Opuscula Math. 44, no. 1 (2024), 79–103
https://doi.org/10.7494/OpMath.2024.44.1.79 Opuscula Mathematica

CONDITIONAL MEAN EMBEDDING AND OPTIMAL FEATURE SELECTION VIA POSITIVE DEFINITE KERNELS

Palle E.T. Jorgensen, Myung-Sin Song, and James Tian

Communicated by P.A. Cojuhari

Abstract. Motivated by applications, we consider new operator-theoretic approaches to conditional mean embedding (CME). Our present results combine a spectral analysis-based optimization scheme with the use of kernels, stochastic processes, and constructive learning algorithms. For initially given non-linear data, we consider optimization-based feature selections. This entails the use of convex sets of kernels in a construction of optimal feature selection via regression algorithms from learning models. Thus, with initial inputs of training data (for a suitable learning algorithm), each choice of a kernel K in turn yields a variety of Hilbert spaces and realizations of features. A novel aspect of our work is the inclusion of a secondary optimization process over a specified convex set of positive definite kernels, resulting in the determination of "optimal" feature representations.

Keywords: positive-definite kernels, reproducing kernel Hilbert space, stochastic processes, frames, machine learning, embedding problems, optimization.

Mathematics Subject Classification: 47N10, 47A52, 47B32, 42A82, 42C15, 62H12, 62J07, 65J20, 68T07, 90C20.

1. INTRODUCTION

Recently, Conditional Mean Embedding (CME) has gained significant attention in various applications [4,5,14–16,18,24,25]. One reason for this is that CME stands at the crossroads of stochastic processes and constructive learning algorithms. In this study, we focus on a novel utilization of CME in analyzing optimization-based selections of positive definite kernels and their associated reproducing kernel Hilbert spaces (as pioneered by Aronszajn [1]). We explore their connections to optimal feature selections through regression algorithms for specific learning models [19, 21, 30].

A function $K : X \times X \to \mathbb{C}$ is said to be a positive definite (p.d.) kernel if, for all $N \in \mathbb{N}$, all $(x_i)_{i=1}^N$ in X and $(c_i)_{i=1}^N$ in $\mathbb{C}$,

$$\sum_{i,j=1}^{N} \overline{c}_i c_j K(x_i, x_j) \geq 0. \tag{1.1}$$

Every p.d. kernel K is associated with a reproducing kernel Hilbert space (RKHS), denoted by $\mathscr{H}_K$, which is the completion of the linear span of functions

$$K_x \coloneqq K(\cdot, x), \quad x \in X \tag{1.2}$$

with respect to the norm

$$\left\| \sum_{i=1}^{N} c_i K(\cdot, x_i) \right\|_{\mathscr{H}_K} = \left(\sum_{i,j=1}^{N} \overline{c}_i c_j K(x_i, x_j) \right)^{1/2}.$$

Moreover, $\mathscr{H}_K$ has the reproducing property:

$$f(x) = \langle K_x, f \rangle_{\mathscr{H}_K}, \quad \forall f \in \mathscr{H}_K, \, \forall x \in X. \tag{1.3}$$

The use of p.d. kernels serves two purposes: Firstly, every p.d. kernel K on $X \times X$ can be considered as the covariance kernel for a centered Gaussian process indexed by X, thus leading to associated probability spaces realized in a generalized path space with σ-algebra and probability measures $\mathbb{P}$. Secondly, selecting a p.d. kernel allows for factorizations via Hilbert space, enabling a wide range of Hilbert space choices that facilitate the realization of features based on initial inputs of training data for a suitable learning algorithm, see e.g., [10].

While earlier approaches fixed a p.d. kernel K in the model, we allow for optimization over carefully selected sets of kernels K within a convex set C of p.d. kernels. Consequently, our "optimal" feature representations depend on a secondary optimization over kernels K within the specified convex set C.

Our present approach to feature selection is motivated in part by machine learning and data mining. These applications are typically driven by the challenges posed by "big data" that needs dimension reduction, which involves transforming data from a high-dimensional space to a lower-dimensional space, by keeping the most correlated data and discarding the less correlated data [23, 26, 27]. The linear case of data transformation encompasses principal component analysis (PCA), while our focus lies in the nonlinear theories offered by kernel theory. Our objective is to achieve adaptive selections of nonlinear mappings that maximize the variance in the data, thereby designing optimal kernels for the given task. Such approaches are particularly valuable for clustering and the selection of highly correlated components within the dataset. For further information on kernel learning, we refer to [2, 20, 28, 29].

The main results of the paper are the following. In Section 3, we present a general framework for computation of optimal solutions, with Theorem 3.4 and Corollary 3.5 offering explicit solution formulas. In Sections 4 and 5, we formulate the notion

of solutions in choices of ambient Hilbert spaces, with explicit formulas given in Theorems 4.4 and 5.9. Associated a priori error bounds are presented in Corollary 4.5. Finally, Section 6 deals with two particular choices of p.d. kernels. In this context, explicit optimizers are presented in Theorem 6.6.

The structure of the paper is the following. In Section 2, we provide an introductory summary of the mathematical concepts that will be essential for our analysis. This section serves as a foundation for the subsequent sections. Following this, we delve into the core of the paper by formulating two distinct versions of the general optimization problem, building upon the concepts introduced in Section 2. These versions involve optimizing over convex sets of kernels denoted as K. In brief outline, the two variants are as follows:

(i) In Section 3, we explore an optimization approach that involves applying the square of the $\mathscr{H}_K$-norm to the optimal feature $f^{\varphi,K}$, which is obtained from the results presented in Theorem 2.1. In particular, this uncovers some cases of non-existence of optimizers.

(ii) In Section 4, we shift our focus to a different criterion of "optimal" and provide a solution formula based on the findings outlined in Corollary 4.2.

Furthermore, in Section 5, our optimization approach is applied to CME. This in turn serves to motivate our affirmative optimization results, especially Theorem 4.4, Corollary 4.5, and Theorem 5.9.

Lastly, in Section 6, we introduce a convex set of p.d. kernels on $\mathbb{R}$. This class includes the Gaussian and Cauchy kernels (6.3), among others. It also allows for a complete characterization for the existence of kernel mean embeddings (Lemmas 5.3 and 6.1). Additionally, we show in Theorem 6.6 that the spectral measures of the associated selfadjoint operators admit explicit representations, as compared to Lemmas 3.1 and 4.3, in the framework of our optimizations.

2. OVERVIEW: SPACES AND OPERATORS

In the discussion that follows, we will rely on various results from analysis and geometry that naturally arise when dealing with p.d. kernels K, feature selections through factorization, and the use of RKHSs $\mathscr{H}_K$ for regression and optimization. While this list of topics is well covered in the literature, the references [7–10, 12] are especially relevant for what we need, and we turn here to some new interdisciplinary directions which are motivated by recent applications.

In summary, the aim of our paper can be illustrated within the following framework: Selecting optimal p.d. kernels K for feature analysis, adapted to large training data

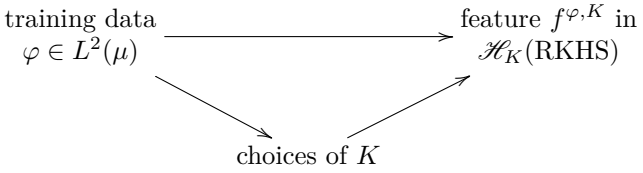
$$\varphi \rightsquigarrow K \rightsquigarrow f \qquad \left.\begin{matrix} \varphi \text{ training data} \\ f \text{ feature selection} \end{matrix}\right\} \text{ depends on choices of p.d. } K.$$

Specifically, we address the following questions:

(i) What is the optimal feature f in $\varphi \rightsquigarrow (\mu, K) \rightsquigarrow f$? Here, μ is a fixed measure on X and $K : X \times X \to \mathbb{C}$ is p.d.

(ii) What is the best choice of K when φ and μ are fixed? How can K be adjusted to optimize the feature f?

The following diagram shows a workflow for learning training data via choices of p.d. kernels, which returns an optimal feature.



The "best" kernel K is determined based on its ability to select the optimal features for a given pair (φ, μ). This notion is formalized as follows:

Theorem 2.1. *Let $T_\mu : \mathscr{H}_K \to L^2(\mu)$ be a bounded linear operator. Then the solution to*

$$f^{\varphi,K} \coloneqq \operatorname{argmin}\left\{\|\varphi - T_\mu f\|^2_{L^2(\mu)} + \alpha \|f\|^2_{\mathscr{H}_K} : f \in \mathscr{H}_K\right\} \tag{2.1}$$

is given by

$$f^{\varphi,K} = T^*_\mu \left(\alpha I + T_\mu T^*_\mu\right)^{-1} \varphi. \tag{2.2}$$

This is a well-known result and can be found in various references such as [4, 13, 17, 22]. An operator-theoretic proof, where T_μ might be unbounded, is given in Section 4.

The solution in (2.1)–(2.2) depends directly on the kernel K. Our first criterion for optimal feature selection corresponds to

$$\max_K \left\|f^{\varphi,K}\right\|^2_{\mathscr{H}_K}. \tag{2.3}$$

To explore this further, one can fix the input φ, and optimize over different choices of (K, μ), that is,

$$\max_{K,\mu} \left\|f^{\varphi,K}\right\|^2_{\mathscr{H}_K}.$$

Alternatively, one can fix the probability measure μ, and optimize over a convex set of admissible kernels K:

Definition 2.2. Given a set X with measure μ, a pair (K, μ) is said to be *admissible* if

$$\mathscr{H}_K \ni K(\cdot, y) \xrightarrow{T_{\mu,K}} K(\cdot, y) \in L^2(\mu), \tag{2.4}$$

extended by linearity, is well defined and closable.

Definition 2.3. Fix μ, let

$$\mathscr{K}(\mu) = \{K : (K, \mu) \text{ is admissible}\}, \tag{2.5}$$
$$\mathscr{K}_b(\mu) = \{K : (K, \mu) \text{ is admissible and } T_{K,\mu} \text{ is bounded}\}. \tag{2.6}$$

Further, for a fixed K, let

$$\mathfrak{M}(K) = \{\mu : (K, \mu) \text{ is admissible}\}.$$

Lemma 2.4. *Suppose (K, μ) is admissible. Then the adjoint operator*

$$T^*_{\mu,K} : L^2(\mu) \to \mathscr{H}_K$$

is specified as follows:

$$\left(T^*_{\mu,K} f\right)(\cdot) = \int K(\cdot, y) f(y) \mu(dy). \tag{2.7}$$

Proof. To verify (2.7), we must show that

$$\left\langle K(\cdot, x), T^*_{\mu,K} f \right\rangle_{\mathscr{H}_K} = \left\langle T_{\mu,K} K(\cdot, x), f \right\rangle_{L^2(\mu)} = \left\langle K(\cdot, x), f \right\rangle_{L^2(\mu)} \tag{2.8}$$

This holds, since

$$\text{LHS}_{(2.8)} = \int K(x, y) f(y) \mu(dy) = \text{RHS}_{(2.8)},$$

by the definition of $T^*_{\mu,K}$ and the reproducing property of $\mathscr{H}_K$. □

To see that $T^*_{\mu,K}$ is well defined, we need the following technical lemma.

Lemma 2.5. *If μ is fixed, then $K \in \mathscr{K}(\mu)$ if and only if*

$$F_\varphi := \int K(\cdot, y) \varphi(y) \mu(dy) \in \mathscr{H}_K, \ \forall \varphi \in L^2(\mu). \tag{2.9}$$

Further, (2.9) *is equivalent to the following: $\forall N \in \mathbb{N}$, $\forall (\alpha_i)_{i=1}^N \subset \mathbb{C}$, $\forall (x_i)_{i=1}^N \subset X$, $\exists C_\varphi < \infty$ with*

$$\left| \sum \alpha_i \int K(x_i, y) \varphi(y) \mu(dy) \right|^2 \le C_\varphi \sum_i \sum_j \overline{\alpha_i} \alpha_j K(x_i, x_j). \tag{2.10}$$

Proof. Assume $F_\varphi \in \mathscr{H}_K$, for all $\varphi \in L^2(\mu)$. Then

$$\text{LHS}_{(2.10)} = \left| \left\langle \sum \alpha_i K_{x_i}, F_\varphi \right\rangle \right|^2 \le \|F_\varphi\|^2_{\mathscr{H}_K} \left\| \sum \alpha_i K_{x_i} \right\|^2_{\mathscr{H}_K} = \text{RHS}_{(2.10))}$$

with $C_\varphi = \|F_\varphi\|^2_{\mathscr{H}_K}$. Here, K_x, $x \in X$, is as in (1.2).

Conversely, if (2.10) holds, then

$$\sum \alpha_k K_{x_i} \longmapsto \sum \overline{\alpha_i} \int K(y, x_i)\, \overline{\varphi(y)} \mu(dy)$$

extends to a unique bounded linear functional l_φ on $\mathscr{H}_K$, and so by Riesz,

$$l_\varphi(f) = \langle \xi, f \rangle_{\mathscr{H}_K}, \quad \forall f \in \mathscr{H}_K.$$

for some $\xi \in \mathscr{H}_K$. Setting $f = K_x$, then

$$\overline{\xi(x)} = l_\varphi(K_x) = \int K(y, x)\, \overline{\varphi(y)} \mu(dy) = \overline{F_\varphi(x)}, \quad \forall x \in X.$$

That is, $F_\varphi = \xi \in \mathscr{H}_K$. □

Lemma 2.6. *Suppose $K(\cdot, \cdot)$ is an integral operator acting on $L^2(X, \mathscr{B}_X, \mu)$, where μ is σ-finite. Then K is p.d. if and only if*

$$\int_X \int_X \overline{\varphi(x)} K(x, y) \varphi(y)\, \mu(dx)\, \mu(dy) \geq 0, \quad \forall \varphi \in L^2(\mu). \tag{2.11}$$

Proof. We first consider the special case where $\varphi = \sum_i \alpha_i \chi_{B_i}$, $\{x_i\}_{i=1}^N \subset X$, $\{B_i\}_{i=1}^N \subset \mathscr{B}_X$. Then (2.11) is equivalent to

$$\sum \sum \overline{\alpha_i} \alpha_j \mu(B_i)\, \mu(B_j)\, K(x_i, x_j) \geq 0,$$

see (1.1). The conclusion follows from this and standard approximations, and we omit the details. □

In the subsequent discussion, we focus on specific cases where explicit expressions can be derived for the optimization problem (2.1)–(2.2). Theorem 2.7 examines the case when the RKHS $\mathscr{H}_K$ is separable and there is a given Parseval frame in $\mathscr{H}_K$. On the other hand, Proposition 2.8 is devoted to the case of atomic measures.

Theorem 2.7. *If $\{f_i\}$ is a Parseval frame (or an orthonormal basis) in $\mathscr{H}_K$, then*

$$K(x, y) = \sum_i f_i(x) \overline{f_i(y)}, \quad \forall (x, y) \in X \times X; \tag{2.12}$$

and

$$\begin{aligned} \left(T_{\mu,K} T_{\mu,K}^* \varphi\right)(x) &= \sum_i \left(\int \overline{f_i(y)} \varphi(y)\, \mu(dy) \right) f_i(x) \\ &= \sum_i \langle \varphi, f_i \rangle_{L^2(\mu)} f_i(x). \end{aligned} \tag{2.13}$$

Moreover,

$$\begin{aligned} f^{\mu,K} &:= \operatorname{argmin} \left\{ \|\varphi - T_{\mu,K} f\|_{L^2(\mu)}^2 + \alpha \|f\|_{\mathscr{H}_K}^2 \right\} \\ &= \sum \left\langle f_i, f^{\mu,K} \right\rangle_{\mathscr{H}_K} f_i. \end{aligned} \tag{2.14}$$

Proof. Assume $\{f_i\}$ is a Parseval frame in $\mathscr{H}_K$. Consider the functions $K_x(\cdot) = K(\cdot, x)$, for $x \in X$. Then

$$K_y(x) = \langle K_x, K_y \rangle_{\mathscr{H}_K} = \sum_i \langle K_x, f_i \rangle \langle f_i, K_y \rangle = \sum_i f_i(x)\overline{f_i(y)},$$

by the reproducing property in $\mathscr{H}_K$ (see (1.3)). This establishes the factorization in (2.12), and the assertions (2.13) and (2.14) follow immediately.

In fact, (2.12) holds if and only if $\{f_i\}$ is a Parseval frame in $\mathscr{H}_K$. However, we will omit the details and refer interested readers to the relevant literature, see e.g., [10]. □

Next, we consider the special case when the measure μ in the pair (K, μ) is atomic.

Proposition 2.8. *Let $\mu = \delta_{x_0}$, then*

$$\left(T_{\delta_{x_0}} T^*_{\delta_{x_0}} f\right)(\cdot) = K(\cdot, x_0) f(x_0),$$
$$\left(T^*_{\delta_{x_0}} T_{\delta_{x_0}} K(\cdot, x)\right)(z) = \overline{K(x_0, z)} K(x_0, x), \quad \forall (x, z) \in X \times X.$$

Proof. One checks that

$$\left(T^*_{\delta_{x_0}} f\right)(\cdot) = \int K(\cdot, y) f(y) \mu(dy) = K(\cdot, x_0) f(x_0),$$

and so

$$T^*_{\delta_{x_0}} T_{\delta_{x_0}} K(\cdot, x) = K(\cdot, x_0) K(x_0, x). \qquad \square$$

To summarize, considering all admissible pairs (K, μ) and the corresponding operators $T_{\mu,K}$ and $T^*_{\mu,K}$, there are two selfadjoint operators (possibly unbounded):

$$L^2(\mu) \xrightarrow{T_{\mu,K} T^*_{\mu,K}} L^2(\mu),$$
$$\mathscr{H}_K \xrightarrow{T^*_{\mu,K} T_{\mu,K}} \mathscr{H}_K.$$

Our focus lies in the spectral decomposition of these operators. Notably, their spectra satisfy that

$$\operatorname{spec}\left(T_{\mu,K} T^*_{\mu,K}\right) \cup \{0\} = \operatorname{spec}\left(T^*_{\mu,K} T_{\mu,K}\right) \cup \{0\}.$$

See e.g., [3, 6]. In Section 4, these general results will be used to provide an operator-theoretic argument for the optimization problem described in (2.1)–(2.2).

3. OPTIMAL FEATURE SELECTIONS

By feature selection we mean a process for reducing the number of input variables, or input data in predictive models. This serves to both reduce the computational cost of modeling, and to improve the performance. It depends on (i) the particular model,

on (ii) choice of statistical measures, and on (iii) data type of both input and output variables. Our present focus is machine learning and optimization via quadratic forms from Section 2, and choice of kernels and associated embeddings into RKHSs. Our results are Theorem 3.4 and Corollary 3.5 below.

Our analysis in the following discussion builds upon the assumptions stated earlier. Specifically, we consider a fixed function φ, which represents the "training data", and a positive σ-finite measure μ. We assume that φ belongs to $L^2(\mu)$.

Our approach to feature selection encompasses both regression using a fixed p.d. kernel K and variations in the choice of p.d. kernels K. Each selection of K leads to a set of admissible features. However, the key lies in identifying a "good" choice of K that produces optimal feature functions $f^{\varphi,K}$, representing distinct and informative features. This distinction is captured by $f^{\varphi,K}$ with large $\|f^{\varphi,K}\|^2_{\mathscr{H}_K}$, indicating a significant variance. Optimal choices of K typically yield successful discrimination by highlighting relevant features that emerge from the specific training data φ. It is important to note that the training data φ remains fixed, while the choice of K determines the features entailed by $\|f^{\varphi,K}\|^2_{\mathscr{H}_K}$.

First, we associate this "variance" $\|f^{\varphi,K}\|^2_{\mathscr{H}_K}$ with the spectral measure of the operator $T_{\mu,K}T^*_{\mu,K}$.

Lemma 3.1. *Given a fixed measure μ and a p.d. kernel $K \in \mathscr{K}(\mu)$, let $f^{\varphi,K}$ be as specified in* (2.2)*. Then*

$$\begin{aligned}\left\|f^{\varphi,K}\right\|^2_{\mathscr{H}_K} &= \left\langle \varphi, T_{\mu,K}T^*_{\mu,K}\left(\alpha + T_{\mu,K}T^*_{\mu,K}\right)^{-2}\varphi\right\rangle_{L^2(\mu)} \\ &= \left\|\left(T_{\mu,K}T^*_{\mu,K}\right)^{1/2}\left(\alpha + T_{\mu,K}T^*_{\mu,K}\right)^{-1}\varphi\right\|^2_{L^2(\mu)} \\ &= \int \frac{x}{(\alpha+x)^2}\left\|Q^{\mu,K}(dx)\,\varphi\right\|^2_{L^2(\mu)},\end{aligned} \tag{3.1}$$

where $Q^{\mu,K}(\cdot)$ is the spectral measure of the operator $T_{\mu,K}T^*_{\mu,K}$, i.e.,

$$T_{\mu,K}T^*_{\mu,K} = \int_0^\infty x\, Q^{\mu,K}(dx).$$

Proof. Let $T := T_{\mu,K}$. Note that

$$TT^*\left(\alpha + TT^*\right)^{-2} : L^2(\mu) \to L^2(\mu)$$

is a bounded operator. We have

$$\begin{aligned}&\left\langle T^*\left(\alpha+TT^*\right)^{-1}\varphi, T^*\left(\alpha+TT^*\right)^{-1}\varphi\right\rangle_{\mathscr{H}_K} \\ &= \left\langle \left(\alpha+TT^*\right)^{-1}\varphi, TT^*\left(\alpha+TT^*\right)^{-1}\varphi\right\rangle_{L^2(\mu)} \\ &= \left\langle \varphi, TT^*\left(\alpha+TT^*\right)^{-2}\varphi\right\rangle_{L^2(\mu)}.\end{aligned}$$

□

Remark 3.2. Note that, if $\left\|T_{\mu,K}T_{\mu,K}^{*}\right\| < \alpha$, then the function $K \longmapsto \left\|f^{\varphi,K}\right\|_{\mathscr{H}_K}^2$ is monotone relative to the order of kernels:

$$K \ll K' \Longleftrightarrow \int \varphi K \varphi \, d\mu \leq \int \varphi K' \varphi \, d\mu.$$

In that case, we need only optimize with respect to the spectral measure associated with the kernel $K \in \mathscr{K}(\mu)$, with μ fixed.

Example 3.3. If $\mu = \delta_{x_0}$ as in Proposition 2.8, then

$$\left(T_{\delta_{x_0}}T_{\delta_{x_0}}^{*}\psi\right)(\cdot) = K(\cdot, x_0)\,\psi(x_0) \in L^2(X, \delta_{x_0}).$$

Moreover,

$$\left\langle \varphi, T_{\delta_{x_0}}T_{\delta_{x_0}}^{*}\psi \right\rangle_{L^2(\delta_{x_0})} = K(x_0, x_0)\,\overline{\varphi(x_0)}\psi(x_0).$$

Similarly, for atomic measures $\mu = \sum_i \alpha_i \delta_{x_i}$, we have

$$\left\langle \varphi, T_\mu T_\mu^{*}\psi \right\rangle_{L^2(\mu)} = \sum_i \sum_j \overline{\alpha_i}\alpha_j K(x_i, x_j)\,\overline{\varphi(x_i)}\psi(x_j)$$

for all $\varphi, \psi \in L^2(\mu)$.

Next, we consider a positive measure μ as in 2.2. By fixing an orthonormal basis (ONB) in the corresponding $L^2(\mu)$, we then arrive at a convex set C_μ of Mercer kernels K. This set C_μ is specified in (3.2). Consequently, these p.d. kernels K and the corresponding RKHSs are determined by the spectral data outlined in (3.4). As a result, the optimal feature vector can be obtained by solving the convex optimization problem for K in C_μ.

Further note that the spectral data used in the case (of Mercer kernels) is a special case of the general structure presented in Lemma 3.1. Indeed, the reader can verify that the optimization algorithm presented below for the case of Mercer kernels extends to more general cases of convex sets of p.d. kernels as per Lemma 3.1.

Theorem 3.4. *Fix μ, and let $K \in \mathscr{K}(\mu)$. Let $\{e_i\}_{i\in\mathbb{N}}$ be an ONB in $L^2(\mu)$, and consider the Mercer kernel*

$$K(x, y) = \sum \lambda_i e_i(x) e_i(y) \tag{3.2}$$

with $\lambda_i > 0$, and $\sum \lambda_i = 1$. In this case,

$$\langle \varphi, T_K T_K^{*}\psi \rangle_{L^2(\mu)} = \sum \lambda_i \langle \varphi, e_i \rangle_{L^2(\mu)} \langle e_i, \psi \rangle_{L^2(\mu)}.$$

Let f^K be the optimal solution as in (2.2)*. Then,*

$$\left\|f^K\right\|_{\mathscr{H}_K}^2 = \sum \frac{\lambda_i}{(\alpha + \lambda_i)^2} \left|\langle \varphi, e_i \rangle\right|^2. \tag{3.3}$$

Moreover, consider the optimization problem:

$$\begin{cases} \max\limits_{(\lambda_i)} \sum \dfrac{\lambda_i}{(\alpha+\lambda_i)^2} c_i, \\ \sum \lambda_i = 1, \quad \lambda_i > 0, \\ \sum c_i = \|\varphi\|^2_{L^2(\mu)}, \quad c_i := |\langle e_i, \varphi\rangle|^2 \geq 0. \end{cases} \tag{3.4}$$

The solution $(\lambda_i^{\max})$ *satisfies that*

$$\frac{\lambda_i^{\max}}{(\alpha+\lambda_i^{\max})^2} = \xi c_i \tag{3.5}$$

for some constant $\xi \in \mathbb{R}_+$.

Proof. The condition in (3.5) follows from an application of the Cauchy–Schwarz inequality.

To show that the solution $(\lambda_i^{\max})$ to (3.5) in fact represents the solution to the optimization problem (3.4), we observe that one term in the l^2 inner product is fixed, so the maximum in (3.4) is attained when quality holds in the corresponding Cauchy–Schwarz inequality.

Further, note that, for every fixed value of the index i, (3.5) is a quadratic equation (see also Figure 1), and the optimal spectral distribution $(\lambda_i^{\max})$ is explicit. The form of the optimal p.d. kernel K then follows by substitution of $(\lambda_i^{\max})$ into (3.2). □

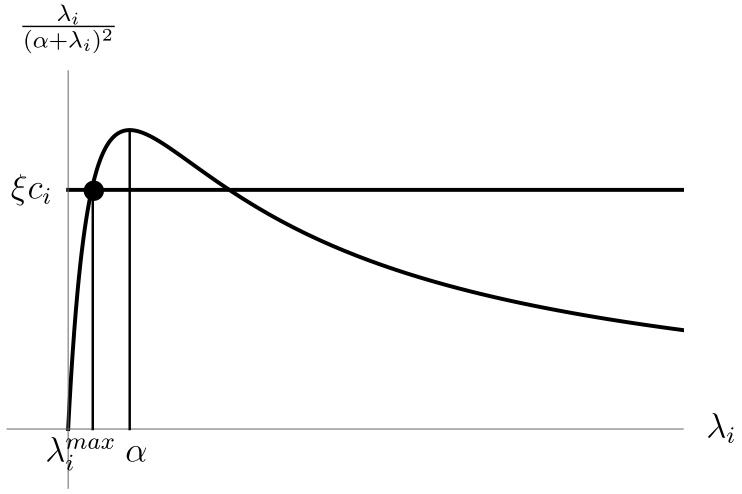


Fig. 1. Spectral distribution in (3.5)

Corollary 3.5. *Consider the finite-dimensional case, i.e., μ is atomic, where $K = \sum_{i=1}^{N} \lambda_i e_i(x) e_i(y)$, with $\{e_i\}_{i=1}^{N}$ an ONB in $L^2(\mu)$. Then the optimization problem*

$$\begin{cases} \max\limits_{(\lambda_i)} \sum\limits_{i=1}^{N} \dfrac{\lambda_i}{(\alpha+\lambda_i)^2} c_i, \\ \sum\limits_{i=1}^{N} \lambda_i = 1, \quad \lambda_i > 0, \\ \sum\limits_{i=1}^{N} c_i = \|\varphi\|^2_{L^2(\mu)}, \quad c_i := |\langle e_i, \varphi\rangle|^2 \geq 0 \end{cases}$$

has solution $(\lambda_i^{\max})$ determined by

$$(\alpha - \lambda_i^{\max})\, c_i = A_N \left(\alpha + \lambda_i^{\max}\right)^3.$$

See Figure 2.

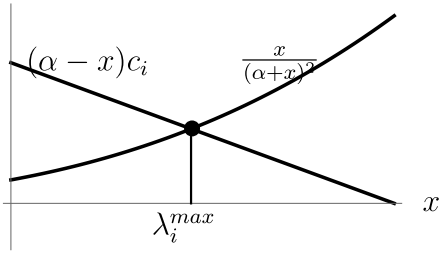


Fig. 2. The solution $(\lambda_i^{\max})$ determined by the intersection of two curves

Moreover, we have

$$\left\|f_N^K\right\|^2_{\mathscr{H}_K} = \sum_{i=1}^{N} \frac{\lambda_i^{\max}}{(\alpha+\lambda_i^{\max})^2} c_i = \sum_{i=1}^{N} \frac{A_N^{2/3} \lambda_i^{\max} c_i^{1/3}}{(\alpha - \lambda_i^{\max})^{2/3}}. \tag{3.6}$$

Proof. Let L be the Lagrangian, where

$$L = \sum_{i=1}^{N} \frac{\lambda_i}{(\alpha+\lambda_i)^2} c_i - A_N \left(\sum_{i=1}^{N} \lambda_i - 1\right).$$

Then,

$$\frac{\partial L}{\partial \lambda_i} = \frac{\alpha^2 - \lambda_i^2}{(\alpha+\lambda_i)^4} c_i - A_N = 0$$

if and only if

$$(\alpha - \lambda_i)\, c_i = A_N \left(\alpha + \lambda_i\right)^3.$$

This yields

$$(\alpha+\lambda_i)^2 = \frac{(\alpha-\lambda_i)^{2/3} c_i^{2/3}}{A_N^{2/3}}$$

so that

$$\sum_{i=1}^{N} \frac{\lambda_i^{\max}}{(\alpha+\lambda_i^{\max})^2} c_i = \sum_{i=1}^{N} \frac{\lambda_i^{\max}}{(\alpha-\lambda_i^{\max})^{2/3} c_i^{2/3}} A_N^{2/3} c_i = \sum_{i=1}^{N} \frac{\lambda_i^{\max}}{(\alpha-\lambda_i^{\max})^{2/3}} A_N^{2/3} c_i^{1/3}$$

which is (3.6). □

4. OPTIMIZATION IN AN AMBIENT HILBERT SPACE

We continue our analysis within the established framework, where the input consists of a fixed measure μ, and a function φ from $L^2(\mu)$ representing the training data. Our focus is on optimal choices for p.d. kernels K that maximize the effectiveness of K-features in the kernel learning process, as discussed in Section 2.

In (2.3), we provide a precise selection criterion for optimal kernels K, along with a solution outlined in Theorem 3.4. In this section, we introduce an alternative criterion (4.5)–(4.6) and derive a solution in Theorem 4.4.

In both cases, when an ONB in $L^2(\mu)$ is chosen, we study the corresponding convex sets of Mercer kernels K as specified in (3.2) and (4.9). Each kernel K is determined by a spectral distribution $\{\lambda_i\}$. The optimization objective is based on (2.2), which incorporates a penalty term in the form of weighted $\mathscr{H}_K$-norm squared, with an assigned parameter α, as shown in (2.1). As a result, we obtain an optimal feature vector f^K for every K.

We further examine the K-variance, measured with the use of the $\mathscr{H}_K$ norm-squared. Two such variance measures are considered, (3.3) and (4.10). In the first case, we observe a singularity blowup when the values of λ_i approach α. In the second case, the dependence on K takes a different form; we show that then the K-variance, as expressed in (4.10), is monotone, as specified in detail in Theorem 4.4 and Corollary 4.5 (spectral *a priori* error-bounds).

Recall that in our general setup for regression optimization, we have arranged that the training data may be represented via an operator T in a Hilbert space. Below we first recall some basic facts from operator theory.

Let $T : \mathscr{H}_1 \to \mathscr{H}_2$ be a closed, densely defined linear operator between Hilbert spaces. On $\mathscr{H}_1 \times \mathscr{H}_2$, define the inner product

$$\left\langle \begin{bmatrix} u_1 \\ u_1 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \right\rangle_{\mathscr{H}_1 \times \mathscr{H}_2} := \alpha \langle u_1, v_1 \rangle_{\mathscr{H}_1} + \langle u_2, v_2 \rangle_{\mathscr{H}_2} \tag{4.1}$$

where α is a positive constant.

Define $W : \mathscr{H}_1 \to \mathscr{H}_1 \times \mathscr{H}_2$ by

$$W(u) = \begin{bmatrix} u \\ Tu \end{bmatrix}, \quad u \in dom(T).$$

This operator is 1–1 and has a bounded inverse on its range, $ran\,(W)$.

Lemma 4.1. *The projection from $\mathscr{H}_1 \times \mathscr{H}_2$ onto $ran\,(W)$ is given by*

$$\begin{bmatrix} \alpha\left(\alpha I_1 + T^*T\right)^{-1} & T^*\left(\alpha I_2 + TT^*\right)^{-1} \\ \alpha T\left(\alpha I_1 + T^*T\right)^{-1} & TT^*\left(\alpha I_2 + TT^*\right)^{-1} \end{bmatrix}. \tag{4.2}$$

Proof. When $\alpha = 1$, the block matrix in (4.2) represents the projection from the direct sum $\mathscr{H}_1 \oplus \mathscr{H}_2$ onto the graph of the operator T. This result can be found in e.g., [11, Corollary 1.55]. The case of $\alpha \neq 1$ is a straightforward variation of the argument presented therein. □

Corollary 4.2. *Let $T : \mathscr{H}_1 \to \mathscr{H}_2$ be as above. Then, for all $v \in \mathscr{H}_2$, we have*

$$\begin{aligned} u^* &= \operatorname{argmin}\left\{\alpha\|u\|^2_{\mathscr{H}_1} + \|Tu - v\|^2_{\mathscr{H}_2} : u \in \mathscr{H}_1\right\} \\ &= \operatorname{argmin}\left\{\|Wu - (0,v)\|^2_{\mathscr{H}_1\times\mathscr{H}_2} : u \in \mathscr{H}_1\right\} \\ &= T^*\left(\alpha I_2 + TT^*\right)^{-1} v. \end{aligned} \tag{4.3}$$

Proof. By the definition of W, we have

$$\alpha\|u\|^2_{\mathscr{H}_1} + \|Tu - v\|^2_{\mathscr{H}_2} = \left\|Wu - \begin{bmatrix} 0 \\ v \end{bmatrix}\right\|^2_{\mathscr{H}_1\times\mathscr{H}_2}.$$

By Lemma 4.1, the projection of $\begin{bmatrix} 0 \\ v \end{bmatrix}$ onto $ran\,(W)$ is given by the vector

$$\begin{bmatrix} T^*\left(\alpha I_2 + TT^*\right)^{-1} v \\ TT^*\left(\alpha I_2 + TT^*\right)^{-1} v \end{bmatrix},$$

which is equal to

$$Wu^* = \begin{bmatrix} u^* \\ Tu^* \end{bmatrix},$$

for a unique u^* in $\mathscr{H}_1$. This gives (4.3). □

Now, return to optimal feature selections. Fix μ, and consider kernels $K \in \mathscr{K}(\mu)$, see Definition 2.3. Define $W_K : \mathscr{H}_K \to \mathscr{H}_K \times L^2(\mu)$ by

$$W_K h = \begin{bmatrix} h \\ T_{K,\mu} h \end{bmatrix}.$$

The inner product on $\mathscr{H}_K \times L^2(\mu)$ is as in (4.1), with parameter $\alpha > 0$.

Fix $\varphi \in L^2(\mu)$, then we get a unique $f^{\varphi,K} \in \mathscr{H}_K$, such that $W_K f^{\varphi,K}$ is the projection of

$$\begin{bmatrix} 0 \\ \varphi \end{bmatrix} \in \mathscr{H}_K \times L^2(\mu) \quad \text{onto} \quad ran\,(W_K).$$

That is,

$$f^{\varphi,K} = T_{K,\mu}^{*}\left(\alpha I_{L^2} + T_{K,\mu}T_{K,\mu}^{*}\right)^{-1}\varphi \tag{4.4}$$

by Corollary 4.2.

This motivates the optimization criterion:

$$\max_{K\in\mathscr{K}(\mu)}\left\{\left\|W_K f^{\varphi,K}\right\|^2_{\mathscr{H}_K\times L^2(\mu)}\right\} \tag{4.5}$$

$$\Updownarrow$$

$$\max_{K\in\mathscr{K}(\mu)}\left\{\alpha\left\|f^{\varphi,K}\right\|^2_{\mathscr{H}_K} + \left\|T_{K,\mu}f^{\varphi,K}\right\|^2_{L^2(\mu)}\right\} \tag{4.6}$$

Below is a modification of Lemma 3.1.

Lemma 4.3. *With μ, K fixed, $K\in\mathscr{K}(\mu)$, let $f^{\varphi,K}$ be as specified in* (4.4)*. Then*

$$\left\|W_K f^{\varphi,K}\right\|^2_{\mathscr{H}_K\times L^2(\mu)} = \int \frac{x}{\alpha+x}\left\|Q^{\mu,K}(dx)\,\varphi\right\|^2_{L^2(\mu)}, \tag{4.7}$$

where $Q^{K,\mu}(dx)$ is the spectral measure of the operator $T_{K,\mu}T_{K,\mu}^{*}$. Especially,

$$\left\|W_K f^{\varphi,K}\right\|^2_{\mathscr{H}_K\times L^2(\mu)} \le \|\varphi\|^2_{L^2(\mu)}. \tag{4.8}$$

Proof. Let $T := T_{K,\mu}$, then

$$\begin{aligned}
\left\|W_K f^{\varphi,K}\right\|^2_{\mathscr{H}_K\times L^2(\mu)} &= \alpha\left\|f^{\varphi,K}\right\|^2_{\mathscr{H}_K} + \left\|Tf^{\varphi,K}\right\|^2_{L^2(\mu)}\\
&= \alpha\left\|T^{*}\left(\alpha+TT^{*}\right)^{-1}\varphi\right\|^2_{\mathscr{H}_K} + \left\|\left(\alpha+TT^{*}\right)^{-1}TT^{*}\varphi\right\|^2_{L^2(\mu)}\\
&= \int\left(\frac{\alpha x}{(\alpha+x)^2} + \frac{x^2}{(\alpha+x)^2}\right)\left\|Q^{\mu,K}(dx)\,\varphi\right\|^2_{L^2(\mu)}\\
&= \int\frac{x}{\alpha+x}\left\|Q^{\mu,K}(dx)\,\varphi\right\|^2_{L^2(\mu)}.
\end{aligned}$$

Finally, (4.8) follows from the fact that $W_K f^{\varphi,K}$ is the projection of $\begin{bmatrix}0\\ \varphi\end{bmatrix}\in\mathscr{H}_K\times L^2(\mu)$ onto $ran\,(W)$. □

Next, we state an analog of Theorem 3.4.

Theorem 4.4. *Fix μ, and let $K\in\mathscr{K}(\mu)$. Let $\{e_i\}_{i\in\mathbb{N}}$ be an ONB in $L^2(\mu)$, and consider the p.d. kernel*

$$K(x,y) = \sum_{i\in\mathbb{N}}\lambda_i e_i(x)e_i(y) \tag{4.9}$$

with $\lambda_i > 0$. Let $f^{\varphi,K}$ be the optimal solution as in (4.4)*. Then,*

$$\left\|W_K f^{\varphi,K}\right\|^2_{\mathscr{H}_K\times L^2(\mu)} = \sum\frac{\lambda_i}{\alpha+\lambda_i}\left|\langle\varphi,e_i\rangle\right|^2. \tag{4.10}$$

Proof. See the proof of Theorem 3.4. □

Corollary 4.5. *Let W_K and $f^{\varphi,K}$ be as in Theorem 4.4, and assume K is bounded. Let $\lambda_- = \inf\{\lambda_i\}$, $\lambda_+ = \sup\{\lambda_i\}$.*

(i) *The following hold:*

$$\frac{\lambda_-}{\alpha+\lambda_-}\left\|\varphi\right\|^2_{L^2(\mu)} \leq \left\|W_K f^{\varphi,K}\right\|^2_{\mathscr{H}_K\times L^2(\mu)} \leq \frac{\lambda_+}{\alpha+\lambda_+}\left\|\varphi\right\|^2_{L^2(\mu)}. \tag{4.11}$$

(ii) *Equivalently, the approximation error satisfies*

$$\frac{\alpha}{\alpha+\lambda_+}\left\|\varphi\right\|^2_{L^2(\mu)} \leq err \leq \frac{\alpha}{\alpha+\lambda_-}\left\|\varphi\right\|^2_{L^2(\mu)}. \tag{4.12}$$

(iii) *By increasing λ_-, $W_K f^{\varphi,K}$ approximates $(0,\varphi)$ in $\mathscr{H}_K \times L^2(\mu)$ arbitrarily well.*

Proof. Notice that the function $f(x) = \frac{x}{\alpha+x}$ in (4.7) is strictly increasing in $(0,\infty)$, so that (4.11) follows from (4.10). The other assertions are immediate. □

Remark 4.6. The difference between the two feature selection methods in Sections 3 and 4 is as follows.

Fix a measure μ, and consider $K \in \mathscr{K}(\mu)$, i.e., all admissible kernels. Let $\mathscr{H}_K$ be the associated RKHS. In both cases, for a given $\varphi \in L^2(\mu)$, the best feature vector in $\mathscr{H}_K$ is the same

$$f^{K,\varphi} = T^*_{K,\mu}\left(\alpha + T_{K,\mu}T^*_{K,\mu}\right)^{-1}\varphi.$$

See (2.2) and (4.4).

However, the criteria for optimization over kernels K are different:

$$\text{Section 3:}\quad \max_{K\in\mathscr{K}(\mu)}\left\{\left\|f^{K,\varphi}\right\|^2_{\mathscr{H}_K}\right\}, \tag{4.13}$$

$$\text{Section 4:}\quad \max_{K\in\mathscr{K}(\mu)}\left\{\alpha\left\|f^{K,\varphi}\right\|^2_{\mathscr{H}_K} + \left\|T_{K,\mu}f^{K,\varphi}\right\|^2_{L^2}\right\}. \tag{4.14}$$

As discussed at the beginning of (4), the vector

$$\left(f^{K,\varphi}, T_{K,\mu}f^{K,\varphi}\right)$$

is the projection of $(0,\varphi)$ in $\mathscr{H}_K \times L^2$ onto the graph of $T_{K,\mu}$. Thus, (4.14) is the norm squared of the projected vector and the corresponding optimization makes use of Hilbert space geometry.

5. APPLICATIONS TO CME

A key feature in *conditional mean embedding* (CME) involves the analysis of systems of random variables, and conditional distributions, which take values in suitable choices of RKHSs, often in infinite dimensions. Hence, conditional expectations, and relative transition operators, will entail choices of p.d. kernels, typically one for each random variable under consideration. The implementation of kernel embedding of

distributions (also called the *kernel mean* or mean map) yields nonparametric outcomes in which a probability distribution is represented as an element of an RKHS. In diverse applications, the use of CME has served as useful tools in for example, problems of sequentially optimizing conditional expectations for objective functions. In such settings, both the conditional distribution and the objective function, while fixed, are assumed to be unknown.

The setting for CME is as follows:

Let X, Y be random variables on a probability space $(\Omega, \mathscr{C}, \mathbb{P})$, taking values in sets A, B, respectively, and has joint measure

$$\mu\left(S_1 \times S_1\right) = \mathbb{P}\left(X^{-1}\left(S_1\right) \cap Y^{-1}\left(S_2\right)\right)$$

for all $S_1 \times S_2 \in \mathscr{B}_A \times \mathscr{B}_B$, the product σ-algebra.

Denote by μ_X, μ_Y the corresponding marginal measures, and let $\mu_{Y|x}$ be the conditional measure defined as

$$\mu_{Y|x}\left(S\right) = \mathbb{P}\left(Y^{-1}\left(S\right) \mid X = x\right)$$

for all $S \in \mathscr{B}_B$ and $x \in A$.

Assume further that K, L are given p.d. kernels on A, B with RKHSs $\mathscr{H}_K$, $\mathscr{H}_L$, respectively.

Lemma 5.1. *For every $x \in A$, set*

$$\pi(x) := \mathbb{E}\left[L\left(\cdot, Y\right) \mid X = x\right] = \int L\left(\cdot, y\right) d\mu_{Y|x}\left(y\right) \tag{5.1}$$

Then, for all $f \in \mathscr{H}_L$, it holds that

$$\left\langle \pi(x), f\right\rangle_{\mathscr{H}_L} = \int \left\langle L\left(\cdot, y\right), f\right\rangle d\mu_{Y|x}\left(y\right) = \mathbb{E}\left[f\left(Y\right) \mid X = x\right].$$

Proof. The integral on the RHS in formula (5.1) is an extension of (2.7) from Lemma 2.4. Moreover, the proof of the lemma follows the ideas in Section 2. □

Definition 5.2. The map $x \mapsto \pi(x)$ in (5.1) is referred to as the kernel mean embedding (KME) of the conditional expectation $\mathbb{E}\left[\,\cdot \mid X = x\right]$, also known as the conditional mean embedding (CME).

We will provide a brief overview of the existence of KME as follows:

Consider a measurable space $(X, \mathscr{B}_X)$, where $\mathscr{B}_X$ is a given σ-algebra of subsets of X. We focus on measures μ on $(X, \mathscr{B}_X)$, with particular emphasis on the case where μ is a probability measure, i.e., $\mu(x) = 1$. Let K be a p.d. kernel on $X \times X$ and let $\mathscr{H}_K$ be the associated RKHS.

Lemma 5.3. *Assuming that*

$$\int_X \int_X \mu(dx)\, K(x,y) \mu(dy) < \infty, \tag{5.2}$$

then the following function

$$T_K(\mu)(\cdot) \coloneqq \int K(\cdot, y)\, \mu(dy) \tag{5.3}$$

belongs to $\mathscr{H}_K$, and

$$\|T_K(\mu)\|^2_{\mathscr{H}_K} = \int_X \int_X \mu(dx)\, K(x,y) \mu(dy). \tag{5.4}$$

Proof. We refer to the cited references for details. □

Corollary 5.4. *Let $W : \Omega \to X$ be a random variable on a given probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and let $\mu = \mu_W = \mathbb{P} \circ W^{-1}$. Assume K is a p.d. kernel on $X \times X$ with the property* (5.2)*. Then, we have the representation*

$$\langle f, T_K(\mu_W) \rangle_{\mathscr{H}_K} = \mathbb{E}(f \circ W)$$

for all $f \in \mathscr{H}_K$.

Now we return to our analysis of CME.

Lemma 5.5. *As in Lemma 5.1, consider $\pi(x)$ for $x \in A$, as per the definition* (5.1) *in Lemma 5.1. Then*

(i) *$\pi(x) \in \mathscr{H}_L$ if and only if*

$$\mathbb{E}[L(Y,Y) \mid X = x] := \mu_{Y|x} L \mu_{Y|x} < \infty, \tag{5.5}$$

where

$$\mu_{Y|x} L \mu_{Y|x} = \iint L(y_1, y_2)\, d\mu_{Y|x}(y_1)\, d\mu_{Y|x}(y_2).$$

(ii) *$\pi \in L^2(A, \mu_X) \otimes \mathscr{H}_L$ if and only if*

$$\mathbb{E}[K(Y,Y)] = \int \left(\mu_{Y|x} L \mu_{Y|x}\right) d\mu_X(x) < \infty.$$

In that case, setting $\widetilde{\pi}(f)(x) := \langle \pi(x), f \rangle_{\mathscr{H}_L}$, then

$$\int \|\widetilde{\pi}(f)\|^2_{\mathscr{H}_K}\, d\mu_X \le \mathbb{E}[K(Y,Y)]\, \|f\|^2_{\mathscr{H}_L}$$

for all $f \in \mathscr{H}_L$.

Proof. Consider the filter of finite measurable partitions $\mathscr{P}(B)$ of the measurable space $(B, \mathscr{B})$, i.e., $\{A_i\}_{i=1}^N$ for some $N < \infty$, with $A_i \in \mathscr{B}$, $A_i \cap A_j = \emptyset$ if $i \neq j$, and $\bigcup_i A_i = B$, then

$$\sum_{i=1}^{N} L(\cdot, y_i)\, \mu_{Y|x}(A_i) \in \mathscr{H}_K \tag{5.6}$$

with

$$\left\| \sum\nolimits_{i=1}^{N} L(\cdot, y_i)\, \mu_{Y|x}(A_i) \right\|_{\mathscr{H}_K}^2 = \sum\nolimits_i \sum\nolimits_j \mu_{Y|x}(A_i)\, L(y_i, y_j)\, \mu_{Y|x}(A_j). \tag{5.7}$$

Since L is assumed measurable, the right-hand side of (5.7) has a limit, as we pass to the limit of the filter of all measurable partitions $\mathscr{P}(B)$, see (5.6), and the limit is well defined and finite if and only if (5.5) holds. This follows from the following computation:

$$\iint \mu_{Y|x}(dy_1)\, L(y_1, y_2)\, \mu_{Y|x}(dy_2) = \mathbb{E}[L(Y, Y) \mid X = x].$$

But since we have "=" in the identity (5.7) for all finite partitions, it follows that (5.5) holds if and only if the integral on the right-hand side in (5.1) is convergent with its values in $\mathscr{H}_L$.

The second part of the lemma is immediate. □

Remark 5.6. The setting of the lemma is a fixed a p.d. kernel L and a measure space $(B, \mathscr{B})$. We have L defined on $B \times B$ and assumed measurable w.r.t. the corresponding product σ-algebra. The key idea behind the justification of the RKHS $\mathscr{H}_L$ valued integral $\pi(x)$ in (5.1) is a rigorous justification of a limit of an approximation by finite sums in $\mathscr{H}_L$, and the limit with respect to the RKHS norm in $\mathscr{H}_L$. This is doable as per our discussion, but the limit will be indexed by a filter of partitions of the measure space $(B, \mathscr{B})$. And the limit is with respect to refinement within the filter of partitions, where refinement defined by recursive subdivision, i.e., subdivisions of one partition are creating a finer partition. Note that the reasoning involves the same kind of limit which is used in the justification of general Ito isometries, and Ito integrals for Gaussian processes.

Question 5.7. Assume $\pi \in L^2(A, \mu_X) \otimes \mathscr{H}_L$. What is the best approximation to choice of CME μ from an $\mathscr{H}_L$-valued RKHS?

One option in the literature is to approximate π from $\mathscr{H}_K \otimes \mathscr{H}_L$. More generally, one may start from an $\mathscr{B}(\mathscr{H}_L)$-valued p.d. kernel $S : A \times A \to \mathscr{B}(\mathscr{H}_L)$, i.e.,

$$\sum_{i,j=1}^{N} \langle u_i, S(x_i, x_j)\, u_j \rangle_{\mathscr{H}_L} \geq 0 \tag{5.8}$$

$\forall (x_i)_{i=1}^N \subset A$, $\forall (u_i)_{i=1}^N \subset \mathscr{H}_L$, and $\forall N \in \mathbb{N}$.

Let $\mathscr{H}_S$ be the Hilbert completion of the set span $\{S(\cdot, x)\, u : x \in A, u \in \mathscr{H}_L\}$ with respect to the inner product

$$\left\langle \sum S(\cdot, x_i)\, u_i, \sum S(\cdot, x_j)\, v_j \right\rangle_{\mathscr{H}_S} := \sum_{i,j} \langle u_i, S(x_i, x_j)\, v_j \rangle_{\mathscr{H}_L}.$$

Then $\mathscr{H}_S$ is an RKHS with the following reproducing property:

For all $F \in \mathscr{H}_S$, $x \in A$ and $u \in \mathscr{H}_L$, we have

$$\langle u, F(x) \rangle_{\mathscr{H}_L} = \langle S(\cdot, x)\, u, F \rangle_{\mathscr{H}_L}.$$

Remark 5.8. In the special case $\mathscr{H}_S = \mathscr{H}_K \otimes \mathscr{H}_L$, we have

$$S(x, y) = K(x, y) I_{\mathscr{H}_L},$$

where K is the scalar valued p.d. kernel of $\mathscr{H}_K$, and $I_{\mathscr{H}_L}$ denotes the identity operator on $\mathscr{H}_L$.

Theorem 5.9. *Assume S is compatible with the marginal distribution of X, then we have*

$$f^{\pi,S} := \operatorname{argmin} \left\{ \|T_S f - \pi\|^2_{L^2(A) \otimes \mathscr{H}_L} + \alpha \|f\|^2_{\mathscr{H}_S} \right\} \tag{5.9}$$

$$= T_S^* \left(\alpha + T_S T_S^*\right)^{-1} \pi. \tag{5.10}$$

Then, we may apply the methods from Section 3 to the problem:

$$\max_S \left\| f^{\pi,S} \right\|^2_{\mathscr{H}_S}.$$

Proof. Illustrating the versatility of Hilbert space operators, the reader will be able to fill in the argument for this formula (5.10), and its implications, following the general framework presented in Sections 2 and 3 above. □

6. A NEW CONVEX SET OF P.D. KERNELS

In this section, we introduce a convex set of positive definite kernels on $\mathbb{R}$, denoted as G_1, also referred to as stationary kernels. These kernels can be represented as $K_g(x, y) = g(x - y)$, where g is the Fourier transform of a probability measure μ on $\mathbb{R}$.

Specifically, given K_g along with its associated RKHS $\mathscr{H}_{K_g}$, and a Borel measure ρ on $\mathbb{R}$, Lemma 6.1 shows that the existence of kernel mean embedding (KME), as discussed in Lemma 5.3, is completely characterized by the Fourier transform of the measure ρ. Furthermore, under suitable conditions, the integral operator $T_\lambda : \mathscr{H}_{K_g} \to L^2(d\lambda)$, where λ denotes Lebesgue measure, is densely defined and closeable. In Theorem 6.6, we show that the spectral measures of the selfadjoint operators $T_\lambda T_\lambda^*$ have explicit representations, in comparison to Lemmas 3.1 and 4.3.

Importantly, this class of kernels includes highly popular choices such as the Gaussian and Cauchy kernels, as illustrated in Example 6.3. These kernels are widely used in machine learning, statistical modeling, and other related areas.

To begin, we establish the necessary notations and definitions. Let $\mathcal{M}(\mathbb{R})$ be the set of all Borel measures on $\mathbb{R}$, and $\mathcal{M}_1(\mathbb{R})$ be the subset of probability measures. For all $\rho \in \mathcal{M}(\mathbb{R})$, let

$$\widehat{\rho}(\xi) = \int_{\mathbb{R}} e^{i\xi x} d\rho(x)$$

denote the Fourier transform.

Consider the following convex set of stationary kernels

$$G_1 = \left\{ \mathbb{R} \times \mathbb{R} \xrightarrow{K_g} \mathbb{C} : K_g(x,y) = g(x-y),\ g = \widehat{\mu},\ \mu \in \mathcal{M}_1(\mathbb{R}) \right\}.$$

Lemma 6.1. *Fix $K_g \in G_1$, and let $\mathscr{H}_{K_g}$ be the corresponding RKHS. Then, for all $\rho \in \mathcal{M}(\mathbb{R})$,*

$$g * d\rho := \int_{\mathbb{R}} K_g(\cdot, y)\, d\rho(y) \in \mathscr{H}_{K_g}$$
$$\Updownarrow$$
$$\int_{\mathbb{R}} |\widehat{\rho}(\xi)|^2 \, d\mu(\xi) < \infty.$$

Proof. Assume $g * d\rho \in \mathscr{H}_{K_g}$, then

$$\begin{aligned}
\|g * d\rho\|^2_{\mathscr{H}_{K_g}} &= \iint \langle K_g(\cdot, y), K_g(\cdot, z) \rangle_{\mathscr{H}_{K_g}} \, d\rho(y)\, d\rho(z) \\
&= \iint g(y-z)\, d\rho(y)\, d\rho(z) \\
&= \int \left(\iint e^{i\xi(y-z)} d\rho(y)\, d\rho(z) \right) d\mu(\xi) \\
&= \int |\widehat{\rho}(\xi)|^2 \, d\mu(\xi) < \infty.
\end{aligned}$$

Conversely, suppose $C := \int |\widehat{\rho}(\xi)|^2 \, d\mu(\xi) < \infty$. Then, for all $\sum c_k K_g(\cdot, x_k)$, we have

$$\begin{aligned}
\sum c_k K_g(\cdot, x_k) \longmapsto \left| \sum c_k (g * d\rho)(x_k) \right|^2 &= \left| \sum c_k \int \int e^{i\xi(x_k - y)} d\rho(y)\, d\mu(\xi) \right|^2 \\
&\le \int \left| \sum c_k e^{i\xi x_k} \right| |\widehat{\rho}(\xi)| \, d\mu(\xi)^2 \\
&\le \int \left| \sum c_k e^{i\xi x_k} \right|^2 d\mu(\xi) \int |\widehat{\rho}(\xi)|^2 \, d\mu(\xi) \\
&= C \cdot \sum_k \sum_l \overline{c_k} c_l K_g(x_k, x_l).
\end{aligned}$$

It follows that $g * d\rho \in \mathscr{H}_{K_g}$ by density and Riesz's theorem. □

Fix $K_g \in G_1$, and let $\mathscr{H}_{K_g}$ be the RKHS. Let $d\lambda$ denote the Lebesgue measure on $\mathbb{R}$. Suppose $\left\{\varphi \in L^2(d\lambda) : \widehat{\varphi} \in L^2(\mu)\right\}$ is dense in $L^2(d\lambda)$. Then, the operator

$$T_\lambda : \mathscr{H}_{K_g} \to L^2(d\lambda), \quad T_\lambda\left(\sum_i c_i K_g(\cdot, x_i)\right) = \sum_i c_i K_g(\cdot, x_i) \tag{6.1}$$

is densely defined and closable, and its adjoint is given by

$$T_\lambda^* : L^2(d\lambda) \to \mathscr{H}_{K_g}, \quad T_\lambda^*(\varphi) = g * \varphi, \ \forall \varphi \in dom\left(T_\lambda^*\right) \tag{6.2}$$

where $dom\left(T_\lambda^*\right) = \left\{\varphi \in L^2(d\lambda) : \widehat{\varphi} \in L^2(\mu)\right\}$.

Corollary 6.2. *For all $\varphi \in dom\left(T_\lambda^*\right)$, we have*

$$\left\|T_\lambda^* \varphi\right\|_{\mathscr{H}_{K_g}}^2 = \int |\widehat{\varphi}(\xi)|^2 \, d\mu(\xi).$$

Proof. This follows from Lemma 6.1 by setting $d\rho = g d\lambda$, and $\widehat{\varphi}$ is the L^2-Fourier transform of φ. □

Example 6.3. Consider the following two p.d. kernels on $\mathbb{R}$:

$$K_1(x, y) = e^{-|x-y|}, \quad K_2(x, y) = e^{-\frac{1}{2}(x-y)^2}. \tag{6.3}$$

Note that

$$g_1(x) = e^{-|x|} = \int e^{i\xi x} d\mu_1(\xi), \quad d\mu_1(\xi) = \frac{1}{\pi}\frac{1}{1+\xi^2} d\xi,$$
$$g_2(x) = e^{-\frac{1}{2}x^2} = \int e^{i\xi x} d\mu_2(\xi), \quad d\mu_2(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} d\xi.$$

Moreover, for K_1, if $\varphi \in L^2(\mathbb{R})$, then

$$\|g_1 * \varphi\|_{\mathscr{H}_{K_1}}^2 = \int_{\mathbb{R}} \frac{|\widehat{\varphi}(\xi)|^2 \, d\xi}{1+\xi^2} = \left\langle \varphi, \left(1 - (d/dx)^2\right)^{-1} \varphi \right\rangle_{L^2(\mathbb{R})}.$$

In other words, the RKHS is the RKHS from the Green's function for $1 - (d/dx)^2$, or $1 - \Delta$ in $\mathbb{R}^n$, $n > 1$.

Given $K_g \in G_1$, the convolution $\varphi \mapsto g * \varphi \in \mathscr{H}_{K_g}$ may be extended to measures or distributions.

Lemma 6.4. *Let $K_g \in G_1$, and $\mathscr{H}_{K_g}$ be the corresponding RKHS. Then,*

$$g(x - \cdot) = g * \delta_x \in \mathscr{H}_{K_g}$$

and

$$g * \delta_x' \in \mathscr{H}_{K_g} \iff \int |\xi|^2 \mu(d\xi) < \infty. \tag{6.4}$$

Note (6.4) *is satisfied for K_2 but not for K_1 in Example 6.3.*

Proof. Using $g * \delta_x = g(x - \cdot)$, we have $g * \delta_x \in \mathscr{H}_{K_g}$ and

$$\|g(x-\cdot)\|^2_{\mathscr{H}_{K_g}} = \langle g(x-\cdot), g(x-\cdot)\rangle_{\mathscr{H}_{K_g}} = g(x-x) = g(0) = 1.$$

Equivalently, $\delta_x \leftrightarrow \widehat{\delta}_x(\xi) = e^{i\xi x}$, $\xi \in \mathbb{R}$, and

$$\int |\widehat{\delta_x}|^2 d\mu = \int \left|e^{ix\xi}\right|^2 \mu(d\xi) = \mu(\mathbb{R}) = g(0) = 1.$$

Similarly, $\delta'_x \leftrightarrow \widehat{\delta'_x}(\xi) = i\xi e^{ix\xi}$. Thus $g * \delta'_x \in \mathscr{H}_{K_g}$ if and only if $\int |\xi|^2 \mu(d\xi) < \infty$. □

Remark 6.5. Given $K_g(x,y) = g(x-y)$, where $g(x) = \int e^{i\xi x}\mu(d\xi)$, and μ is a finite positive Borel measure on $\mathbb{R}$, the reproducing property of $\mathscr{H}_{K_g}$ below may be verified using Fourier-inversion:

$$\langle g(x-\cdot), \varphi * g\rangle_{\mathscr{H}_K} = (\varphi * g)(x), \quad \forall x \in \mathbb{R}.$$

Proof. Indeed, we have

$$\begin{aligned}\langle g(x-\cdot), \varphi * g\rangle_{\mathscr{H}_K} &= \int \overline{e^{i\xi x}}\widehat{\varphi}(\xi)\mu(d\xi) \\ &= \int_{\mathbb{R}} e^{i\xi x}\widehat{\varphi * g}(\xi)d\xi = (\varphi * g)(x).\end{aligned}$$ □

Theorem 6.6. *Fix $K_g \in G_1$, and let $\mathscr{H}_{K_g}$ be the RKHS. Let $T_\lambda : \mathscr{H}_{K_g} \to L^2(d\lambda)$ be as in* (6.1)*, where $d\lambda$ denotes the Lebesgue measure on $\mathbb{R}$. Let $f^{g,\lambda}$ be the solution in Theorem 2.1, i.e.,*

$$f^{g,\lambda} = T^*_\lambda (\alpha + T_\lambda T^*_\lambda)^{-1}\varphi$$

where $\varphi \in L^2(d\lambda)$ is fixed. Then, by the L^2-Fourier transform, we have

$$\widehat{T_\lambda f^{g,\lambda}} = \left[T_\lambda T^*_\lambda (\alpha + T_\lambda T^*_\lambda)^{-1}\varphi\right]^\wedge = \frac{\widehat{g}}{\alpha + \widehat{g}}\widehat{\varphi}.$$

Moreover, the optimal selections from Lemmas 3.1 and 4.3, respectively, admit the following explicit spectral representations:

$$\left\|f^{g,\lambda}\right\|^2_{\mathscr{H}_{K_g}} = \int \frac{\widehat{g}(\lambda)}{(\alpha + \widehat{g}(\lambda))^2}\left|\widehat{\varphi}(\lambda)\right|^2 d\lambda, \tag{6.5}$$

$$\left\|W_{K_g} f^{g,\lambda}\right\|^2_{\mathscr{H}_{K_g}\times L^2(d\lambda)} = \int \frac{\widehat{g}(\lambda)}{\alpha + \widehat{g}(\lambda)}\left|\widehat{\varphi}(\lambda)\right|^2 d\lambda. \tag{6.6}$$

Proof. By the definition of T^*_λ from (6.2), it follows that $T_\lambda T^*_\lambda \varphi = g * \varphi \in L^2(d\lambda)$, and so $\widehat{T_\lambda T^*_\lambda}\varphi = \widehat{g}\widehat{\varphi}$.

It follows from this, that

$$\left\|f^{g,\lambda}\right\|^2_{\mathscr{H}_{K_g}} \underset{(3.1)}{=} \left\|(T_\lambda T^*_\lambda)^{1/2}(\alpha + T_\lambda T^*_\lambda)^{-1}\varphi\right\|^2_{L^2(d\lambda)}$$

$$= \int \frac{\widehat{g}(\lambda)}{(\alpha + \widehat{g}(\lambda))^2} |\widehat{\varphi}(\lambda)|^2 \, d\lambda$$

and on the other hand,

$$\left\| W_{K_g} f^{g,\lambda} \right\|^2_{\mathscr{H}_{K_g} \times L^2(d\lambda)} \underset{(4.7)}{=} \left\langle \varphi, T_\lambda T_\lambda^* \left(\alpha + T_\lambda T_\lambda^*\right)^{-1} \varphi \right\rangle_{L^2(d\lambda)}$$
$$= \int \frac{\widehat{g}(\lambda)}{\alpha + \widehat{g}(\lambda)} |\widehat{\varphi}(\lambda)|^2 \, d\lambda. \qquad \square$$

Example 6.7. Let K_1 and K_2 be the p.d. kernels from (6.3). The formulas in (6.5)–(6.6) take explicit forms, summarized in Table 1.

Table 1
The p.d. kernels K_1 and K_2

$K_g = g(x-y)$	$g = \widehat{\mu}$	μ	$\frac{\widehat{g}}{(\alpha+\widehat{g})^2}$	$\frac{\widehat{g}}{\alpha+\widehat{g}}$
$K_1 = e^{-\lvert x-y\rvert}$	$g_1 = e^{-\lvert x\rvert}$	$d\mu_1 = \frac{1}{1+\xi^2} d\xi$	$\frac{1+\xi^2}{(1+\alpha(1+\xi^2))^2}$	$\frac{1}{1+\alpha(1+\xi^2)}$
$K_2 = e^{-\frac{1}{2}(x-y)^2}$	$g_2 = e^{-\frac{1}{2}x^2}$	$d\mu_2 = e^{-\frac{1}{2}\xi^2} d\xi$	$\frac{e^{\frac{1}{2}x^2}}{\left(1+\alpha e^{\frac{1}{2}x^2}\right)^2}$	$\frac{1}{1+\alpha e^{\frac{1}{2}\xi^2}}$

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc. **68** (1950), no. 3, 337–404.

[2] J. Cerviño, J.A. Bazerque, M. Calvo-Fullana, A. Ribeiro, *Multi-task reinforcement learning in reproducing kernel Hilbert spaces via cross-learning*, IEEE Trans. Signal Process. **69** (2021), 5947–5962.

[3] N. Dunford, J.T. Schwartz, *Linear Operators. Part II*, Wiley Classics Library, John Wiley & Sons, Inc., New York, 1988.

[4] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, M. Pontil, *Conditional mean embeddings as regressors*, [in:] *Proceedings of the 29th International Coference on International Conference on Machine Learning*, Omnipress, Madison, WI, USA, 2012, ICML'12, 1803–1810.

[5] D. He, J. Cheng, K. Xu, *High-dimensional variable screening through kernel-based conditional mean dependence*, J. Statist. Plann. Inference **224** (2023), 27–41.

[6] P. Jorgensen, F. Tian, *Non-commutative Analysis*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2017.

[7] P. Jorgensen, F. Tian, *Decomposition of Gaussian processes, and factorization of positive definite kernels*, Opuscula Math. **39** (2019), no. 4, 497–541.

[8] P. Jorgensen, F. Tian, *Realizations and factorizations of positive definite kernels*, J. Theoret. Probab. **32** (2019), no. 4, 1925–1942.

[9] P. Jorgensen, F. Tian, *Sampling with positive definite kernels and an associated dichotomy*, Adv. Theor. Math. Phys. **24** (2020), no. 1, 125–154.

[10] P. Jorgensen, F. Tian, *Reproducing kernels: harmonic analysis and some of their applications*, Appl. Comput. Harmon. Anal. **52** (2021), 279–302.

[11] P. Jorgensen, F. Tian, *Infinite-dimensional Analysis – Operators in Hilbert Space; Stochastic Calculus via Representations, and Duality Theory*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2021.

[12] P. Jorgensen, F. Tian, *Reproducing kernels and choices of associated feature spaces, in the form of L^2-spaces*, J. Math. Anal. Appl. **505** (2022), no. 2, 125535.

[13] P.E.T. Jorgensen, M.-S. Song, J. Tian, *Positive definite kernels, algorithms, frames, and approximations*, (2021), arXiv:2104.11807.

[14] I. Klebanov, I. Schuster, T.J. Sullivan, *A rigorous theory of conditional mean embeddings*, SIAM J. Math. Data Sci. **2** (2020), no. 3, 583–606.

[15] T. Lai, Z. Zhang, Y. Wang, *A kernel-based measure for conditional mean dependence*, Comput. Statist. Data Anal. **160** (2021), Paper no. 107246.

[16] T. Lai, Z. Zhang, Y. Wang, L. Kong, *Testing independence of functional variables by angle covariance*, J. Multivariate Anal. **182** (2021), Paper no. 104711.

[17] Y.J. Lee, C.A. Micchelli, J. Yoon, *On multivariate discrete least squares*, J. Approx. Theory **211** (2016), 78–84.

[18] G. Lever, J. Shawe-Taylor, R. Stafford, C. Szepesvári, *Compressed conditional mean embeddings for model-based reinforcement learning*, AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (2016), 1779–1787.

[19] D.K. Lim, N.U. Rashid, J.G. Ibrahim, *Model-based feature selection and clustering of RNA-seq data for unsupervised subtype discovery*, Ann. Appl. Stat. **15** (2021), no. 1, 481–508.

[20] C.-K. Lu, P. Shafto, *Conditional deep Gaussian processes: Multi-fidelity kernel learning*, Entropy **2021**, 23(11), 1545.

[21] E. Mehmanchi, A. Gómez, O.A. Prokopyev, *Solving a class of feature selection problems via fractional 0–1 programming*, Ann. Oper. Res. **303** (2021), 265–295.

[22] C.A. Micchelli, M. Pontil, Q. Wu, D.-X. Zhou, *Error bounds for learning the kernel*, Anal. Appl. (Singap.) **14** (2016), no. 6, 849–868.

[23] P. Niyogi, S. Smale, S. Weinberger, *A topological view of unsupervised learning from noisy data*, SIAM J. Comput. **40** (2011), no. 3, 646–663.

[24] J. Park, K. Muandet, *A measure-theoretic approach to kernel conditional mean embeddings*, arXiv:2002.03689.

[25] S. Ray Chowdhury, R. Oliveira, F. Ramos, *Active learning of conditional mean embeddings via Bayesian optimisation*, [in:] J. Peters, D. Sontag (eds) *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, PMLR, 2020, volume 124 of Proceedings of Machine Learning Research, 1119–1128.

[26] S. Smale, Y. Yao, *Online learning algorithms*, Found. Comput. Math. **6** (2006), 145–170.

[27] S. Smale, D.-X. Zhou, *Geometry on probability spaces*, Constr. Approx. **30** (2009), 311–323.

[28] P. Xu, Y. Wang, X. Chen, Z. Tian, *COKE: communication-censored decentralized kernel learning*, J. Mach. Learn. Res. **22** (2021), Paper no. 196.

[29] Y. Zhang, Y.-C. Chen, *Kernel smoothing, mean shift, and their learning theory with directional data*, J. Mach. Learn. Res. **22** (2021), Paper no. 154.

[30] P. Zhao, L. Lai, *Minimax rate optimal adaptive nearest neighbor classification and regression*, IEEE Trans. Inform. Theory **67** (2021), no. 5, 3155–3182.

Palle E.T. Jorgensen (corresponding author)
palle-jorgensen@uiowa.edu

Department of Mathematics
The University of Iowa
Iowa City, IA 52242–1419, U.S.A.

Myung-Sin Song
msong@siue.edu

Department of Mathematics and Statistics
Southern Illinois University Edwardsville
Edwardsville, IL 62026, U.S.A.

James Tiang
jft@ams.org

Mathematical Reviews
416–4th Street Ann Arbor, MI 48103–4816, U.S.A.

Received: June 3, 2023.
Accepted: July 5, 2023.