Optimal Learning Rates for the Conditional Mean Embedding

Anonymous Author(s)
Affiliation
Address
email

Abstract

We address the consistency of a ridge regression estimate of the conditional mean embedding (CME), which is an embedding of the conditional distribution of Y given X into a target reproducing kernel Hilbert space $\mathcal{H}_Y$. The CME allows us to take conditional expectations of target RKHS functions, and has been employed in nonparametric causal and Bayesian inference. Our results focus on the misspecified setting, where the target CME is in the space of Hilbert-Schmidt operators acting from an input interpolation space between $\mathcal{H}_X$ and L_2, to $\mathcal{H}_Y$. We derive a novel and adaptive statistical learning rate for the empirical CME estimator under this setting. We further establish a lower bound on the learning rate, which reveals that when the target CME operator is smooth, the obtained upper bound is optimal.

1 Introduction

Approximation of the conditional expectation operator is a central issue in the statistical learning community, and many approaches have been proposed [41, 16, 17, 18]. Given random variables X and Y with a joint distribution $\mathbb{P}$, the conditional expectation operator for a function f is defined

$$[Pf](x) := \mathbb{E}_{\mathbb{P}}(f(Y)|X = x) = \int_E f(y) d\mathbb{P}(Y|X = x).$$

Conventional parametric models to approximate P often involve density estimation and expensive numerical analysis. Hence, recent studies attempt to explore a new framework to approximate P via kernel methods. Specifically, given kernels k_X and k_Y with corresponding reproducing kernel Hilbert space $\mathcal{H}_X$ and $\mathcal{H}_Y$ for X and Y respectively, we may define the conditional meaning embedding (CME) operator as $F_*(x) := \mathbb{E}_{\mathbb{P}}(k_Y(\cdot, Y)|X = x)$, and we may employ the reproducing property to obtain $[Pf](x) = \langle f, F_*(x)\rangle_{\mathcal{H}_Y}$ for any $f \in \mathcal{H}_Y$. The advantage of the CME framework is that it allows the straightforward evaluation of conditional expectations of any function in $\mathcal{H}_Y$. The CME framework has been applied successfully to many learning problems such as probabilistic inference [32], reinforcement learning [25, 12] and causal inference [22, 28].

Despite these successful applications, there have been two main challenges in establishing a rigorous theory of conditional mean embeddings. The first challenge, remarkably, has been in establishing a principled and sufficiently general definition of the conditional mean embedding itself. The CME was originally introduced as an operator mapping from $\mathcal{H}_X$ to $\mathcal{H}_Y$ [9, 34]. This definition has the benefit of elegance, and of a straightforward expression in terms of feature covariances and cross-covariances. A disadvantage is that the definition requires the conditional mean $\mathbb{E}(g(Y)|X = \cdot) \in \mathcal{H}_X, \forall g \in \mathcal{H}_Y$: we call this the *well-specified scenario*. This strong assumption may be violated in practice (see [14]

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

and [10, Section 3.1] for illustrations), and significantly restricts the class of distributions on which we can define a CME.

An alternative approach, due to [11], is to express the conditional mean embedding as the solution of a least-squares regression problem in a vector valued RKHS [4, 5]. In subsequent work, [26] establish a rigorous measure-theoretic definition of the conditional mean embedding as a square integrable function on $\mathcal{H}_Y$, which is the definition we will use in the present work. Both [11, 26] connect this CME definition to the original operator-mapping definition by means of a surrogate loss, which upper bounds the regression loss. A direct connection remained elusive until the work of [23], who show that in the well-specified case, the CME can be arbitrarily well approximated by a Hilbert-Schmidt operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$, thus connecting the operator-theoretic and measure-theoretic definitions. The connection in the misspecified case remains to be established, however.

The second challenge has been in obtaining consistency results and the optimal learning rates for empirical estimates of the CME. An early consistency analysis of the sample estimator, due to [33], requires very strong smoothness assumptions. A more refined analysis, due to [11], attains the optimal $O(\log n/n)$ learning rate for the sample estimator, but only in the case where $\mathcal{H}_Y$ is finite dimensional. For the infinite dimensional RKHS, [28] and [26] establish consistency in the well-specified case, with learning rates of $O(n^{-1/6})$ and $O(n^{-1/4})$. Subsequent work [39] yields a further improvement in the statistical learning rate, and consistency in the misspecified setting, building on recent developments in regularized regression under misspecification [8]. A limitation of [39] is that it requires an explicit relation between the smoothness of the target CME and the size of the RKHS. In particular, when the kernel has slow eigenvalue decay (as in the case of Matérn kernels, for example), the setting of their result is very close to the well-specified scenario. Finally, to our knowledge, there is presently no result establishing a matching lower bound for the CME learning rate. Hence, whether the obtained upper rate is optimal remains unknown.

In the present work, we address the above mentioned challenges. Building on [26, 23] and the interpolation space theory results of [38], we establish an **operator-theoretic definition of the CME in the misspecified case**, and show that the target CME can be well approximated by Hilbert-Schmidt operators. In doing so, we avoid the well-specified requirement, and simply assume the target CME operator to be defined in respect of an intermediate interpolation space, strictly larger than the original RKHS. Moreover, we show that the space of Hilbert-Schmidt operators defined on the interpolation space can approximate the CME operator with arbitrary accuracy in operator norm (see Theorem 2).

Next, we establish **consistency and convergence rates of the CME sample estimator in the misspecified setting**, building on [8, 39]. In particular, under certain benign conditions, we obtain the optimal $O(\log n/n)$ learning rate up to logarithmic factor. This matches with the current optimal analysis from [11] without the restrictive assumption of finite dimensional $\mathcal{H}_Y$. Thanks to our operator-theoretic definition of the CME, and unlike [39], we do not require an a-priori relation between the rate of kernel eigenvalue decay and the smoothness of the conditional mean operator (i.e., our results apply generally in the misspecified setting). Finally, in Theorem 5, we provide a novel **lower bound on the CME learning rate**, which reveals that the obtained upper rate is optimal in the setting of a smooth CME operator.

2 Background

Throughout the paper, we consider two random variables X, Y defined on the measurable space $(E, \mathcal{F}_E)$ where E is a second countable locally compact Hausdorff space and $\mathcal{F}_E$ its Borel σ-field. We let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space with expectation operator $\mathbb{E}$. Let π and ν be the pushforward of $\mathbb{P}$ under X and Y respectively, i.e., $X \sim \pi$ and $Y \sim \nu$. To be rigorous, we use the Markov kernel $p : E \times \mathcal{F}_E \to \mathbb{R}_+$ to define the conditional distribution introduced before as

$$\mathbb{P}[Y \in A | X = x] = \int_A p(x, dy),$$

for all $x \in E$ and the events $A \in \mathcal{F}_E$. In addition, we denote the space of real-valued Lebesgue square integrable functions on $(E, \mathcal{F}_E)$ with respect to π as $L_2(E, \mathcal{F}_E, \pi; \mathbb{R}) = L_2(\pi)$ and similarly

to ν as $L_2(E, \mathcal{F}_E, \nu; \mathbb{R}) = L_2(\nu)$. We denote B for a separable Banach space with norm $\|\cdot\|_B$ and H for a separable real Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$. We write $\mathcal{L}(B, B')$ as the Banach space of bounded linear operators from B to another Banach space B', equipped with the operator norm $\|\cdot\|_{B \to B'}$. When $B = B'$, we simply write $\mathcal{L}(B)$ instead. We also let $L_p(\Omega, \mathcal{F}, \pi; B)$ denote the space of strongly $\mathcal{F} - \mathcal{F}_B$ measurable and Bochner p-integrable functions $f : \Omega \to B$ for $1 < p \leq \infty$. $H \otimes H'$ denotes the tensor product of Hilbert spaces H, H'. The Hilbert space $H \otimes H'$ is the completion of the algebraic tensor product with respect to the inner product $\langle x_1 \otimes x_1', x_2 \otimes x_2' \rangle_{H \otimes H'} = \langle x_1, x_2 \rangle_H \langle x_1', x_2' \rangle_{H'}$ for $x_1, x_2 \in H$ and $x_1', x_2' \in H'$. Finally, we denote the p-Schatten class $S_p(H, H')$ to be the space of all compact operators C from H to H' such that $\|C\|_{S_p(H,H')} := \left\|(\sigma_i(C))_{i \in J}\right\|_{\ell_p}$ is finite. Here $\|(\sigma_i(C))_{i \in J}\|_{\ell_p}$ is the ℓ_p sequence space norm of the sequence of the strictly positive singular values of C indexed by the countable set J. For $p = 2$, $S_2(H, H')$ is the Hilbert space of Hilbert-Schmidt operators from H to H'.

Reproducing Kernel Hilbert Spaces, Covariance Operators: Consider a random variable X defined on $(E, \mathcal{F}_E)$. We let $k_X : E \times E \to \mathbb{R}$ be a symmetric and positive definite kernel function and $\mathcal{H}_X$ be a vector space of $E \to \mathbb{R}$ functions, endowed with a Hilbert space structure via an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_X}$. k_X is a reproducing kernel of $\mathcal{H}_X$ if and only if: 1. $\forall x \in X, k_X(\cdot, x) \in \mathcal{H}_X$; 2.$\forall x \in X$ and $\forall f \in \mathcal{H}_X, f(x) = \langle f, k_X(x, \cdot) \rangle_{\mathcal{H}_X}$. A space $\mathcal{H}_X$ which possesses a reproducing kernel is called a reproducing kernel Hilbert space (RKHS)[1]. We denote the canonical feature map of $\mathcal{H}_X$ as $\phi_X(x) = k_X(\cdot, x)$. Similarly for $\mathcal{H}_Y$, we have $\phi_Y(y)$.

We now introduce some facts about the interplay between $\mathcal{H}_X$ and $L_2(\pi)$ which has been extensively studied by [29, 30],[7] and [38]. We first define the not necessarily injective embedding $I_\pi : \mathcal{H}_X \to L_2(\pi)$, mapping a function $f \in \mathcal{H}_X$ to its π-equivalence class $[f]_\pi$. This is well-defined, Hilbert-Schmidt, and the Hilbert-Schmidt norm satisfies

$$\|I_\pi\|_{S_2(\mathcal{H}_X, L_2(\pi))} = \|k_X\|_{L_2(\pi)} := \left(\int_E k_X(x, x) \mathrm{d}\pi(x) \right)^{1/2} < \infty$$

Moreover, the adjoint operator $S_\pi := I_\pi^* : L_2(\pi) \to \mathcal{H}_X$ is an integral operator with respect to the kernel k_X, i.e. for $f \in L_2(\pi)$ and $x \in E$ we have

$$(S_\pi f)(x) = \int_X k_X(x, x') f(x') \, \mathrm{d}\pi(x')$$

Next, we define the self-adjoint and positive semi-definite integral operators

$$L_X := I_\pi S_\pi : L_2(\pi) \to L_2(\pi) \quad \text{and} \quad C_{XX} := S_\pi I_\pi : \mathcal{H}_X \to \mathcal{H}_X$$

These operators are trace class and their trace norms satisfy

$$\|L_X\|_{S_1(L_2(\pi))} = \|C_{XX}\|_{S_1(\mathcal{H}_X)} = \|I_\pi\|^2_{\mathcal{H}_X \to L_2(\pi)} = \|S_\pi\|^2_{L_2(\pi) \to \mathcal{H}_X}.$$

Vector-valued RKHS We also give a brief overview of the vector-valued reproducing kernel Hilbert space (vRKHS). Since the construction is very technical, we only introduce the basic properties and refer the reader to [4] and [5] for more details.

Definition 1. *Let E be a nonempty set and H be a real Hilbert space. Let function $K : E \times E \to \mathcal{L}(H)$ be an operator valued positive-semidefinite (psd) kernel such that $K(x, x') = K(x', x)^*$ for all $x, x' \in E$, and for all $x_1, \ldots, x_n \in E$ and $h_i, h_j \in H$,*

$$\sum_{i,j=1}^{n} \langle h_i, K(x_i, x_j) h_j \rangle_H \geq 0.$$

Fix K, $x \in E$, and $h \in H$, $[K_x h](\cdot) := K(\cdot, x)h$ defines a function from E to H. We now consider

$$\mathcal{G}_{\text{pre}} := \operatorname{span}\{K_x h \mid x \in E, h \in H\}$$

with inner product on $\mathcal{G}_{\text{pre}}$ by linearly extending the expression

$$\langle K_x h, K_{x'} h' \rangle_{\mathcal{G}} := \langle h, K(x, x') h' \rangle_H. \tag{1}$$

Let $\mathcal{G}$ be the completion of $\mathcal{G}_{\text{pre}}$ with respect to this inner product. We call $\mathcal{G}$ the vRKHS induced by the kernel K. The space $\mathcal{G}$ is a Hilbert space consisting of functions from E to H with the reproducing property

$$\langle F(x), h\rangle_H = \langle F, K_x h\rangle_{\mathcal{G}}\,, \tag{2}$$

for all $F \in \mathcal{G}, h \in H$ and $x \in E$. For all $F \in \mathcal{G}$ we obtain

$$\|F(x)\|_H \leq \|K(x,x)\|^{1/2}\|F\|_{\mathcal{G}}, \quad x \in E.$$

Since the inner product given by Eq. (1) implies that K_x is a bounded operator for all $x \in E$. For all $F \in \mathcal{G}$ and $x \in E$, we can rewrite the reproducing property Eq. (2) as $F(x) = K_x^* F$. The linear operators $K_x : H \to \mathcal{G}$ and $K_x^* : \mathcal{G} \to H$ are bounded with

$$\|K_x\| = \|K_x^*\| = \|K(x,x)\|^{1/2}$$

and we have $K_x^* K_{x'} = K\left(x, x'\right), x, x' \in E$.

Recall X and Y are two random variables defined on $(E, \mathcal{F}_E)$ with marginal distribution π and ν. We let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be the RKHS with kernel k_X and k_Y respectively. In the following, we will denote $\mathcal{G}$ as the vRKHS induced by the kernel $K : E \times E \to \mathcal{L}(\mathcal{H}_Y)$ with

$$K(x,x') := k_X(x,x')\mathrm{Id}_{\mathcal{H}_Y}, x, x' \in E.$$

An important property of $\mathcal{G}$ is that elements in $\mathcal{G}$ with kernel K are isometric to Hilbert-Schmidt operators between $\mathcal{H}_X$ and $\mathcal{H}_Y$.

Theorem 1. *Let $\mathcal{H}_X$ and $\mathcal{H}_Y$ be real-valued RKHS with kernel k_X and k_Y respectively. Recall we denote the tensor product space between $\mathcal{H}_X$ and $\mathcal{H}_Y$ as $\mathcal{H}_Y \otimes \mathcal{H}_X$. For $f_Y \in \mathcal{H}_Y$ and $g_X \in \mathcal{H}_X$, define the map Θ as*

$$\left[\Theta\left(f_Y \otimes g_X\right)\right](x) := g_X(x) f_Y = \left(f_Y \otimes g_X\right)\phi_X(x).$$

We then have that Θ defines an isometric isomorphism between $\mathcal{H}_Y \otimes \mathcal{H}_X$ and $\mathcal{G}$.

A proof of Theorem 1 can be found in [23, Theorem 4.4]. The isometric isomorphism Θ induces the operator reproducing property stated below

Corollary 1. *Recall we denote $S_2(\mathcal{H}_X, \mathcal{H}_Y)$ as the 2-Schatten class of compact operators from $\mathcal{H}_X$ to $\mathcal{H}_Y$. For every function $F \in \mathcal{G}$ there exists an operator $C := \Theta^{-1}(F) \in S_2(\mathcal{H}_X, \mathcal{H}_Y)$ such that*

$$F(x) = C\phi_X(x) \in \mathcal{H}_Y,$$

for all $x \in E$ with $\|C\|_{S_2(\mathcal{H}_X,\mathcal{H}_Y)} = \|F\|_{\mathcal{G}}$ and vice versa. Conversely, for any pair $F \in \mathcal{G}$ and $C \in S_2(\mathcal{H}_X, \mathcal{H}_Y)$, we have $C = \Theta^{-1}(F)$ as long as $F(x) = C\phi_X(x)$.

The proof of Corollary 1 is a simple extension of Lemma 15 in [6] and Corollary 4.5 in [23]. Corollary 1 shows that the vRKHS $\mathcal{G}$ is generated via the space of Hilbert-Schmidt operators $S_2(\mathcal{H}_X, \mathcal{H}_Y)$

$$\mathcal{G} = \left\{F : E \to \mathcal{H}_Y | F = C\phi_X(\cdot), C \in S_2(\mathcal{H}_X, \mathcal{H}_Y)\right\}.$$

Conditional Mean Embedding: A particular advantage of kernel methods is its convenience of operating probability distributions, see [24, 27] for examples. This is through the so called kernel mean embedding [1, 31, 13]. Suppose X has distribution π on E, assuming the integrability condition $\int_X \sqrt{k_X(x,x)}d\pi(x) < \infty$, we define the kernel mean embedding $\mu_X(\cdot) = \int_X k_X(\cdot, x)d\pi(x)$. It is easy to show that for each $f \in \mathcal{H}_X$, $\int_X f(x)d\pi(x) = \langle f, \mu_X\rangle_{\mathcal{H}_X}$. Replacing π with the conditional distribution, we obtain the kernel conditional mean embedding as defined in [26].

Definition 2. *Given $p(x, dy)$, the conditional distribution of Y given X, we define the $\mathcal{H}_Y$-valued conditional mean embedding (CME) as*

$$F_*(x) := \int_E \phi_Y(y)p(x,dy) = \mathbb{E}\left[\phi_Y(Y)|X = x\right] \in L_2(E, \mathcal{F}_E, \pi; \mathcal{H}_Y) \tag{3}$$

By reproducing property, we have $\mathbb{E}[f_Y(Y)|X=x] = \langle f_Y, F_*(x)\rangle_{\mathcal{H}_Y}, \forall f_Y \in \mathcal{H}_Y$ and $x \in E$.

The approximation of F_* is a key concept in kernel methods. Assuming C_{XX} is injective, [34, 9, 10] show that F_* has a closed form expression via

$$F_*(x) = C_{YX} C_{XX}^{\dagger} \phi_X(x),$$

where $C_{YX} = \mathbb{E}(\phi_Y \otimes \phi_X)$ and $C^{\dagger}$ denotes the pseudoinverse of C. In addition, estimation of F_* can also be interpreted as the following optimization problem [11, 26, 39]

$$\hat{C}_{Y|X,\lambda} := \underset{C \in S_2(\mathcal{H}_X, \mathcal{H}_Y)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|\phi_Y(y_i) - C\phi_X(x_i)\|_{\mathcal{H}_Y}^2 + \lambda \|C\|_{HS}^2, \tag{4}$$

$\hat{F}_\lambda(x) := \hat{C}_{Y|X,\lambda}\phi_X(x)$, where λ is the regularization parameter and $\|\cdot\|_{HS}$ is the Hilbert-Schmidt operator norm. Implicit in the construction, however, is the assumption that

$$\mathbb{E}\left[f_Y(Y)|X=\cdot\right] \in \mathcal{H}_X, \forall f_Y \in \mathcal{H}_Y. \tag{5}$$

This is a strong assumption [10, 14] and constitutes the so-called *well-specified* case $F_*(\cdot) \in \mathcal{G}$. Our next task is to characterize the Hilbert spaces used to define the misspecified case, and to define the CME in this setting.

Interpolation Space, misspecified CME: We now review the results of [38, 8] that set out the eigendecompositions of L_X and C_{XX}, and apply these in constructing the interpolation spaces used for the misspecified setting. We define an index set I, a non-increasing sequence $(\mu_i)_{i\in I} > 0$, and a family $(e_i)_{i\in I} \in \mathcal{H}_X$, such that $([e_i]_\pi)_{i\in I}$ is an orthonormal basis (ONB) of $L_2(\pi)$ and $(\mu_i^{1/2} e_i)_{i\in I}$ is an ONB of $\mathcal{H}_X$, we have

$$L_X = \sum_{i\in I} \mu_i \langle \cdot, [e_i]_\pi \rangle_{L_2(\pi)} [e_i]_\pi, \qquad C_{XX} = \sum_{i\in I} \mu_i \langle \cdot, \mu_i^{\frac{1}{2}} e_i \rangle_{\mathcal{H}_X} \mu_i^{\frac{1}{2}} e_i.$$

For $\alpha \geq 0$, we define the α-interpolation space [38] by

$$[\mathcal{H}]_X^\alpha := \left\{ \sum_{i\in I} a_i \mu_i^{\alpha/2} [e_i]_\pi : (a_i)_{i\in I} \in \ell_2(\mathbb{N}) \right\} \subseteq L_2(\pi),$$

equipped with the α-power norm

$$\left\| \sum_{i\in I} a_i \mu_i^{\alpha/2} [e_i]_\pi \right\|_{[\mathcal{H}]_X^\alpha} := \left\| (a_i)_{i\in I} \right\|_{\ell_2(\mathbb{N})} = \left(\sum_{i\in I} a_i^2 \right)^{1/2}.$$

For $(a_i)_{i\in I} \in \ell_2(\mathbb{N})$, the α-interpolation space becomes a Hilbert space with inner product defined as

$$\left\langle \sum_i a_i (\mu_i^{\alpha/2} [e_i]_\pi), \sum_i b_i (\mu_i^{\alpha/2} [e_i]_\pi) \right\rangle_{[\mathcal{H}]_X^\alpha} = \sum_i a_i b_i.$$

Moreover, $\left(\mu_i^{\alpha/2} [e_i]_\pi\right)_{i\geq 1}$ forms an ONB of $[\mathcal{H}]_X^\alpha$ and consequently $[\mathcal{H}]_X^\alpha$ is a separable Hilbert space. In the following, we use the abbreviation $\|\cdot\|_\alpha := \|\cdot\|_{[\mathcal{H}]_X^\alpha}$. For $\alpha = 0$ we have $[\mathcal{H}]_X^0 = \overline{\operatorname{ran} I_\pi} \subseteq L_2(\pi)$ with $\|\cdot\|_0 = \|\cdot\|_{L_2(\pi)}$. Moreover, for $\alpha = 1$ we have $[\mathcal{H}]_X^1 = \operatorname{ran} I_\pi$ and $[\mathcal{H}]_X^1$ is isometrically isomorphic to the closed subspace $(\ker I_\pi)^\perp$ of $\mathcal{H}_X$ via I_π, i.e. $\|[f]_\pi\|_1 = \|f\|_{\mathcal{H}_X}$ for $f \in (\ker I_\pi)^\perp$. For $0 < \beta < \alpha < 1$, and assuming I_π is injective, we have

$$[\mathcal{H}]_X \subset [\mathcal{H}]_X^\alpha \subset [\mathcal{H}]_X^\beta \subset [\mathcal{H}]_X^0 = L_2(\pi).$$

Finally, if $\sum_i \mu_i^\alpha e_i^2(x) < \infty$, $\mathcal{H}_X^\alpha$ is an RKHS with corresponding kernel defined as

$$k_X^\alpha(\cdot, x) = \sum_i \mu_i^\alpha e_i(\cdot) e_i(x). \tag{6}$$

For a function $f_X^\alpha = \sum_i a_i \mu_i^{\alpha/2} e_i \in \mathcal{H}_X^\alpha$, we recover the reproducing property via

$$\langle f_X^\alpha, k_X^\alpha(\cdot, x)\rangle_{\mathcal{H}_X^\alpha} = \langle \sum_i a_i \mu_i^{\alpha/2} e_i, \sum_i \mu_i^\alpha e_i(\cdot) e_i(x)\rangle_{\mathcal{H}_X^\alpha} = \sum_i a_i \mu_i^{\alpha/2} e_i(x) = f_X^\alpha(x).$$

Remark 1. *We point out the distinction between $\mathcal{H}_X^\alpha$ and $[\mathcal{H}]_X^\alpha$. In particular, $\mathcal{H}_X^\alpha$ denotes the interpolating RKHS consisting of continuous functions, while $[\mathcal{H}]_X^\alpha$ is the interpolating Hilbert space, where elements are defined through equivalence class. Under our construction, we have $\mathcal{H}_X^\alpha \subseteq [\mathcal{H}]_X^\alpha$.*

We use the interpolation space to address the CME in the context of misspecified case. In particular, we consider the vRKHS $\mathcal{G}^\alpha$ induced by the kernel $K^\alpha : E \times E \to \mathcal{L}(\mathcal{H}_Y)$ via

$$K^\alpha(x, x') := k_X^\alpha(x, x')\mathrm{Id}_{\mathcal{H}_Y}.$$

Assuming $\mathcal{H}_X^\alpha$ is an RKHS and utilizing the same analysis from Theorem 1 and Corollary 1, we have the following result.

Corollary 2. *For every function $F^\alpha \in \mathcal{G}^\alpha$, there exists an operator $C^\alpha := (\Theta^\alpha)^{-1}(F^\alpha) \in S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$ such that*

$$F^\alpha(x) = C^\alpha \phi_X^\alpha(x) \in \mathcal{H}_Y,$$

for all $x \in E$ with $\|C^\alpha\|_{S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)} = \|F^\alpha\|_{\mathcal{G}^\alpha}$ and vice versa. Conversely, for any pair $F^\alpha \in \mathcal{G}^\alpha$ and $C^\alpha \in S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$, we have $C^\alpha = (\Theta^\alpha)^{-1}(F^\alpha)$ as long as $F^\alpha(x) = C^\alpha \phi_X^\alpha(x)$. We have that Θ^α defines an isometric isomorphism between $\mathcal{H}_Y \otimes \mathcal{H}_X^\alpha$ and $\mathcal{G}^\alpha$.

Corollary 2 shows that the vRKHS $\mathcal{G}^\alpha$ is generated via $S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$

$$\mathcal{G}^\alpha = \{F^\alpha : E \to \mathcal{H}_Y | F^\alpha = C^\alpha \phi_X^\alpha(\cdot), C^\alpha \in S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y), \phi_X^\alpha(x) = k_X^\alpha(\cdot, x)\}. \tag{7}$$

Throughout the paper, we assume that $F_* \in \mathcal{G}^\alpha$, indicating $\mathbb{E}(f_Y(Y)|X = \cdot) \in \mathcal{H}_X^\alpha, \forall f_Y \in \mathcal{H}_Y$. Comparing to Eq. (5), we can see that our assumption is strictly weaker in the sense that $\mathcal{H}_X \subset \mathcal{H}_X^\alpha$, for some $\alpha < 1$.

3 Approximation of CME with Interpolation Space

In this section, we demonstrate that the operator F_* can be arbitrarily well approximated in L_2 norm by vectors in the vRKHS $\mathcal{G}^\alpha$, without imposing any smoothness assumption on F_*.

Before we present our theoretical analysis, we first state some technical assumptions on the previously defined RKHS and kernels.

1. The RKHS $\mathcal{H}_X$ and $\mathcal{H}_Y$ are separable. This is satisfied if E is a Polish space and k_X and k_Y are continuous [36].
2. $k_X(\cdot, x)$ and $k_Y(\cdot, y)$ are $\mathcal{F}_\mathbb{R}$ measurable for all $x, y \in E$.
3. The feature maps of k_X and k_Y satisfy $\mathbb{E}\left[\|\phi_X(X)\|_{\mathcal{H}_X}^2\right] < \infty$ and $\mathbb{E}\left[\|\phi_Y(Y)\|_{\mathcal{H}_Y}^2\right] < \infty$, respectively. This can be trivially satisfied if $\sup_{x,y \in E} k_X(x, x), k_Y(y, y) < \infty$.
4. Let $C_0(E)$ be the space of continuous functions vanishing at infinity. Then $\mathcal{H}_X, \mathcal{H}_Y \in C_0(E)$. This can be satisfied if $k_X(x, x)$ and $k_Y(y, y)$ are bounded, $k_X(\cdot, x) \in C_0(E)$ and $k_Y(\cdot, y) \in C_0(E)$.
5. $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense in $L_2(\pi)$ and $L_2(\nu)$ respectively, i.e., k_X and k_Y are universal.

Remark 2. *Assumptions 1-3 ensure that the inclusion mappings I_π, I_μ are bounded and compact, and that $\mathcal{H}_X$ and $\mathcal{H}_Y$ can be continuously embedded into $L_2(\pi)$ and $L_2(\nu)$, respectively. Assumptions 4-5 implies that the RKHS $\mathcal{H}_X$ and $\mathcal{H}_Y$ are dense in $L_2(\pi)$ and $L_2(\nu)$ respectively. In addition, Assumptions 1-5 together imply that the RKHS $\mathcal{G}$ induced by kernel $K = kId_{\mathcal{H}_Y}$ is dense in $L_2(E, \mathcal{F}_E, \nu, \mathcal{H}_Y)$. We also remark that the above assumptions are not restrictive in practice, since well-known kernels such as the Gaussian, Laplacian and Matérn kernels satisfy all of the above assumptions for $E \subseteq \mathbb{R}^d$ and arbitrary probability measures π and ν on $(E, \mathcal{F}_E)$.*

With the above assumptions made, our Theorem 2 demonstrates that the CME operator F_* can be arbitrarily approximated by operators in $S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$.

Theorem 2. *Let Assumptions* 1-5 *be satisfied, for every* $\delta > 0$*, there exists a Hilbert-Schmidt operator* $C^\alpha \in S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$ *such that*

$$\|F_* - C^\alpha \phi_X^\alpha(\cdot)\|^2_{L_2(E,\mathcal{F}_E,\pi,\mathcal{H}_Y)} < \delta.$$

In Eq (7) we prove that for any operator C^α in $S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$, there is a corresponding vector F^α in the vRKHS $\mathcal{G}^\alpha$ induced by the kernel $K^\alpha = k^\alpha \mathrm{Id}$. As a result, Theorem 2 shows that the constructed interpolation space $\mathcal{G}^\alpha$ is rich enough to approximate the CME operator F_* up to arbitrary accuracy under mild assumptions. However, there are infinitely many operators in $S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$, and how to find the appropriate C^α remains unclear. The following lemma provides a way to find C^α.

Lemma 1. *Let* $\alpha, \beta \in (0, 2]$ *and* $\beta < \alpha$*, the following two holds*

i). For any $f^\alpha \in \mathcal{H}_X^\alpha$ *and* $C^\alpha \in S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$*, there is a corresponding* $f^\beta \in \mathcal{H}_X^\beta$*,* $C^\beta : \mathcal{H}_X^\beta \to \mathcal{H}_Y$ *and* $C^\beta = C^\alpha C_{XX}^{\alpha-\beta}$ *such that*

$$C^\beta f^\beta = C^\alpha f^\alpha.$$

ii). We have $C^\beta \in S_2(\mathcal{H}_X^\beta, \mathcal{H}_Y)$.

By Lemma 1, we can see that if for some $C_{Y|X} \in S_2(\mathcal{H}_X, \mathcal{H}_Y)$ such that $C_{Y|X} k(\cdot, x) = \mathbb{E}[\phi_Y(Y)|X = x] = F_*(x)$, we will have that $C_{Y|X}^\beta k^\beta(\cdot, x) = C_{Y|X} C_{XX}^{1-\beta} k^\beta(\cdot, x) = C_{Y|X} k(\cdot, x) = F_*(x)$. Previous result [34] shows that $C_{Y|X} = C_{YX} C_{XX}^\dagger$. However, this requires that $\langle f_Y, C_{Y|X}\phi_X(\cdot)\rangle_{\mathcal{H}_Y} \in \mathcal{H}_X$, for all $f_Y \in \mathcal{H}_Y$. In contrast, if we use $C_{Y|X}^\beta$ instead, this would only require $\langle f_Y, C_{Y|X}^\beta \phi_X^\beta(\cdot)\rangle_{\mathcal{H}_Y} \in \mathcal{H}_X^\beta$, which is strictly weaker than the previous requirement.

Remark 3. *We point out that a simpler version of part i) in Lemma 1 in the case of* $\alpha = 1$ *has been proved in [39]. Part ii) in Lemma 1 is novel and is a crucial result that enables us to study the learning rate in the misspecified case.*

Our following result establishes the connection between the CME and the Hilbert-Schmidt operator space $S_2(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$ and points out the condition when the approximation error can be zero.

Theorem 3. *Under Assumptions* 1-5*, the following statements are equivalent:*

i) $F_*(\cdot) = \mathbb{E}[\phi_Y(Y)|X = \cdot] \in \mathcal{G}^\alpha$*;*

ii) There exists an operator $C^\alpha \in S(\mathcal{H}_X^\alpha, \mathcal{H}_Y)$ *such that*

$$[(C^\alpha)^* f_Y](x) = \langle (C^\alpha)^* f_Y, \phi_X^\alpha(x)\rangle_{\mathcal{H}_X^\alpha} = \langle f_Y, C^\alpha \phi_X^\alpha(x)\rangle_{\mathcal{H}_Y} = \mathbb{E}[f_Y(Y)|X = x]$$

for all $x \in E$ *and* $f_Y \in \mathcal{H}_Y$.

Both i) and ii) imply the following

iii) There exists an operator C^α *such that* $\|F_* - C^\alpha \phi_X^\alpha(\cdot)\|^2_{L_2(E,\mathcal{F}_E,\pi,\mathcal{H}_Y)} = 0$.

By comparing point ii) in Theorem 3 to the CME property [34], one can see that the operator C^α is exactly the CME operator introduced in [34]. In particular, in the well-specified case where $\alpha = 1$, we have the closed form of the CME operator as $C = C_{YX} C_{XX}^\dagger$ [34, 15].

4 Optimal Learning Rate for CME

In this section, we derive the learning rate for $\left\|\hat{F}_\lambda(X) - F_*(X)\right\|^2_{L_2(E,\mathcal{F}_E,\pi,\mathcal{H}_Y)}$. We first state additional assumptions that are needed in our derivation. As our assumptions match those of [8], we include the corresponding labels from [8] for ease of reference.

6. (EVD+) Recall $(\mu_i)_{i\in I}$ are the eigenvalues of C_{XX}, for some constant $c_1, c_2 > 0$ and $p \in (0, 1)$, we assume that

$$c_1 i^{-1/p} \le \mu_i \le c_2 i^{-1/p}.$$

7. (EMB) We assume that for $\alpha \in (p, 1]$, the inclusion map $I_\pi^\alpha : \mathcal{H}_X^\alpha \hookrightarrow L_\infty(\pi)$ is continuous and has bounded norm $A > 0$.

8. (SRC) There exists $0 < \beta \leq 2$ such that $F_* = C_{Y|X}^\beta k_X^\beta(\cdot, x) \in \mathcal{G}^\beta$.

9. (MOM) We assume that there are constants $\sigma, R > 0$ such that

$$\mathbb{E}_{Y|x}\left[\left(\{k_Y(\cdot, Y) - \mathbb{E}_{Y|x}[k_Y(\cdot, Y)]\}\right)^q\right] \leq \frac{1}{2} q! \sigma^2 R^{q-2}.$$

Assumption 6 is a standard assumption on the eigenvalue decays (see more details in [3, 8, 39]). Assumption 7 is referred as the embedding property in [8] and can be loosely understood as requiring that $\sum_i \mu_i^\alpha < \infty$ (see [8]). Therefore, the smaller α is, the smoother of the corresponding RKHS $\mathcal{H}_X$. Since we assume k_X to be bounded, the embedding property always hold true when $\alpha \geq 1$. Assumption 8 is often referred as the source condition in literature ([3, 8, 19, 20]). It imposes the smoothness assumption on the target CME operator F_*. In particular, when $\beta \geq 1$, the source condition implies that F_* has a representative from $\mathcal{G}$, indicating the well specified scenario. However, once we let $\beta < 1$, we are in the hard learning scenario, which is the main interest in this manuscript. Finally, Assumption 9 is a typical noise condition that we impose on the underlying distribution $\mathbb{P}$ (see [3, 8, 39] for more details).

Under these assumptions, our next result provide an upper bound on the learning rate.

Theorem 4. *Let Assumption* 1-3 *and* 6-9 *hold, for some constant* c *and* $r > 1$*, define the parameter*

$$\lambda_n = \Theta\left(\left(\frac{n}{\log^r n}\right)^{-\frac{1}{\max\{\beta+p, \alpha\}}}\right),$$

then we have for some constant c*, with probability greater than* $1 - \delta$*,*

$$\left\|\hat{F}_\lambda(X) - F_*(X)\right\|^2_{L_2(E, \mathcal{F}_E, \pi, \mathcal{H}_Y)} \leq c \log\left(\delta^{-1}\right) \left(\frac{n}{\log^r n}\right)^{-\frac{\beta}{\max\{\alpha, \beta+p\}}}.$$

Theorem 4 provides the finite sample learning rate for the empirical CME estimator defined in Eq. (4). It states that the learning rate for $\hat{F}_\lambda$ is governed by the interplay between p, α, and β. Intuitively, p describes the decay rate of the eigenvalues $(\mu_i)_{i \in I}$, and α determines the boundedness of the interpolation kernel and has maximum value of 1 according to our assumption. β characterizes the smoothness of the target CME operator.

The exponent $\beta / \max\{\alpha, \beta + p\}$ explicitly provides the learning rate for the CME operator. For example, if we have $\alpha \leq \beta$, we obtain a learning rate of $\beta/(\beta + p)$. In particular, if Gaussian kernel is used where p is arbitrarily close to 0, we can see that our learning rate can even achieve $\log^r n/n$. In addition, if a kernel with slow eigenvalue decay is used, such as the Matérn kernel, we can obtain the minimax optimal learning rate $n^{-1/2}$ up to logarithmic factors if we have $p \leq \beta$. Finally, in the worst case where β is close to 0, the learning rate can be arbitrarily slow.

We point out that our obtained upper bound on the learning rate has a similar form to [39]. In particular, by considering the γ-norm for some $0 < \gamma < \beta$, [39] obtain an upper bound of $O\left(\frac{n}{\log^r n}\right)^{-\frac{\beta-\gamma}{\max\{\alpha, \beta+p\}}}$. There are two main differences between their result and ours. A first difference is that we consider the L_2 difference between $\hat{F}_\lambda$ and F_*, while [39] study the difference between $\hat{C}_{Y|X,\lambda}$ and $C_{Y|X}^\beta$. A second difference, which is arguably the more important of the two, is that [39] address the misspecified case by assuming that the smoothness of the target CME F_* is related to the size of the RKHS $\mathcal{H}_X$, i.e., $\beta > p$. This can be limiting, especially when non-smooth kernels such as Matérn kernels are used, since in these cases p is close to 1. In contrast, by realizing that the vRKHS induced by the kernel K^β can be identified with the Hilbert-Schmidt operator space, i.e., $C^\beta \in S_2(\mathcal{H}_X^\beta, \mathcal{H}_Y)$, we can avoid making this strong assumption (with implications for the proof, as briefly discussed next).

Sketch of Proof We first notice that

$$\begin{aligned}\left\|\hat{F}_\lambda(x) - F_*(x)\right\|_{\mathcal{H}_Y} &= \left\|\hat{F}_\lambda(x) - F_{Y|X,\lambda}(x) + F_{Y|X,\lambda}(x) - F_*(x)\right\|_{\mathcal{H}_Y} \\ &\leq \left\|\hat{F}_\lambda(x) - F_{Y|X,\lambda}(x)\right\|_{\mathcal{H}_Y} + \left\|F_{Y|X,\lambda}(x) - F_*(x)\right\|_{\mathcal{H}_Y}.\end{aligned}$$

where we define $F_{Y|X,\lambda}(x) = C_{YX}(C_{XX} + \lambda)^{-1}k_X(\cdot, x)$. It is established that $\hat{F}_\lambda(x) = \hat{C}_{YX}(\hat{C}_{XX} + \lambda)^{-1}k_X(\cdot, x)$. Hence we can see that the error for the first term is mainly due to the sample approximation. We therefore refer the first term as the *Variance*, for which [39] establish consistency. Since we assume $F_*(x) = C^\beta_{Y|X}k^\beta_X(\cdot, x)$, the second term can be written as

$$F_{Y|X,\lambda}(x) - F_*(x) = -\sum_{i\in I}\frac{\lambda}{\mu_i + \lambda}C^\beta_{Y|X}\mu^\beta_i e_i(x)e_i.$$

We refer to this term as the *Bias*. Our proof of convergence of the bias adapts the proof in [28, Theorem 6], and utilizes the fact that $C^\beta_{Y|X}$ is Hilbert-Schmidt to obtain a sharp rate.

Our final theorem provides a lower convergence rate, which allows us to confirm the optimality of our learning rate.

Theorem 5. *Suppose we are given a dataset $D := \{(x_i, y_i)\}_{i=1}^n$. Let Assumptions 1-3 hold. Let π be the marginal distribution defined on E such that Assumption 6 and 7 hold for kernel k_X with $0 < p \leq \alpha \leq 1$. Then for all $0 < \beta \leq 2$ and all constants σ, R, c_1, there exists constants $c_2, c_3, c_4 > 0$ such that for all learning methods $D \to \hat{F}_D$ and all constant $\tau > 0$, there is a distribution $\mathbb{P}_0$ defined on $E \times \mathcal{H}_Y$ with its marginal distribution as π such that $\|F_*\|_{L_2(E,\mathcal{F}_E,\pi;\mathcal{H}_Y)} \leq c_1$ and Assumption 8 and 9 are satisfied with σ and R and with probability not less than $1 - c_2\tau^{1/c_3}$,*

$$\|\hat{F}_D - F_*\|^2_{L_2(E,\mathcal{F}_E,\pi;\mathcal{H}_Y)} \geq \tau^2 c_4 n^{-\frac{\max\{\alpha,\beta\}}{\max\{\alpha,\beta\}+p}}.$$

Theorem 5 states that under Assumption 1-3 and 6-9, there is no learning method can achieve a learning rate faster than $n^{-\frac{\max\{\alpha,\beta\}}{\max\{\alpha,\beta\}+p}}$. To our knowledge, this is the first analysis that demonstrates the lower rate for CME learning. In the context of regularized regression, [3, 37, 2] provide a similar lower bound on the learning rate. However, a key difference in our analysis is that the output of the regression learning now lives in an potentially infinite dimensional RKHS, rather than in $\mathbb{R}$. Our analysis reveals that in the case where $\alpha \leq \beta$, the obtained upper rate in Theorem 4 is optimal. However, the optimal rate for $\beta < \alpha$ still remains a challenge (even when the output is $\mathbb{R}$).

5 Conclusion

In this paper, we provide a rigorous theoretical foundation for approximating the CME operator, as well as studying the statistical learning rate. Utilizing recently developed interpolation space techniques, we first demonstrate that the Hilbert-Schmidt operator space $S_2(\mathcal{H}^\alpha_X, \mathcal{H}_Y)$ defined on the interpolation space $\mathcal{H}^\alpha_X$ and the target Hilbert space $\mathcal{H}_Y$ can approximate the CME operator F_* with arbitrary accuracy. This then allows us to define the target CME operator to live in the larger interpolation space $\mathcal{G}^\alpha$, in contrast to the well-specified setting where $F_* \in \mathcal{G}$. By doing so, we are able to study the convergence rate of the empirical CME operator in the misspecified scenario. In the most benign case, we derive the optimal learning rate $O(\log n/n)$ up to logarithmic factors without requiring finite dimensionality on $\mathcal{H}_Y$. We also show that when less smooth kernels are used, the learning rate can achieve the minimax optimal rate $O(n^{-1/2})$. Finally, we provide a matching lower bound on the learning rate, which reveals that the obtained upper rate is optimal when $\beta > \alpha$.

Looking beyond, our current interpolation space setting indicates that the convergence rate can be arbitrarily slow if $\beta \to 0$. This prevents the learning of the constant function, which plays a crucial role in completing the theory of CME as pointed out by [14]. Addressing this challenge is an important future direction of research.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[2] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

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

[4] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408, 2006.

[5] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

[6] C. Ciliberto, L. Rosasco, and A. Rudi. A consistent regularization approach for structured prediction. *Advances in neural information processing systems*, 29, 2016.

[7] E. De Vito, L. Rosasco, and A. Caponnetto. Discretization error analysis for tikhonov regularization. *Analysis and Applications*, 4(01):81–99, 2006.

[8] S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:205–1, 2020.

[9] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

[10] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

[11] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors-supplementary. *arXiv preprint arXiv:1205.4656*, 2012.

[12] S. Grunewalder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in mdps with rkhs embeddings. *arXiv preprint arXiv:1206.4655*, 2012.

[13] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

[14] I. Klebanov, I. Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

[15] I. Klebanov, B. Sprungk, and T. Sullivan. The linear conditional expectation in hilbert space. *Bernoulli*, 27(4):2267–2299, 2021.

[16] S. Klus, P. Koltai, and C. Schütte. On the numerical approximation of the perron-frobenius and koopman operator. *arXiv preprint arXiv:1512.05997*, 2015.

[17] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28(3):985–1010, 2018.

[18] M. Korda and I. Mezić. On convergence of extended dynamic mode decomposition to the koopman operator. *Journal of Nonlinear Science*, 28(2):687–710, 2018.

[19] J. Lin and V. Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In *International Conference on Machine Learning*, pages 3092–3101. PMLR, 2018.

[20] J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

[21] H. Q. Minh. Regularized divergences between covariance operators and gaussian measures on hilbert spaces. *Journal of Theoretical Probability*, 34(2):580–643, 2021.

[22] J. Mitrovic, D. Sejdinovic, and Y. W. Teh. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018.

[23] M. Mollenhauer and P. Koltai. Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*, 2020.

[24] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[25] Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert space embeddings of pomdps. *arXiv preprint arXiv:1210.4887*, 2012.

[26] J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*, 33:21247–21259, 2020.

[27] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[28] R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] S. Smale and D.-X. Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41(3):279–305, 2004.

[30] S. Smale and D.-X. Zhou. Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

[31] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[32] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *Signal Processing Magazine, IEEE*, 30(4):98–111, 2013.

[33] L. Song, A. Gretton, and C. Guestrin. Nonparametric tree graphical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 765–772. JMLR Workshop and Conference Proceedings, 2010.

[34] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.

[35] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[36] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[37] I. Steinwart, D. R. Hush, C. Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

[38] I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, 2012.

[39] P. Talwai, A. Shameli, and D. Simchi-Levi. Sobolev norm learning rates for conditional mean embeddings. *arXiv preprint arXiv:2105.07446*, 2021.

[40] A. B. Tsybakov. Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

[41] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Page 9.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] To the best of our knowledge, our research does not impose any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Section 3 provides all the assumptions for our research.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proof of all the results are provided in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]