdot\tilde{x}(\psi(z,z_p))}) \in \mathcal{A}_2,$$
$$b(z,\omega) = \mathrm{diag}(\mathrm{e}^{-\sqrt{-1}\omega\cdot\psi(z,z_1)}, \ldots, \mathrm{e}^{-\sqrt{-1}\omega\cdot\psi(z,z_p)}) \in \mathcal{A}_2.$$

Proposition 5.9 *The $\mathcal{A}_2$-valued map k defined as Eq.* (8) *is an $\mathcal{A}_2$-valued positive definite kernel.*

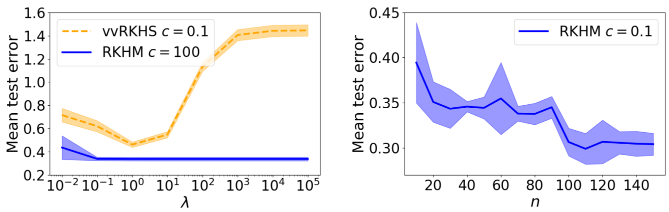
The following proposition shows k is a generalization of $\check{k}$, which means we finally generalize the ($\mathbb{C}$-valued) convolution kernel $\tilde{k}$ to an $\mathcal{A}_2$-valued positive definite kernel. This allows us to generalize the relationship between the CNNs and the convolutional kernel to that of a C^*-algebra-valued version of the CNNs and the C^*-algebra-valued convolutional kernel k.

Proposition 5.10 *If $a_i = I$, then the $\mathcal{A}_2$-valued positive definite kernel k defined as Eq.* (8) *is reduced to the $\mathcal{A}_1$-valued convolutional kernel $\check{k}$ defined as Eq.* (7).

6 NUMERICAL RESULTS

6.1 Experiments with Synthetic Data

We compared the performances of supervised learning in RKHMs and vvRKHSs. We generated n samples



(a) Regularization parameter λ (b) Number of samples n

Figure 4: Mean test error versus hyperparameters (Mean value $\pm$ standard deviation of 5 runs).

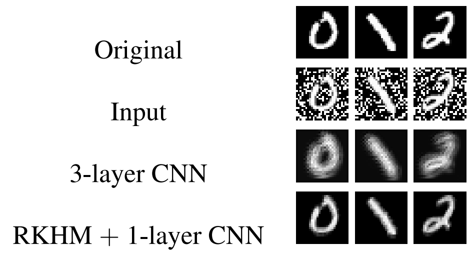
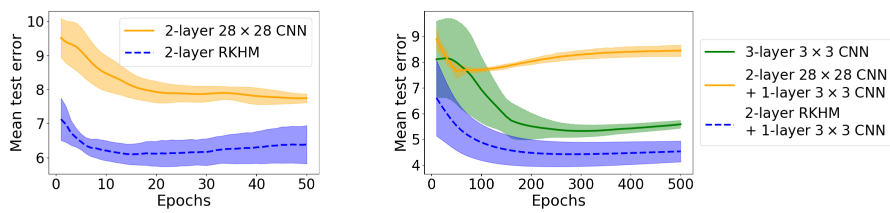


Figure 5: Comparison between RKHM and CNN



(a) 2-layer models (b) 3-layer models

Figure 6: Mean test error versus the number of epochs (Mean value $\pm$ standard deviation of 5 runs).

Table 1: Comparison between an RKHM and vvRKHSs (Mean value $\pm$ standard deviation of 5 runs)

		Mean error
vvRKHS, Gaussian ($\tilde{k}(x,y) = \mathrm{e}^{-c\|x-y\|^2}$)	$k = \tilde{k}I$	0.640 ± 0.122
	$k = \tilde{k}T$	0.603 ± 0.028
	Nonsep	0.650 ± 0.051
vvRKHS, Laplacian ($\tilde{k}(x,y) = \mathrm{e}^{-c\|x-y\|}$)	$k = \tilde{k}I$	0.538 ± 0.027
	$k = \tilde{k}T$	0.590 ± 0.021
	Nonsep	0.650 ± 0.048
vvRKHS, Polynomial ($\tilde{k}(x,y) = \sum_{i=1}^{3}(1 - cx \cdot y)^i$)	$k = \tilde{k}I$	0.800 ± 0.032
	$k = \tilde{k}T$	0.539 ± 0.012
	Nonsep	0.539 ± 0.012
RKHM ($k(x,y) = \sum_{i=1}^{3} R_x^*(I - cQ_x^*)^i(I - cQ_y)^i R_y$)		$\mathbf{0.343 \pm 0.022}$

$$T = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \text{Nonsep: } k(x_1, x_2)_{i,j} = \tilde{k}(x_{1,i}, x_{2,j})$$

$x_1, \ldots, x_n$ in $[0,1]^2$ each of whose elements is independently drawn from the uniform distribution on $[0,1]$. For a generated sample $x_i = [x_{i,1}, x_{i,2}]$, we added noise $\xi_i \in \mathbb{R}^2$, each of whose elements is independently drawn from the Gaussian distribution with mean 0 and standard deviation 0.1. We generated the corresponding output sample y_i as $y_i = [\sin(\tilde{x}_{i,1} + \tilde{x}_{i,2}), \sin(\tilde{x}_{i,1} + \tilde{x}_{i,2}) + \sin(0.5(\tilde{x}_{i,1} + \tilde{x}_{i,2}))] \in \mathbb{R}^2$, where $\tilde{x}_i = x_i + \xi_i$. We learned a function f that maps x_i to y_i in different RKHMs and vvRKHSs and different values of the regularization parameter λ. To compare the performances, we generated 100 test input samples $\hat{x}_1, \ldots, \hat{x}_{100}$ in $[0,1]^2$ each of whose elements is independently drawn from the uniform distribution on $[0,1]$. We also generated $\hat{y}_1, \ldots, \hat{y}_{100}$ given by $\hat{y}_i = [\sin(\hat{x}_{i,1} + \hat{x}_{i,2}), \sin(\hat{x}_{i,1} + \hat{x}_{i,2}) + \sin(0.5(\hat{x}_{i,1} + \hat{x}_{i,2}))]$. We computed the mean error $1/100 \sum_{i=1}^{100} \|f(\hat{x}_i) - \hat{y}_i\|$. The results for $n = 30$ are illustrated in Table 1 and Figure 4. Regarding Table 1, we executed a cross-validation grid search to find the best parameters c and λ, where c is a parameter in the positive definite kernels and λ is the regularization parameter. Regarding Figure 4 (a), we set c as the parameter found by the cross-validation and computed the error for different values of λ. We remark that the mean error for the RKHM becomes large as λ becomes large, but because of the scale of the vertical axis, we cannot see the change clearly in the figure. We can see that RKHM outperforms vvRKHSs. We also show the relationship between the mean error and the number of samples in Figure 4 (b). We can see that the mean error becomes small as the number of samples becomes large.

Regarding the learning in RKHMs, for $i = 1, \ldots, n$, we transformed $x_i \in [0,1]^2$ into $\mathrm{circ}(x_i) \in Circ(2)$. Then, we set $\mathcal{A}_1 = Circ(2)$ and $\mathcal{A}_2 = \mathbb{C}^{2\times 2}$. We computed the solution of the minimization problem (2) and obtained a function $\hat{f} \in \mathcal{M}_k$ that maps $\mathrm{circ}(x_i)$ to $\mathrm{circ}(y_i)$. Since the output of the learned function $\hat{f}$ takes its value on $\mathcal{A}_2$, we computed the mean value of $(1,1)$ and $(2,2)$ entries of $\hat{f}(\hat{x}_i)$ for obtaining the first element of the output vector in $\mathbb{R}^2$ and that of $(1,2)$ and $(2,1)$ entries for the second element. Regarding the C^*-algebra-valued kernel for RKHM, we set $k(x,y) = \sum_{i=1}^{3} R_x^*(I - cQ_x^*)^i(I - cQ_y)^i R_y$ for $x \in \mathcal{A}_1$, where $x = Q_x R_x$ is the QR decomposition of x.

6.2 Experiments with MNIST

We compared our method with CNNs using MNIST (LeCun et al., 1998). Our objective is to find a function that maps a noisy image to its original image using an RKHM and a CNN. For $i = 1, \ldots, 20$, we generated training samples as follows: We added noise to each pixel of an original image y_i and generated a noisy image x_i. The noise is drawn from the normal distribution with mean 0 and standard deviation 0.01. Moreover, each digit (0–9) is contained in the training sample set equally (i.e., the number of samples for each digit is 2). The image size is 28×28. We represent input and output images x_i and y_i as the circulant matrices $\mathrm{circ}(x_i)$ and $\mathrm{circ}(y_i)$ whose first rows are x_i and y_i. Then, we learned the function in the RKHM associated with a polynomial kernel $k(x, y) = (a_3^* \sigma(xa_1 + a_2)^* + a_4^*)(\sigma(ya_1 + a_2)a_3 + a_4)$, where $\sigma(x) = (I - cQ_x)R_x + (I - cQ_x)^3 R_x$. Since k has 4 $\mathcal{A}_2$-valued parameters, it corresponds to a generalization of 2-layer CNN with 28×28 filters (see Subsection 5.1). Regarding the parameters a_i, we used the simple gradient descent method and optimized them. We generated 100 noisy images for test samples in the same manner as the training samples and computed the mean error with respect to them. For comparison, we also trained a 2-layer CNN with 28×28 filters with the same training samples. The results are illustrated in Figure 6 (a). We can see that the RKHM outperforms the CNN. Moreover, we combined the RKHM with a 1-layer CNN with a 3×3 filter, whose inputs are the outputs of the function learned in the RKHM. We also trained a 3-layer CNN with 3×3 filters and a 2-layer CNN with 28×28 filters combined with a 1-layer CNN with a 3×3 filter. The results are illustrated in Figures 5 and 6 (b). We can see that by replacing convolutional layers with an RKHM, we can achieve better performance. RKHMs and convolutional layers with 28×28 filters capture global information of images. According to the results of the CNN with 28×28 filters and the RKHM in Figure 6 (b), we can see that the RKHM can capture global information of the images more effectively. On the other hand, convolutional layers with 3×3 filters capture local information. Since the 2-layer RKHM combined with a 1-layer CNN with a 3×3 filter outperforms a 3-layer CNN with 3×3 filters, we conclude that the combination of the RKHM and CNN captures the global and local information more effectively.

7 CONCLUSION AND DISCUSSION

We investigated supervised learning in RKHM and provided a new twist and insights for kernel methods. We constructed C^*-algebra-valued kernels from the perspective of C^*-algebra, which is suitable, for example, for analyzing image data. We investigated the generalization bound and computational complexity for RKHM learning and showed the connection with existing methods. RKHMs enable us to construct larger representation spaces than the case of RKHSs and vvRKHSs, and generalize operations such as convolution. This fact implies the representation power of RKHMs goes beyond that of existing frameworks.

There are several interesting topics for future work. First, more results for infinite-dimensional C^*-algebras would be helpful for analyzing functional data, as Hashimoto et al. (2021) proposed. Results about generalization bounds in Subsection 4.2.1 are not applicable to infinite-dimensional C^*-algebras since the trace is not available for general bounded linear operators. Finding a nice transformation of elements in general C^*-algebras into a scalar value and deriving a bound based on the transformation can help us understand the generalization property of kernel methods for functional data. We may also define a C^*-algebra-valued generalization error and bound it directly with the order in C^*-algebras. Then, improving results in Subsection 4.2.1 using the features of C^*-algebras can tell us more advantages of applying C^*-algebras to kernel methods.

Studying excess risk is another interesting topic (see, for example, Marteau-Ferey et al. (2019b,a). We documented an argument how we can apply existing results to our problem for finite-dimensional C^*-algebras in Section C in SM. Similar to the case of generalization error, generalizing and improving these results using the features of C^*-algebras is interseting.

Finally, applying group C^*-algebras on non-abelian groups would enable us to investigate group equivalent kernel methods and neural networks (for example, Sonoda et al. (2022)). We discussed the connection between RKHMs over $C^*(\mathbb{Z}/p\mathbb{Z})$ and CNNs. Since the product in group C^*-algebras is defined by the group convolution, generalizing our results for an abelian group $\mathbb{Z}/p\mathbb{Z}$ to a non-abelian group help us understand the connection between RKHMs and group equivalent neural networks.

Acknowledgements

Hachem Kadri is partially supported by grant ANR-19-CE23-0011 from the French National Research Agency. Masahiro Ikeda is partially supported by grant JPMJCR1913 from JST CREST.

References

Aizerman, M. A., Braverman, E. M., and Rozonoer, L. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

Álvarez, M. A., Rosasco, L., Lawrence, N. D., et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Aronszajn, N. Theory of reproducing kernels. *Transactions* *of the American mathematical society*, 68(3):337–404, 1950.

Audiffren, J. and Kadri, H. Stability of multi-task kernel regression algorithms. In *Proceedings of the 5th Asian Conference on Machine Learning (ACML)*, pp. 1–16, 2013.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual workshop on Computational learning theory (COLT)*, pp. 144–152, 1992.

Brouard, C., Szafranski, M., and d'Alché Buc, F. Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. *Journal of Machine Learning Research*, 17:1–48, 2016.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Chanda, B. and Majumder, D. D. *Digital Image Processing and Analysis*. PHI Learning, 2nd edition, 2011.

Christmann, A. and Steinwart, I. *Support Vector Machines*. Springer, 2008.

Gray, R. M. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

Hashimoto, Y., Ishikawa, I., Ikeda, M., Komura, F., Katsura, T., and Kawahara, Y. Reproducing kernel Hilbert C^*-module and kernel mean embeddings. *Journal of Machine Learning Research*, 22(267):1–56, 2021.

Hashimoto, Y., Wang, Z., and Matsui, T. C^*-algebra net: a new approach generalizing neural network parameters to C^*-algebra. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Heo, J. Reproducing kernel Hilbert C^*-modules and kernels associated with cocycles. *Journal of Mathematical Physics*, 49(10):103507, 2008.

Hestenes, M. R. and Stiefel, E. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.

Huusari, R. and Kadri, H. Entangled kernels - beyond separability. *Journal of Machine Learning Research*, 22(24): 1–40, 2021.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., and Audiffren, J. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016.

Kirillov, A. A. *Elements of the Theory of Representations*. Springer, 1976.

Laforgue, P., Lambert, A., Brogat-Motte, L., and d'Alché-Buc, F. Duality in RKHSs with infinite dimensional outputs: Application to robust losses. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Lance, E. C. *Hilbert C^*-modules – a Toolkit for Operator Algebraists*. London Mathematical Society Lecture Note Series, vol. 210. Cambridge University Press, 1995.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Mairal, J. End-to-end kernel learning with supervised convolutional kernel networks. In *Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. Convolutional kernel networks. In *Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

Manuilov, V. M. and Troitsky, E. V. Hilbert C^* and W^*-modules and their morphisms. *Journal of Mathematical Sciences*, 98(2):137–201, 2000.

Marteau-Ferey, U., Bach, F., and Rudi, A. Globally convergent newton methods for ill-conditioned generalized self-concordant losses. In *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2019a.

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, 2019b.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT)*, 2016.

Micchelli, C. A. and Pontil, M. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Minh, H. Q., Bazzani, L., and Murino, V. A unifying framework in vector-valued reproducing kernel Hilbert spaces for manifold regularization and co-regularized multi-view learning. *Journal of Machine Learning Research*, 17(25):1–72, 2016.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT press, 2018.

Moslehian, M. S. Vector-valued reproducing kernel Hilbert C^*-modules. *Complex Analysis and Operator Theory*, 16(1):Paper No. 2, 2022.

Murphy, G. J. *C^*-Algebras and Operator Theory*. Academic Press, 1990.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Proceedings of the Advances in Neural Information Processing Systems 20 (NIPS)*, 2007.

Sangnier, M., Fercoq, O., and d'Alché-Buc, F. Joint quantile regression in vector-valued RKHSs. In *Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory (COLT)*, 2001.

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

Sindhwani, V., Minh, H. Q., and Lozano, A. C. Scalable matrix-valued kernel learning for high-dimensional nonlinear multivariate regression and granger causality. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

Sonoda, S., Ishikawa, I., and Ikeda, M. Universality of group convolutional neural networks based on ridgelet analysis on groups. In *Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2022.

Trefethen, L. N. and Bau, D. *Numerical Linear Algebra*. SIAM, 1997.

Van Loan, C. *Computational Frameworks for the Fast Fourier Transform*. SIAM, 1992.

Notation

The typical notations in this paper are listed in Table A.

A Hilbert C^*-module

We provide definitions and a lemma related to Hilbert C^*-module.

Definition A.1 (C^*-module) *Let $\mathcal{M}$ be an abelian group with an operation $+$. If $\mathcal{M}$ is equipped with a (right) $\mathcal{A}$-multiplication, then $\mathcal{M}$ is called a (right) C^**-module *over $\mathcal{A}$.*

Definition A.2 ($\mathcal{A}$-valued inner product) *Let $\mathcal{M}$ be a C^*-module over $\mathcal{A}$. A $\mathbb{C}$-linear map with respect to the second variable $\langle\cdot,\cdot\rangle_{\mathcal{M}}:\mathcal{M}\times\mathcal{M}\to\mathcal{A}$ is called an $\mathcal{A}$-valued* inner product *if it satisfies the following properties for $u,v,w\in\mathcal{M}$ and $a,b\in\mathcal{A}$:*

1. $\langle u, va+wb\rangle_{\mathcal{M}} = \langle u,v\rangle_{\mathcal{M}}\, a + \langle u,w\rangle_{\mathcal{M}}\, b$,
2. $\langle v,u\rangle_{\mathcal{M}} = \langle u,v\rangle_{\mathcal{M}}^*$,
3. $\langle u,u\rangle_{\mathcal{M}} \geq_{\mathcal{A}} 0$,
4. *If* $\langle u,u\rangle_{\mathcal{M}} = 0$, *then* $u=0$.

Definition A.3 ($\mathcal{A}$-valued absolute value and norm) *Let $\mathcal{M}$ be a C^*-module over $\mathcal{A}$. For $u\in\mathcal{M}$, the $\mathcal{A}$*-valued absolute value *$|u|_{\mathcal{M}}$ on $\mathcal{M}$ is defined by the positive element $|u|_{\mathcal{M}}$ of $\mathcal{A}$ such that $|u|_{\mathcal{M}}^2=\langle u,u\rangle_{\mathcal{M}}$. The nonnegative real-valued norm $\|\cdot\|_{\mathcal{M}}$ on $\mathcal{M}$ is defined by $\|u\|_{\mathcal{M}}=\big\||u|_{\mathcal{M}}\big\|_{\mathcal{A}}$.*

Similar to the case of Hilbert spaces, the following Cauchy–Schwarz inequality for $\mathcal{A}$-valued inner products is available (Lance, 1995, Proposition 1.1).

Lemma A.4 (Cauchy–Schwarz inequality) *Let $\mathcal{M}$ be a Hilbert $\mathcal{A}$-module. For $u,v\in\mathcal{M}$, the following inequality holds:*

$$|\langle u,v\rangle_{\mathcal{M}}|_{\mathcal{A}}^2 \;\leq_{\mathcal{A}} \|u\|_{\mathcal{M}}^2\,\langle v,v\rangle_{\mathcal{M}}\,.$$

B Proofs

We provide the proofs of the propositions and lemmas in the main thesis.

Lemma 2.4 *The group C^*-algebra $C^*(\mathbb{Z}/p\mathbb{Z})$ is C^*-isomorphic to $Circ(p)$.*

Proof Let $f:C^*(\mathbb{Z}/p\mathbb{Z})\to Circ(p)$ be a map defined as $f(x)=\mathrm{circ}(x(0),\ldots,x(p-1))$. Then, f is linear and invertible. In addition, we have

$$\begin{aligned}
f(x)f(y) &= \mathrm{circ}\left(\sum_{z\in\mathbb{Z}/p\mathbb{Z}} x(0-z)y(z),\ldots,\sum_{z\in\mathbb{Z}/p\mathbb{Z}} x(p-1-z)y(z)\right)\\
&= \mathrm{circ}((x\cdot y)(0),\ldots,(x\cdot y)(p-1)) = f(x\cdot y),\\
f(x)^* &= \mathrm{circ}(\overline{x(0)},\overline{x(p-1)},\ldots,\overline{x(1)}) = f(x^*),\\
\|f(x)\| &= \left\|F\,\mathrm{diag}\left(\sum_{z\in\mathbb{Z}/p\mathbb{Z}} x(z)\mathrm{e}^{2\pi\sqrt{-1}z\cdot 0/p},\ldots,\sum_{z\in\mathbb{Z}/p\mathbb{Z}} x(z)\mathrm{e}^{2\pi\sqrt{-1}z(p-1)/p}\right)F^*\right\| = \|x\|,
\end{aligned}$$

where the last formula is derived by Lemma 2.3. Thus, f is a C^*-isomorphism. □

In the following, for a probability space Ω and a random variable (measurable map) $g:\Omega\to\mathbb{C}$, the integral of g is denoted by $\mathrm{E}[g]$.

Table A: Notation table

$\mathcal{A}$	A C^*-algebra
$\mathcal{A}_+$	The subset of $\mathcal{A}$ composed of all positive elements in $\mathcal{A}$
$\le_{\mathcal{A}}$	For $a, b \in \mathcal{A}$, $a \le_{\mathcal{A}} b$ means $b - a$ is positive
$\lneq_{\mathcal{A}}$	For $a, b \in \mathcal{A}$, $a \lneq_{\mathcal{A}} b$ means $b - a$ is positive and nonzero
$\lvert \cdot \rvert_{\mathcal{A}}$	The $\mathcal{A}$-valued absolute value in $\mathcal{A}$ defined as $\lvert a\rvert_{\mathcal{A}} = (a^* a)^{1/2}$ for $a \in \mathcal{A}$.
$\mathcal{X}$	An input space
$\mathcal{Y}$	An output space
k	An $\mathcal{A}$-valued positive definite kernel
ϕ	The feature map endowed with k
$\mathcal{M}_k$	The RKHM associated with k
$\mathbf{G}$	The $\mathcal{A}$-valued Gram matrix defined as $\mathbf{G}_{i,j} = k(x_i, x_j)$ for given samples $x_1, \ldots, x_n \in \mathcal{X}$
F	The discrete Fourier transform (DFT) matrix, whose (i, j)-entry is $\omega^{(i-1)(j-1)}/\sqrt{p}$

Lemma 4.3 *For a positive $\mathcal{A}$-valued random variable $c : \Omega \to \mathcal{A}_+$, we have* $\mathrm{E}[c^{1/2}] \le_{\mathcal{A}} \mathrm{E}[c]^{1/2}$.

Proof For any $\epsilon > 0$, let $x_0 = \mathrm{E}[c + \epsilon 1_{\mathcal{A}}]$, $a = 1/2 x_0^{-1/2}$, and $b = 1/2 x_0^{1/2}$. Then, we have $a x_0 + b = x_0^{1/2}$ and for any $x \in \mathcal{A}_+$, we have

$$\begin{aligned}(ax+b)^*(ax+b) - x &= \frac{1}{4} x x_0^{-1} x + \frac{1}{4} x + \frac{1}{4} x + \frac{1}{4} x_0 - x\\ &= \left(\frac{1}{2} x_0^{-1/2} x - \frac{1}{2} x_0^{1/2}\right)^* \left(\frac{1}{2} x_0^{-1/2} x - \frac{1}{2} x_0^{1/2}\right) = (ax - b)^*(ax - b) \ge_{\mathcal{A}} 0.\end{aligned}$$

Thus, we have $ax + b = \lvert ax + b\rvert_{\mathcal{A}} \ge_{\mathcal{A}} \lvert x^{1/2}\rvert_{\mathcal{A}} = x^{1/2}$. Therefore, we have

$$\mathrm{E}\big[(c + \epsilon 1_{\mathcal{A}})^{1/2}\big] \le_{\mathcal{A}} \mathrm{E}[a(c + \epsilon 1_{\mathcal{A}}) + b] = a x_0 + b = x_0^{1/2} = \mathrm{E}[(c + \epsilon 1_{\mathcal{A}})]^{1/2}.$$

Since $\epsilon > 0$ is arbitrary and $\mathcal{A}_+$ is closed, we have $\mathrm{E}[c^{1/2}] \le_{\mathcal{A}} \mathrm{E}[c]^{1/2}$. ☐

Proposition 4.2 *Let $B > 0$ and let $\mathcal{F} = \{f \in \mathcal{M}_k \mid \|f\|_k \le B\}$ and let $C = \|\, E[\lvert\sigma_i\rvert_{\mathcal{A}}^2] \,\|_{\mathcal{A}}$. Then, we have*

$$\hat{R}(\mathcal{F}, \boldsymbol{\sigma}, \mathbf{x}) \le_{\mathcal{A}} \frac{B\sqrt{C}}{n} \left(\sum_{i=1}^n \|k(x_i, x_i)\|_{\mathcal{A}} \right)^{1/2} 1_{\mathcal{A}}.$$

Proof By Lemma 4.3, we have

$$\begin{aligned}\hat{R}(\mathcal{F}, \boldsymbol{\sigma}, \mathbf{x}) &= \mathrm{E}\left[\sup_{f \in \mathcal{F}} \left\lvert \frac{1}{n} \sum_{i=1}^n f(x_i)^* \sigma_i \right\rvert_{\mathcal{A}}\right] = \frac{1}{n} \mathrm{E}\left[\sup_{f \in \mathcal{F}} \left\lvert \left\langle f, \sum_{i=1}^n \phi(x_i) \sigma_i \right\rangle_k \right\rvert_{\mathcal{A}}\right]\\ &\le \frac{1}{n} \mathrm{E}\left[\sup_{f \in \mathcal{F}} \left\lvert \sum_{i=1}^n \phi(x_i) \sigma_i \right\rvert_k \|f\|_k\right] = \frac{1}{n} \mathrm{E}\left[\left\lvert \sum_{i=1}^n \phi(x_i) \sigma_i \right\rvert_k B\right] = \frac{B}{n} \mathrm{E}\left[\left(\sum_{i,j=1}^n \sigma_i^* k(x_i, x_j) \sigma_j \right)^{1/2}\right]\\ &\le \frac{B}{n} \mathrm{E}\left[\sum_{i,j=1}^n \sigma_i^* k(x_i, x_j) \sigma_j \right]^{1/2} = \frac{B}{n} \left(\sum_{i=1}^n \mathrm{E}[\sigma_i^* k(x_i, x_i) \sigma_i] \right)^{1/2} \le \frac{B}{n} \left(\sum_{i=1}^n \mathrm{E}[\sigma_i^* \sigma_i] \|k(x_i, x_i)\|_{\mathcal{A}} \right)^{1/2}\\ &\le \frac{B}{n} \left(\sum_{i=1}^n C 1_{\mathcal{A}} \|k(x_i, x_i)\|_{\mathcal{A}} \right)^{1/2} = \frac{B\sqrt{C}}{n} \left(\sum_{i=1}^n \|k(x_i, x_i)\|_{\mathcal{A}} \right)^{1/2} 1_{\mathcal{A}},\end{aligned}$$

where the third inequality is derived by the Cauchy–Schwartz inequality (Lemma A.4). ☐

In the following, we put $\mathcal{A}=\mathbb{C}^{p\times p}$. The following lemmas are applied for the proofs of Lemmas 4.5 and 4.6.

Lemma B.1 *Let $a,b\in\mathbb{R}^{p\times p}$ or $a,b\in\mathcal{A}_+$. If $a\le_{\mathcal{A}} b$, then* $\operatorname{tr}(a)\le\operatorname{tr}(b)$.

Proof Since $b-a\in\mathcal{A}_+$, we have $0\le\operatorname{tr}(b-a)=\operatorname{tr}(b)-\operatorname{tr}(a)$. □

Lemma B.2 *Let $\mathcal{S}$ be a subset of $\mathcal{A}_+$. Then,* $\operatorname{tr}(\sup_{s\in\mathcal{S}} s)\ge\sup_{s\in\mathcal{S}}\operatorname{tr}(s)$. *If there exists $\hat{s}\in\mathcal{S}$ such that* $\sup_{s\in\mathcal{S}} s=\hat{s}$, *then* $\operatorname{tr}(\sup_{s\in\mathcal{S}} s)=\sup_{s\in\mathcal{S}}\operatorname{tr}(s)$.

Proof Let $\epsilon>0$. Then, there exists $t\in\mathcal{S}$ such that

$$(1-\epsilon)\sup_{s\in\mathcal{S}}\operatorname{tr}(s)\le\operatorname{tr}(t)\le\operatorname{tr}(\sup_{s\in\mathcal{S}} s).$$

Since $\epsilon>0$ is arbitrary, we have $\operatorname{tr}(\sup_{s\in\mathcal{S}} s)\ge\sup_{s\in\mathcal{S}}\operatorname{tr}(s)$.

If there exists $\hat{s}\in\mathcal{S}$ such that $\sup_{s\in\mathcal{S}} s=\hat{s}$, then we have

$$\operatorname{tr}(\sup_{s\in\mathcal{S}} s)=\operatorname{tr}(\hat{s})\le\sup_{s\in\mathcal{S}}\operatorname{tr}(s).$$

□

Lemma B.3 *Let $a\in\mathbb{R}^{p\times p}$. Then,* $\operatorname{tr}(a)\le\operatorname{tr}(|a|_{\mathcal{A}})$.

Proof Let $\lambda_1,\ldots,\lambda_p$ be eigenvalues of a, and let $\kappa_1,\ldots,\kappa_p$ be singular values of a. Then, by Weyl's inequality, we have

$$\operatorname{tr}(a)=\sum_{i=1}^p\lambda_i\le\sum_{i=1}^p|\lambda_i|\le\sum_{i=1}^p\kappa_i=\operatorname{tr}(|a|_{\mathcal{A}}).$$

□

We now show Lemmas 4.5 and 4.6.

Lemma 4.5 *Let $s_1,\ldots,s_n$ be $\{-1,1\}$-valued Rademacher variables (i.e. independent uniform random variables taking values in $\{-1,1\}$) and let $\sigma_1,\ldots,\sigma_n$ be i.i.d. $\mathcal{A}$-valued random variables each of whose element is the Rademacher variable. Let $\mathbf{s}=\{s_i\}_{i=1}^n$, and $\mathbf{z}=\{(x_i,y_i)\}_{i=1}^n$. Then, we have*

$$\hat{R}(\operatorname{tr}\mathcal{G}(\mathcal{F}),\mathbf{s},\mathbf{z})\le L\operatorname{tr}\hat{R}(\mathcal{F},\boldsymbol{\sigma},\mathbf{x}).$$

Proof For $f_1,f_2\in\mathcal{F}$, we have

$$\begin{aligned}\operatorname{tr}(|f_1(x_i)-y_i|_{\mathcal{A}}^2)-\operatorname{tr}(|f_2(x_i)-y_i|_{\mathcal{A}}^2)&=\operatorname{tr}((f_1(x_i)-y_i+f_2(x_i)-y_i)^*(f_1(x_i)-y_i-f_2(x_i)+y_i))\\&\le\|f_1(x_i)-y_i+f_2(x_i)-y_i\|_{\mathcal{A}}\,\|f_1(x_i)-y_i-f_2(x_i)+y_i\|_{\mathrm{HS}},\end{aligned}$$

where the first equality holds since for $a_1,a_2\in\mathbb{R}^{p\times p}$, $\operatorname{tr}(a_1^*a_2)=\operatorname{tr}(a_2^*a_1)$ and $\|\cdot\|_{\mathrm{HS}}$ is the Hilbert–Schmidt norm in $\mathbb{C}^{p\times p}$. In addition, we have

$$\begin{aligned}\|f_1(x_i)-y_i+f_2(x_i)-y_i\|_{\mathcal{A}}&=\|\left\langle\phi(x_i),f_1+f_2\right\rangle_k-2y_i\|_{\mathcal{A}}\\&\le\|k(x_i,x_i)\|^{1/2}\|f_1+f_2\|_k+2\|y_i\|_{\mathcal{A}}\le2(B\sqrt{D}+E)=\frac{L}{\sqrt{2}}.\end{aligned}$$

Thus, by setting $\psi_i(f) = \mathrm{tr}(|f(x_i) - y_i|^2)$, $\phi_i(f) = L/\sqrt{2} f(x_i)$, and $\|\cdot\| = \|\cdot\|_{\mathrm{HS}}$ in Theorem 3 of Maurer (2016), we obtain

$$\begin{aligned}
&\mathrm{E}\bigg[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n s_i \,\mathrm{tr}(|f(x_i)-y_i|_{\mathcal{A}}^2)\bigg] \le \sqrt{2}\frac{L}{\sqrt{2}}\mathrm{E}\bigg[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \langle \sigma_i, f(x_i)\rangle_{\mathrm{HS}}\bigg]\\
&\le L\mathrm{E}\bigg[\sup_{f\in\mathcal{F}} \mathrm{tr}\bigg|\frac{1}{n}\sum_{i=1}^n f(x_i)^*\sigma_i\bigg|_{\mathcal{A}}\bigg] \le L\,\mathrm{tr}\,\mathrm{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(x_i)^*\sigma_i\bigg|_{\mathcal{A}}\bigg].
\end{aligned}$$

□

Lemma 4.6 *Let $z:\Omega\to\mathcal{X}\times\mathcal{Y}$ be a random variable and let $g\in\mathcal{G}(\mathcal{F})$. Under the same notations and assumptions as Lemma 4.5, for any $\delta\in(0,1)$, with probability $\ge 1-\delta$, we have*

$$\mathrm{tr}\bigg(\mathrm{E}[g(z)] - \frac{1}{n}\sum_{i=1}^n g(x_i,y_i)\bigg) \le 2\hat{R}(\mathrm{tr}\,\mathcal{G}(\mathcal{F}),\mathbf{s},\mathbf{z}) + 3\sqrt{2M}p\sqrt{\frac{\log(2/\delta)}{n}}.$$

Proof For a random variable $S=(z_1,\ldots,z_n):\Omega\to(\mathcal{X}\times\mathcal{Y})^n$, let $\Phi(S)=\sup_{g\in\mathcal{G}(\mathcal{F})}\mathrm{tr}(\mathrm{E}[g(z)] - 1/n\sum_{i=1}^n g(z_i))$. For $i=1,\ldots,n$, let $S_i=(z_1',\ldots,z_n')$, where $z_j=z_j'$ for $j\ne i$ and $z_i\ne z_i'$. Then, we have

$$\begin{aligned}
\Phi(S)-\Phi(S_i) &\le \sup_{g\in\mathcal{G}(\mathcal{F})}\mathrm{tr}\bigg(\mathrm{E}[g(z)]-\frac{1}{n}\sum_{j=1}^n g(z_j)\bigg) - \sup_{g\in\mathcal{G}(\mathcal{F})}\mathrm{tr}\bigg(\mathrm{E}[g(z)]-\frac{1}{n}\sum_{j=1}^n g(z_j')\bigg)\\
&\le \frac{1}{n}\sup_{g\in\mathcal{G}(\mathcal{F})}\mathrm{tr}\bigg(\sum_{j=1}^n g(z_j) - \sum_{j=1}^n g(z_j')\bigg) = \frac{1}{n}\sup_{g\in\mathcal{G}(\mathcal{F})}\mathrm{tr}(g(z_i)-g(z_i'))\\
&\le \frac{p}{n}\sup_{g\in\mathcal{G}(\mathcal{F})}\|g(z_i)-g(z_i')\|_{\mathcal{A}} \le \frac{2\sqrt{M}p}{n}.
\end{aligned}$$

The final inequality is derived since

$$\begin{aligned}
\|g(z_i)\|_{\mathcal{A}} &= \||f(x_i)-y_i|_{\mathcal{A}}^2\|_{\mathcal{A}} = \||\langle\phi(x_i),f\rangle_k - y_i|_{\mathcal{A}}^2\|_{\mathcal{A}}\\
&\le \big(\|k(x_i,x_i)\|^{1/2}\|f\|_k + \|y_i\|_{\mathcal{A}}\big)^2 \le DB^2 + 2\sqrt{D}BE + E^2.
\end{aligned}$$

By McDiarmid's inequality, for any $\delta\in(0,1)$, with probability $\ge 1-\delta/2$, we have

$$\Phi(S) - \mathrm{E}[\Phi(S)] \le \sqrt{\frac{1}{2}\sum_{i=1}^n\bigg(\frac{2\sqrt{M}p}{n}\bigg)^2\log\frac{2}{\delta}} = \sqrt{2M}p\sqrt{\frac{\log\frac{2}{\delta}}{n}}$$

Thus, for any $g\in\mathcal{G}(\mathcal{F})$, we have

$$\mathrm{tr}\bigg(\mathrm{E}[g(z)] - \frac{1}{n}\sum_{i=1}^n g(z_i)\bigg) \le \Phi(S) \le \mathrm{E}[\Phi(S)] + \sqrt{2M}p\sqrt{\frac{\log\frac{2}{\delta}}{n}}.$$

For the remaining part, the proof is the same as that of Theorem 3.3 of Mohri et al. (2018). Since

$$\Phi(S) = \sup_{g\in\mathcal{G}(\mathcal{F})}\bigg(\mathrm{E}[\mathrm{tr}(g(z))] - \frac{1}{n}\sum_{i=1}^n \mathrm{tr}(g(z_i))\bigg),$$

we replace g in the proof of Theorem 3.3 in Mohri et al. (2018) by $z\mapsto\mathrm{tr}(g(z))$ in our case and derive

$$\begin{aligned}
\mathrm{tr}\bigg(\mathrm{E}[g(z)] - \frac{1}{n}\sum_{i=1}^n g(x_i,y_i)\bigg) &\le 2\mathrm{E}\bigg[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n s_i\,\mathrm{tr}\,|f(x_i)-y_i|_{\mathcal{A}}^2\bigg] + 3\sqrt{2M}p\sqrt{\frac{\log\frac{2}{\delta}}{n}}\\
&\le 2\hat{R}(\mathrm{tr}\,\mathcal{G}(\mathcal{F}),\mathbf{s},\mathbf{z}) + 3\sqrt{2M}p\sqrt{\frac{\log\frac{2}{\delta}}{n}},
\end{aligned}$$

which completes the proof. $\square$

Proposition 4.7 *For $a_{i,j} \in \mathcal{A}_1$, the computational complexity for computing $(\mathbf{G}+\lambda I)^{-1}\mathbf{y}$ by direct methods for solving linear systems of equations is $O(np^2 \log p + n^3 p)$.*

Proof Since all the elements of $\mathbf{G}$ and $\mathbf{y}$ are in $\mathcal{A}_1$, we have

$$(\mathbf{G}+\lambda I)^{-1}\mathbf{y} = (\mathbf{F}\mathbf{\Lambda}_{\mathbf{G}+\lambda I}^{-1}\mathbf{F}^*)\mathbf{F}\mathbf{\Lambda}_{\mathbf{y}}F^* = \mathbf{F}\mathbf{\Lambda}_{\mathbf{G}+\lambda I}^{-1}\mathbf{\Lambda}_{\mathbf{y}}F^*,$$

where $\mathbf{F}$ is the $\mathbb{C}^{p\times p}$-valued $n\times n$ diagonal matrix whose diagonal elements are all F. In addition, $\mathbf{\Lambda}_{\mathbf{G}+\lambda I}$ is the $\mathcal{A}_1$-valued $n\times n$ whose (i,j)-entry is $\Lambda_{k(x_i,x_j)}$, and $\mathbf{\Lambda}_{\mathbf{y}}$ is the vector in $\mathcal{A}_1^n$ whose ith element is Λ_{y_i}. If we use the fast Fourier transformation, then the computational complexity of computing Fy for $y \in \mathbb{C}^{p\times p}$ is $O(p^2\log p)$. Moreover, since the computational complexity of multiplication $\Lambda_x\Lambda_y$ for $x,y\in\mathcal{A}_1$ is $O(p)$, using Gaussian elimination and back substitution, the computational complexity of computing $\mathbf{\Lambda}_{\mathbf{G}+\lambda I}^{-1}\mathbf{\Lambda}_{\mathbf{y}}$ is $O(n^3p)$. As a result, the total computational complexity is $O(np^2\log p + n^3p)$. $\square$

Proposition 4.9 *Let $a_{i,j}\in\mathcal{A}_2$ whose number of nonzero elements is $O(p\log p)$. Then, the computational complexity for 1 iteration step of CG method is $O(n^2p^2\log p)$.*

Proof The computational complexity for computing 1 iteration step of CG method is equal to that of computing $(\mathbf{G}+\lambda I)\mathbf{b}$ for $\mathbf{b}\in\mathcal{A}_2^n$. For $b\in\mathcal{A}_2$, the computational complexity of computing $k(x_i,x_j)b$ is $O(p^2\log p)$ since those of computing $a_{i,j}b$ and x_ib are both $O(p^2\log p)$. (For x_ib, we use fast Fourier transformation.) Therefore, the computational complexity of computing $(\mathbf{G}+\lambda I)\mathbf{b}$ is $O(n^2p^2\log p)$. $\square$

Proposition 5.1 *For $a_1,\ldots,a_L,b_1,\ldots,b_L\in\mathcal{A}_1$ and $\sigma_1,\ldots,\sigma_L:\mathcal{A}_1\to\mathcal{A}_1$ each of which has an expansion $\sigma_j(x)=\sum_{l=1}^{\infty}\alpha_{j,l}x^l$ with $\alpha_{j,l}\ge 0$, let $\hat{k}:\mathcal{X}\times\mathcal{X}\to\mathcal{A}_1$ be defined as*

$$\hat{k}(x,y) = \sigma_L(b_L^*b_L + \sigma_{L-1}(b_{L-1}^*b_{L-1} + \cdots + \sigma_2(b_2^*b_2 + \sigma_1(b_1^*b_1 + x^*a_1^*a_1y)a_2^*a_2)\cdots\times a_{L-1}^*a_{L-1})a_L^*a_L). \tag{3}$$

Then, $\hat{k}$ is an $\mathcal{A}_1$-valued positive definite kernel.

Proof Let $l:\mathcal{X}\times\mathcal{X}\to\mathcal{A}_1$ be an $\mathcal{A}_1$-valued positive definite kernel and $\sigma:\mathcal{A}_1\to\mathcal{A}_1$ be a map that has an expansion $\sigma(x)=\sum_{j=1}^{\infty}\alpha_jx^j$ with $\alpha_j\ge 0$. Then, $\sigma\circ l$ is also an $\mathcal{A}_1$-valued positive definite kernel. Indeed, for $d_1,\ldots,d_n\in\mathcal{A}_1$ and $x_1,\ldots,x_n\in\mathcal{X}$, we have

$$\sum_{i,j=1}^{n} d_i^*\sigma(l(x_i,x_j))d_j = \sum_{i,j=1}^{n}\sum_{s=1}^{\infty}\alpha_s d_i^* l(x_i,x_j)^s d_j \ge_{\mathcal{A}_1} 0.$$

Since $(x,y)\mapsto b_1^*b_1+x^*a_1^*a_1y$ is an $\mathcal{A}_1$-valued positive definite kernel, $(x,y)\mapsto\sigma_1(b_1^*b_1+x^*a_1^*a_1y)$ is also an $\mathcal{A}_1$-valued positive definite kernel. Moreover, since $\sigma_1(b_1^*b_1+x^*a_1^*a_1y)$ and a_2 are in $\mathcal{A}_1$, $(x,y)\mapsto b_2^*b_2+\sigma_1(b_1^*b_1+x^*a_1^*a_1y)a_2a_2^*$ is also an $\mathcal{A}_1$-valued positive definite kernel. We iteratively apply the above result and obtain the positive definiteness of $\hat{k}$. $\square$

Proposition 5.2 *The $\mathcal{A}_1$-valued positive definite kernel $\hat{k}$ defined as Eq.* (3) *is composed of the sum of $\mathcal{A}_1$-valued polynomial kernels.*

Proof Since $(x,y)\mapsto b_1^*b_1+x^*a_1^*a_1y$ is an $\mathcal{A}_1$-valued polynomial kernel and $\sigma:\mathcal{A}_1\to\mathcal{A}_1$ is a map that has an expansion $\sigma(x)=\sum_{j=1}^{\infty}\alpha_jx^j$, $\hat{k}$ is composed of the sum of $\mathcal{A}_1$-valued polynomial kernels. $\square$

Proposition 5.4 *Define* $\hat{k} : \mathcal{X} \times \mathcal{X} \to \mathcal{A}_1$ *as*

$$\hat{k}(x, y) = \int_{\mathbb{R}} \int_{\mathbb{R}^m} c_x(\omega, \eta)^* c_y(\omega, \eta) \, \mathrm{d}\lambda_\beta(\omega) \mathrm{d}\lambda_\sigma(\eta), \tag{6}$$

where $\mathrm{d}\lambda_\beta(\omega) = \beta \mathrm{e}^{-\frac{\beta^2 \omega^2}{2}} \mathrm{d}\omega$ *for* $\beta > 0$ *and*

$$c_x(\omega, \eta) = \mathrm{circ}\left(|x(z_1)| \mathrm{e}^{\sqrt{-1}\omega \cdot z_1} \mathrm{e}^{\sqrt{-1}\eta \cdot \tilde{x}(z_1)}, \cdots, |x(z_p)| \mathrm{e}^{\sqrt{-1}\omega \cdot z_p} \mathrm{e}^{\sqrt{-1}\eta \cdot \tilde{x}(z_p)} \right),$$

for $x \in \mathcal{X}$*,* $\omega \in \mathbb{R}^m$*, and* $\eta \in \mathbb{R}$*. Then,* $\hat{k}$ *is an* $\mathcal{A}_1$*-valued positive definite kernel, and for any* $l = 1, \dots, p$*,* $\tilde{k}$ *is written as*

$$\tilde{k}(x, y) = \frac{1}{p} \sum_{i,j=1}^{p} \hat{k}(x, y)_{i,j} = \sum_{j=1}^{p} \hat{k}(x, y)_{l,j},$$

where $\hat{k}(x, y)_{i,j}$ *is the* (i, j)*-entry of* $\hat{k}(x, y)$*.*

Proof The positive definiteness of $\hat{k}$ is trivial. As for the relationship between $\tilde{k}$ and $\hat{k}$, we have

$$\begin{aligned}
\hat{k}(x, y)_{i,j} &= \int_{\mathbb{R}} \int_{\mathbb{R}^m} \sum_{l=1}^{p} |x(z_{p-i+2+l})| \mathrm{e}^{-\sqrt{-1}\omega \cdot z_{p-i+2+l}} \mathrm{e}^{-\sqrt{-1}\eta \cdot \tilde{x}(z_{p-i+2+l})} \\
&\qquad \times |y(z_{p-j+2+l})| \mathrm{e}^{\sqrt{-1}\omega \cdot z_{p-j+2+l}} \mathrm{e}^{\sqrt{-1}\eta \cdot \tilde{y}(z_{p-j+2+l})} \mathrm{d}\lambda_\beta(\omega) \mathrm{d}\lambda_\sigma(\eta) \\
&= \sum_{l=1}^{p} |x(z_{p-i+2+l})| \, |y(z_{p-j+2+l})| \mathrm{e}^{-\frac{1}{2\sigma^2} |\tilde{x}(z_{p-i+2+l}) - \tilde{y}(z_{p-j+2+l})|^2} \mathrm{e}^{-\frac{1}{2\beta^2} \| z_{p-i+2+l} - z_{p-j+2+l} \|^2}.
\end{aligned}$$

Thus, we have

$$\sum_{i=1}^{p} \hat{k}(x, y)_{i,j} = \sum_{i,l=1}^{p} |x(z_{l+j-i})| \, |y(z_l)| \mathrm{e}^{-\frac{1}{2\beta^2} |\tilde{x}(z_{l+j-i}) - \tilde{y}(z_l)|^2} \mathrm{e}^{-\frac{1}{2\sigma^2} \| z_{l+j-i} - z_l \|^2} = \tilde{k}(x, y).$$

□

Proposition 5.9 *The* $\mathcal{A}_2$*-valued map* k *defined as Eq.* (8) *is an* $\mathcal{A}_2$*-valued positive definite kernel.*

Proof For $n \in \mathbb{N}$, $c_1, \dots, c_n \in \mathcal{A}_2$, and $x_1, \dots, x_n \in \mathcal{A}_1^d$, we have

$$\begin{aligned}
&\sum_{i,j=1}^{n} c_i^* k(x_i, x_j) c_j \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}^m} \sum_{i=1}^{n} \sum_{z \in \Omega} c_i^* a_1^* \mathbf{x}_i(z) a_2^* b(z, \omega)^* a_3^* \tilde{\mathbf{x}}_i(z, \eta)^* a_4^* \sum_{j=1}^{n} \sum_{z' \in \Omega} a_4 \tilde{\mathbf{x}}_j(z', \eta) a_3 b(z', \omega) a_2 \mathbf{x}_j(z') a_1 c_j \, \mathrm{d}\lambda_\beta(\omega) \mathrm{d}\lambda_\sigma(\eta),
\end{aligned}$$

which is positive semi-definite. □

C Regarding excess risk

We discuss the excess risk (Marteau-Ferey et al., 2019b,a) of our supervised problem. In this section, we focus on the case $\mathcal{A} = \mathbb{C}^{p \times p}$. In this case, we can also regard $\mathcal{A}^n$ as a Hilbert space equipped with the Hilbert–Schmidt inner product. Thus, taking the trace of our problem (2), we get

$$\mathrm{tr} \min_{\mathbf{c} \in \mathcal{A}^n} \left(\sum_{i=1}^{n} \tilde{\ell}_{z_i}(\mathbf{c}) + \lambda \mathbf{c}^* \mathbf{G} \mathbf{c} \right) = \min_{\mathbf{c} \in \mathcal{A}^n} \mathrm{tr} \left(\sum_{i=1}^{n} \tilde{\ell}_{z_i}(\mathbf{c}) + \lambda \mathbf{c}^* \mathbf{G} \mathbf{c} \right) = \min_{\mathbf{c} \in \mathcal{A}^n} \left(\sum_{i=1}^{n} \mathrm{tr}(\tilde{\ell}_{z_i}(\mathbf{c})) + \lambda \, \mathrm{tr}(\mathbf{c}^* \mathbf{G} \mathbf{c}) \right),$$

where $z_i = (x_i, y_i)$, $\tilde{\ell}_{z_i}(\mathbf{c}) = |\Phi(x_i)^* \mathbf{c} - y_i|_{\mathcal{A}}^2$, and $\Phi(x_i)^* = [k(x_i, x_1), \dots, k(x_i, x_n)]$. The first equality is derived by Lemma B.2. Therefore, by setting $\ell = \mathrm{tr} \, \tilde{\ell}$, our problem is reduced to the problem discussed in Section 2 by Marteau-Ferey et al. (2019b). Therefore, their results enable us to obtain an excess risk bound.