Counterfactual Mean Embedding: A Kernel Method for Nonparametric Causal Inference

Krikamol Muandet[*] **Motonobu Kanagawa**[†]
Max Planck Institute for Intelligent Systems
Tübingen, Germany
firstname.lastname@tuebingen.mpg.de

Sorawit Saengkyongam
Agoda
Bangkok, Thailand
sorawitj@gmail.com

Sanparith Marukatat
NECTEC, NSTDA
Pathumthani, Thailand
sanparith.marukatat@nectec.or.th

Abstract

This paper introduces a novel Hilbert space representation of a counterfactual distribution—called counterfactual mean embedding (CME)—with applications in nonparametric causal inference. Counterfactual prediction has become an ubiquitous tool in machine learning applications, such as online advertisement, recommendation systems, and medical diagnosis, whose performance relies on certain interventions. To infer the outcomes of such interventions, we propose to embed the associated counterfactual distribution into a reproducing kernel Hilbert space (RKHS) endowed with a positive definite kernel. Under appropriate assumptions, the CME allows us to perform causal inference over the entire landscape of the counterfactual distribution. The CME can be estimated consistently from observational data without requiring any parametric assumption about the underlying distributions. We also derive a rate of convergence which depends on the smoothness of the conditional mean and the Radon-Nikodym derivative of the underlying marginal distributions. Our framework can deal with not only real-valued outcome, but potentially also more complex and structured outcomes such as images, sequences, and graphs. Lastly, our experimental results on off-policy evaluation tasks demonstrate the advantages of the proposed estimator.

1 Introduction

To infer causal relation, it is natural to state the problem in terms of counterfactual question, *e.g.*, *would the patient has recovered had the medical treatment been different?* This school of thought is influenced predominantly by the potential outcome framework [28]. It has been studied extensively in classical statistics and has a wide range of applications in social science, econometrics, and epidemiology. Moreover, important applications of machine learning such as online advertisement and recommendation system can be reformulated under this framework [6, 21, 29]. Although a randomized experiment—which is considered a gold standard in causal inference—can in principle be employed for these applications, it can be too expensive, time-consuming, or unethical to implement in practice. Hence, this work will focus on what are known as observational studies [26, 28].

[*] Done partially while KM was at the Department of Mathematics, Mahidol University, Thailand.

[†] Done partially while MK was at the Institute of Statistical Mathematics, Tokyo.

Preprint. Work in progress.

In observational studies, we are interested in inferring causal relation between a treatment T and an outcome Y, which have been recorded along with a covariate X. In an online advertisement, for example, the covariate usually encodes information about the users and the associated queries. The treatment may be an ad placement and the outcome is determined by whether or not the user clicks the advertisement. Throughout, we denote the space of treatments by $\mathcal{T}$, the space of covariates by $\mathcal{X}$, and the space of possible outcomes by $\mathcal{Y}$. For $\mathbf{x} \in \mathcal{X}$ and $t \in \mathcal{T}$, $Y_t(\mathbf{x})$ denotes the *potential outcome* for $\mathbf{x}$ under the treatment $T = t$. Likewise, we denote the *counterfactual outcome* for $\mathbf{x}$ under the treatment $t^* \neq t$ after the treatment $T = t$ is already applied by $Y_{t^*}(\mathbf{x})$. That is, $Y_{t^*}(\mathbf{x})$ is defined after we already observe the value of $Y_t(\mathbf{x})$. We refer to the distribution of $Y_{t^*}(\mathbf{x})$ as the counterfactual distribution. Then, inferring causal relation between T and Y would have been as straightforward as calculating the difference $Y_t(\mathbf{x}) - Y_{t^*}(\mathbf{x})$ had we known the value of both $Y_t(\mathbf{x})$ and $Y_{t^*}(\mathbf{x})$. However, as implied in the preceding statement, it is virtually impossible to observe both of them simultaneously due to the fundamental problem of causal inference. Instead, we must resort to a logged dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{t}_i, \mathbf{y}_i)\}_{i=1}^n$ where $(\mathbf{x}_i, \mathbf{t}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{T} \times \mathcal{Y}$.

From machine learning perspective, attempts have recently been made to formulate the problem above as a supervised learning [19, 30, 37]. For example, one possibility is to fit a predictor $h : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ directly to the dataset $\mathcal{D}$ and then use it to estimate Y_{t^*} [16, 19]. Unfortunately, the learnt predictor h is almost always biased because only one potential outcome is observed for a given covariate $\mathbf{x}$. Moreover, the treatment assignment mechanism is not always under our control and in general could depend on a hidden confounder. Hence, learning causal relation from logged data $\mathcal{D}$ differs fundamentally from standard supervised learning. In [19, 30], the authors propose to reduce this bias by also learning a joint representation of covariates in treatment and control groups. Other well-known techniques such as inverse propensity score (IPS) weighting [20, 6], doubly robust estimator [10], and deep learning [19, 14] have also been applied successfully.

In this work, we propose a novel representation of counterfactual distributions of Y_{t^*} called counterfactual mean embedding (CME). The CME relies on kernel mean embedding [4, 31, 24] which maps probability distributions into a reproducing kernel Hilbert space (RKHS). Based on this representation, causal inference can be performed over the entire landscape of counterfactual distribution using the kernel arsenal. We show that this representation can be estimated *consistently* from observational data. Furthermore, we establish a convergence rate of the estimator which depends on the smoothness regarding the conditional distribution and the Radon-Nikodym derivative of the underlying marginal distributions. Interestingly, we found that our estimator is guaranteed to converge reasonably fast, if *either* the Radon-Nikodym derivative *or* the conditional mean is smooth. This property resembles that of *doubly robust* estimator [7, 10]. Our framework is nonparametric as it requires no parametric assumption about the underlying distributions. Since the CME depends only on Gram matrices evaluated on the data, it can potentially be applied to more complex and structured outcomes such as images, sequences, and graphs. Lastly, we demonstrate the effectiveness of the proposed estimator on simulated data as well as real-world policy evaluation tasks.

The rest of the paper is organized as follows. Section 2 introduces kernel methods and Hilbert space embedding of distributions, which form the backbone of this work. Section 3 introduces counterfactual learning and then provides a generalization of Hilbert space embedding to counterfactual distributions. In this section, we also present the theoretical results and the application of CME in off-policy evaluation. Finally, experimental results on both simulated and real data are provided in Section 4.

2 Hilbert space embedding of distributions

Let $\mathcal{X}$ be a nonempty set and $\mathscr{H}$ be a reproducing kernel Hilbert space (RKHS) of functions $f : \mathcal{X} \to \mathbb{R}$. The RKHS $\mathscr{H}$ is uniquely determined by a symmetric, positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and possesses two important properties [1]: (i) for any $\mathbf{x} \in \mathcal{X}$, the function $k(\mathbf{x}, \cdot) : \mathbf{x}' \mapsto k(\mathbf{x}, \mathbf{x}')$ is an element of $\mathscr{H}$, (ii) the inner product in $\mathscr{H}$ satisfies the reproducing property, *i.e.*, for all $f \in \mathscr{H}$ and $\mathbf{x} \in \mathcal{X}$, $f(\mathbf{x}) = \langle k(\mathbf{x}, \cdot), f \rangle_{\mathscr{H}}$. Let $\mathscr{P}$ denote the set of probability measures on a measurable space $\mathcal{X}$. Then, the *kernel mean embedding* (KME) of $\mathbb{P} \in \mathscr{P}$ can be defined as [4, 31, 24]

$$\mu : \mathscr{P} \to \mathscr{H}, \quad \mathbb{P} \mapsto \mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\mathbf{x}, \cdot) \, \mathrm{d}\mathbb{P}(\mathbf{x}). \tag{1}$$

It is well-defined if k is measurable and bounded, *i.e.*, $\sup_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x},\mathbf{x}) < \infty$. By reproducing property of $\mathscr{H}$, $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}}\rangle_{\mathscr{H}}$ for any $f \in \mathscr{H}$. Given an i.i.d. sample $\{\mathbf{x}_i\}_{i=1}^{n}$ from $\mathbb{P}$, the empirical estimate of (1) is given by $\hat{\mu}_{\mathbb{P}} := \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i,\cdot)$. The $\sqrt{n}$-consistency of $\hat{\mu}_{\mathbb{P}}$ has been established in [32, Theorem 27] and also in [13, 22, 39].

By virtue of (1), a well-known discrepancy measure called *maximum mean discrepancy* (MMD) between two distributions $\mathbb{P}$ and $\mathbb{Q}$ can be defined as $\mathrm{MMD}^2[\mathscr{H},\mathbb{P},\mathbb{Q}] = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathscr{H}}$. The MMD has been applied extensively in *two-sample testing* [5, 13]. The RKHS $\mathscr{H}$ (and the associated kernel k) is said to be *characteristic* if $\mathrm{MMD}^2[\mathscr{H},\mathbb{P},\mathbb{Q}] = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. In which case, the map (1) is injective which implies that $\mu_{\mathbb{P}}$ captures all necessary information about $\mathbb{P}$. Examples of characteristic kernels include Gaussian kernel $k(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|_2^2/2\sigma^2), \sigma > 0$ and Laplacian kernels $k(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|_2/\sigma), \sigma > 0$ [11, 35].

The KME can be extended to conditional distributions $\mathbb{P}(Y|X)$ via the notion of covariance operators in RKHS [34, 33]. Let (X,Y) be a random variable taking value on $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{Y}$ is another measurable space, and $(\mathscr{H},k)$ and $(\mathscr{F},\ell)$ be RKHSs with measurable kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume that $\mathbb{E}_X[k(X,X)] < \infty$ and $\mathbb{E}_Y[\ell(Y,Y)] < \infty$. The (uncentered) *covariance operator* $\mathcal{C}_{YX} : \mathscr{H} \to \mathscr{F}$, which encodes the information of the joint distribution on (X,Y), is defined as $\mathcal{C}_{YX}f := \mathbb{E}_{X,Y}[\ell(\cdot,Y)f(X)] \in \mathcal{F}$ for $f \in \mathscr{H}$. Alternatively, $\mathcal{C}_{YX}$ can be expressed in terms of a tensor product $\mathcal{C}_{YX} = \mathbb{E}_{X,Y}[\ell(\mathbf{y},\cdot)\otimes k(\mathbf{x},\cdot)]$. If $X = Y$, we write the covariance operator as $\mathcal{C}_{XX} : \mathscr{H} \to \mathscr{H}$. See Appendix C.1 and, *e.g.*, [11, 42] for furthers details on covariance operators.

The *conditional mean embedding* of $\mathbb{P}(Y|X = \mathbf{x})$ for some $\mathbf{x} \in \mathcal{X}$ is defined as $\mu_{Y|\mathbf{x}} := \mathbb{E}_{Y|\mathbf{x}}[\ell(Y,\cdot)\,|\,X = \mathbf{x}] = \mathcal{U}_{Y|X}k(\mathbf{x},\cdot)$ where $\mu_{Y|\mathbf{x}}$ is an element in $\mathscr{F}$ and $\mathcal{U}_{Y|X}$ is an operator from $\mathscr{H}$ to $\mathscr{F}$ [34, 33]. By the reproducing property of $\mathscr{F}$, $\mathbb{E}_{Y|\mathbf{x}}[g(Y)|X = \mathbf{x}] = \langle g, \mu_{Y|\mathbf{x}}\rangle_{\mathscr{F}}$ for all $g \in \mathscr{F}$. The embeddings $\mathcal{U}_{Y|X}$ and $\mu_{Y|\mathbf{x}}$ can respectively be expressed in terms of the covariance operators as $\mathcal{U}_{Y|X} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ and $\mu_{Y|\mathbf{x}} := \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}k(\mathbf{x},\cdot)$ under certain assumptions [34, 33, 24].[3] In what follows, we treat $\mathcal{U}_{Y|X}$ and $\mu_{Y|\mathbf{x}}$ as RKHS representations of $\mathbb{P}(Y|X)$ and $\mathbb{P}(Y|X = \mathbf{x})$. See, *e.g.*, [33, 24] and references therein for applications of conditional mean embedding.

3 Counterfactual mean embedding

Throughout, we will assume w.l.o.g. that $\mathcal{T} = \{0,1\}$. Then, the causal effect in potential outcome framework is usually characterized by an *individual treatment effect* (ITE), given by $\mathrm{ITE}(\mathbf{x}) := Y_1(\mathbf{x}) - Y_0(\mathbf{x})$, for a covariate $\mathbf{x}$. Since we can never observe both $Y_0(\mathbf{x})$ and $Y_1(\mathbf{x})$ at the same time, one often resort to the *average treatment effect* (ATE) defined by $\mathrm{ATE} := \mathbb{E}_{\mathbf{x}\sim\mathbb{P}(\mathbf{x})}[Y_1(\mathbf{x}) - Y_0(\mathbf{x})]$ instead. Unlike the ITE, the ATE can be estimated empirically as $\widehat{\mathrm{ATE}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_1(\mathbf{x}_i) - \frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_0(\tilde{\mathbf{x}}_j)$ where $\{\mathbf{y}_1(\mathbf{x}_i)\}_{i=1}^{n}$ and $\{\mathbf{y}_0(\tilde{\mathbf{x}}_j)\}_{j=1}^{n}$ are the post-treatment outcomes. Other summary statistics such as ratio and quantile of the distribution have also been investigated [15].

To make causal claims, we require the following assumptions, which are common in observational studies.

Assumption 1. *(**A1**) Stable unit treatment value assumption (SUTVA): the outcome of subject i is independent of the outcomes of other individuals and their received treatments. (**A2**) Conditional exogeneity/unconfoundedness/ignorability: $Y_0, Y_1 \perp\!\!\!\perp T|X$. In other words, given the covariates X the outcome is independent of the treatment assignment. This assumption implies that the distributions of $Y_j \,|\, X, T = j$ and $Y_j \,|\, X$ agree [27, 15, 18]. (**A3**) The common support assumption: $\mathcal{X}_1 \subseteq \mathcal{X}_0$.*

Under these assumptions, we can claim that $\mathrm{ATE} = 0$ if T has no causal effect on Y. Nevertheless, in many contexts, *e.g.*, applied econometrics, the outcome distribution may change in ways that cannot be revealed by an examination of the averages. For example, wage distributions tend to be skewed to the right in which case mean effects would deem inappropriate [23]. This motivates us to consider the estimator of the entire outcome distribution.

[3] This holds under the assumption that $\mathbb{E}_{Y|X}[g(Y)|X = \cdot] \in \mathscr{H}$ for all $g \in \mathscr{F}$ [34, 33, 12]. Note that we nevertheless do *not* require this condition for our theoretical analysis; see Sec. 3.2

3.1 Hilbert space representation of counterfactual distribution

We first introduce the notion of counterfactual distribution and then define corresponding embedding in the RKHS. As a working example, let us consider the following question taken from [9]: "*what would women wages have prevailed if they face the men wage schedule, or vice versa?*". This type of questions cannot be addressed by randomized experiments, due to an ethical or practical issue.

Let $\mathbb{P}_{Y\langle 0|0\rangle}$ and $\mathbb{P}_{X_0}$ be probability distributions defined on $\mathcal{Y}$ and $\mathcal{X}$, and assume they respectively represent *observed distributions* of wages and features for men. Similarly, let $\mathbb{P}_{Y\langle 1|1\rangle}$ and $\mathbb{P}_{X_1}$ be the corresponding observed distributions for women. Then, the *counterfactual distribution* of wages that would have prevailed for women had they faced the men's wage schedule is defined by [9] as

$$\mathbb{P}_{Y\langle 0|1\rangle}(\mathbf{y}) := \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}). \tag{2}$$

Assumption (A3) ensures that the above integral is well defined. Note that $\mathbb{P}_{Y\langle 0|1\rangle}$ does not arise as a distribution from any observable distribution and hence is not observable in practice. Through (2), we are able to consider counterfactual scenarios, where changes may occur in the covariate distributions, or in the conditional distribution of the outcome given covariates. In this paper, we focus on the effect of changes in the covariate distributions (*i.e.*, the change from $\mathbb{P}_{X_0}$ to $\mathbb{P}_{X_1}$) to the outcomes. This scenario is also related to domain adaptation problems in machine learning [3, 41].

Definition 1 (Counterfactual mean embedding)**.** *Assume that (A3) holds. Then an RKHS embedding of the counterfactual distribution* (2) *is defined by*

$$\mu_{Y\langle 0|1\rangle} := \int \ell(\mathbf{y},\cdot)\,\mathrm{d}\mathbb{P}_{Y\langle 0|1\rangle}(\mathbf{y}) = \iint \ell(\mathbf{y},\cdot)\,\mathrm{d}\mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}). \tag{3}$$

The counterfactual distribution corresponding to the individual treatment effect (ITE) can be obtained by restricting $\mathbb{P}_{X_1}$ to the Dirac measure $\delta_{\mathbf{x}}$ for some $\mathbf{x} \in \mathcal{X}$. In what follows, we define $\mathbb{P}_{Y^*\langle 0|1\rangle}$ as the true interventional distribution associated with the treatment. Then, the following lemma assigns a causal interpretation to the CME, which follows from [9, Lemma 2.1] and Definition 1.

Lemma 1 (Causal interpretation)**.** *Suppose that (A2) holds, i.e.,* $Y_0, Y_1 \perp\!\!\!\perp T|X$ *almost surely for* X*, and that (A3) holds. Then* $\mu_{Y\langle 0|1\rangle} = \mu_{Y^*\langle 0|1\rangle}$ *where* $\mu_{Y^*\langle 0|1\rangle}$ *is the embedding of* $\mathbb{P}_{Y^*\langle 0|1\rangle}$.

Lemma 1 equips $\mu_{Y\langle 0|1\rangle}$ with an arsenal to perform causal inference, *i.e.*, it can be viewed as a representation of the actual interventional distribution associated with the specified treatment. If we further assume that ℓ is characteristic, then $\mu_{Y\langle 0|1\rangle}$ captures all necessary information about $\mathbb{P}_{Y^*\langle 0|1\rangle}$.

In practice, it is not possible to obtain a sample from $\mathbb{P}_{Y\langle 0|1\rangle}$, and therefore $\mu_{Y\langle 0|1\rangle}$ cannot be estimated directly, which differs from the case of the standard mean embedding $\hat{\mu}_{\mathbb{P}}$. We therefore propose the following estimator (4), which instead uses samples from $\mathbb{P}_{Y_0|X_0}$ and $\mathbb{P}_{X_1}$ to estimate $\mu_{Y\langle 0|1\rangle}$. The estimator is essentially (or superficially) the *kernel sum rule*, so the proof of the following proposition can be found in [33, 12]. Note that there is a conceptual difference between our estimator and the standard kernel sum rule: Our estimator is considered as the kernel sum rule *equipped with* a causal interpretation, provided that the assumptions in Lemma 1 hold.

Proposition 1. *Given samples* $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ *from* $\mathbb{P}_{Y_0 X_0}(\mathbf{x}, \mathbf{y})$ *and* $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$ *from* $\mathbb{P}_{X_1}(\mathbf{x})$*, let* $\widehat{\mathcal{C}}_{YX} := \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{y}_i,\cdot)\otimes k(\mathbf{x}_i,\cdot)$ *and* $\widehat{\mathcal{C}}_{XX} := \frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\cdot)\otimes k(\mathbf{x}_i,\cdot)$ *be empirical covariance operators and let* $\hat{\mu}_{X_1} := \frac{1}{m}\sum_{i=1}^m k(\mathbf{x}'_i,\cdot)$ *be the empirical kernel mean. Then an empirical estimator of* $\mu_{Y\langle 0|1\rangle}$ *is defined and expressed as*

$$\hat{\mu}_{Y\langle 0|1\rangle} := \widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon I)^{-1}\hat{\mu}_{X_1} = \Phi(\mathbf{K} + n\varepsilon\mathbf{I})^{-1}\tilde{\mathbf{K}}\mathbf{1}_m \tag{4}$$

where $\varepsilon > 0$ *is a regularization constant,* $\mathbf{1}_m = (1/m, \ldots, 1/m)^\top$, $\mathbf{K} \in \mathbb{R}^{n\times n}$ *with* $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\tilde{\mathbf{K}} \in \mathbb{R}^{n\times m}$ *with* $\tilde{\mathbf{K}}_{ij} = k(\mathbf{x}_i, \mathbf{x}'_j)$*, and* $\Phi = (\ell(\mathbf{y}_1,\cdot), \ldots, \ell(\mathbf{y}_n,\cdot))^\top \in \mathcal{F}^n$*. Note that we can write* $\hat{\mu}_{Y\langle 0|1\rangle} := \sum_{i=1}^n \beta_i \ell(\mathbf{y}_i,\cdot)$ *where* $\boldsymbol{\beta} = (\mathbf{K} + n\varepsilon\mathbf{I})^{-1}\tilde{\mathbf{K}}\mathbf{1}_n$.

Last but not least, it is generally challenging to perform model selection (in our case, the choice of k, ℓ and ε) in counterfactual prediction. To the best of our knowledge, no systematic method has been used in this setting. To this end, we modify the classical cross validation procedure and we use it for model selection. Details are omitted to conserve space (see Appendix B).

3.2 Theoretical analysis: consistency and convergence rates

In the sequel we will use the following notation. Let $L_2(\mathbb{P}_{X_0})$ be the Hilbert space of square-integrable functions,[4] and $\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}$ be the product measure of $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_0}$ on the product space $\mathcal{X} \times \mathcal{X}$. For convergence analysis, we make the following assumption.

Assumption 2. **(i)** *$\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces, k and ℓ are measurable kernels on $\mathcal{X}$ and $\mathcal{Y}$, and $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_1}$ are probability measures on $\mathcal{X}$.* **(ii)** *k and $\mathbb{P}_{X_0}$ satisfy $\int k(x,x)\,\mathrm{d}\mathbb{P}_{X_0}(x) < \infty$.* **(iii)** *$\mathbb{P}_{X_1}$ is absolutely continuous w.r.t. $\mathbb{P}_{X_0}$ with the Radon-Nikodym derivative $g := \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$ satisfying $g \in L_2(\mathbb{P}_{X_0})$.* **(iv)** *A function $\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $\theta(x, \tilde{x}) := \mathbb{E}[\ell(Y_0, \tilde{Y}_0) | X_0 = x, \tilde{X}_0 = \tilde{x}]$, where $(\tilde{X}_0, \tilde{Y}_0)$ is an independent copy of (X_0, Y_0), satisfies $\theta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$.* **(v)** *Let $n = m$, and samples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ and $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$ in Prop. 4 satisfy $\|\widehat{\mathcal{C}}_{XY} - \mathcal{C}_{XY}\| = O_p(n^{-1/2})$, $\|\widehat{\mathcal{C}}_{XX} - \mathcal{C}_{XX}\| = O_p(n^{-1/2})$, and $\|\hat{\mu}_{X_1} - \mu_{X_1}\|_{\mathscr{H}} = O_p(n^{-1/2})$ as $n \to \infty$.*

In Assumption 2, the condition **(i)** is a minimum assumption, and the condition **(ii)** is satisfied for instance if k is bounded. The condition **(iii)** requires the support of $\mathbb{P}_{X_1}$ be included in that of $\mathbb{P}_{X_0}$, and thus enforces the common support assumption (A3) in Assumption 1. In **(iv)** the function θ encodes the information of the conditional distribution $\mathbb{P}_{Y_0|X_0}$ of Y_0 given X_0, and the assumption $\theta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$ is satisfied for instance if the kernel ℓ is bounded. The condition **(v)** requires that samples are $\sqrt{n}$-consistent, and is satisfied when $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ are i.i.d. with $\mathbb{P}_{Y_0 X_0}(\mathbf{x}, \mathbf{y})$, and $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$ are i.i.d. with $\mathbb{P}_{X_1}(\mathbf{x})$, for instance. However, the condition **(v)** does not require that samples be independent, and can be satisfied even when $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ and $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$ are given as time-series data, for example, if they satisfy an appropriate stationarity condition. We assume $n = m$ for simplicity of presentation.

Theorem 1 below established the consistency of the CME estimator (4), where we also require that the RKHS $\mathscr{H}$ be dense in $L_2(\mathbb{P}_{X_0})$, which is satisfied by commonly used kernels such as Gaussian and Matérn kernels[5] on $\mathcal{X} = \mathbb{R}^d$. The proof can be found in Appendix C.2.

Theorem 1 (Consistency)**.** *Assume that $\mathscr{H}$ is dense in $L_2(\mathbb{P}_{X_0})$, and that Assumption 2 is satisfied. Then if ε_n decays to zero sufficiently slowly as $n \to \infty$, we have $\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathscr{F}} \to 0$ in probability as $n \to \infty$.*

Theorem 1 does not require any parametric assumption on $\mathbb{P}_{Y_0 X_0}(\mathbf{x}, \mathbf{y})$ and $\mathbb{P}_{X_1}(\mathbf{x})$. Besides, it can be considered as a version of [12, Theorem 8] that proves the consistency of the kernel sum rule. Unlike ours, however, [12] assumes that the function θ belongs to the tensor-product RKHS $\mathscr{H} \otimes \mathscr{H}$; this is a strong condition that may not be satisfied in practice. On the other hand, for θ we only require that $\theta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$, which is satisfied if ℓ is bounded, as mentioned.

To derive convergence rates, we need to define the following concepts, whose details can be found in Appendix C.1. Define an integral operator $T : L_2(\mathbb{P}_{X_0}) \to L_2(\mathbb{P}_{X_0})$ by $(Tf)(x) := \int k(x, \tilde{x}) f(\tilde{x})\,\mathrm{d}\mathbb{P}_{X_0}(\tilde{x})$. Under the condition **(ii)** in Assumption 2, T can be written as $Tf = \sum_{i=1}^{\infty} \mu_i \langle f, [e_i]_\sim \rangle_{L_2(\mathbb{P}_{X_0})}$ for any $f \in L_2(\mathbb{P}_{X_0})$ with convergence in $L_2(\mathbb{P}_{X_0})$, where $(\mu_i)_{i=1}^{\infty} \subset (0, \infty)$ and $([e_i]_\sim)_{i=1}^{\infty}$ is an orthonormal system in $L_2(\mathbb{P}_{X_0})$. Then for a constant $\alpha > 0$, the α-th power of T is defined as $T^\alpha f := \sum_{i=1}^{\infty} \mu_i^\alpha \langle f, [e_i]_\sim \rangle_{L_2(\mathbb{P}_{X_0})}$ for $f \in L_2(\mathbb{P}_{X_0})$. Define further an integral operator $T \otimes T : L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}) \to L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$ by $(T \otimes Th)(x_1, x_2) := \iint k(x_1, \tilde{x}_1) k(x_2, \tilde{x}_2) h(\tilde{x}_1, \tilde{x}_2)\,\mathrm{d}\mathbb{P}_{X_0}(\tilde{x}_1)\,\mathrm{d}\mathbb{P}_{X_0}(\tilde{x}_2)$ for $h \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$, and let $(T \otimes T)^\beta$ be the β-th power of $T \otimes T$ for $\beta > 0$. Denote by $\mathrm{Range}(A)$ the range of an operator A. Theorem 2 below establishes convergence rates of our estimator; the proof can be found in Appendix C.3.

Theorem 2 (Convergence rates)**.** *Let Assumption 2 be satisfied. For g and θ as defined in Assumption 2, assume that $g \in \mathrm{Range}(T^\alpha)$ for $0 < \alpha \leq 1$, and that $\theta \in \mathrm{Range}((T \otimes T)^\beta)$ for $0 < \beta \leq 1$. Then for $\varepsilon_n = cn^{-1/(1+\beta+\max(1-\alpha,\alpha))}$ with $c > 0$ being arbitrary but independent of n, we have*

$$\left\| \widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mu_{Y\langle 0|1\rangle} \right\|_{\mathcal{F}} = O_p\left(n^{-(\alpha+\beta)/(2(1+\beta+\max(1-\alpha,\alpha)))} \right) \quad (n \to \infty).$$

[4] More precisely, each element in $L_2(\mathbb{P}_{X_0})$ is a $\mathbb{P}_{X_0}$-equivalent class of functions; see Appendix C.1.

[5] For Matérn kernels, this holds since the resulting RKHSs are norm-equivalent to Sobolev spaces, which contain all functions in the RKHSs of Gaussian kernels.

Algorithm 1 Kernel Policy Evaluation (KPE) for kernels k and ℓ

```
Input: a sample $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ from $\pi_0$ and a sample $\{(\mathbf{u}_j^*, \mathbf{a}_j^*)\}_{j=1}^m$ from $\pi_*$.
Compute $\boldsymbol{\beta} = (\mathbf{K} + n\epsilon I)^{-1}\widetilde{\mathbf{K}}\mathbf{1}_m$ for $\mathbf{K}_{ij} = k((\mathbf{u}_i, \mathbf{a}_i), (\mathbf{u}_j, \mathbf{a}_j))$, $\widetilde{\mathbf{K}}_{ij} = k((\mathbf{u}_i, \mathbf{a}_i), (\mathbf{u}_j^*, \mathbf{a}_j^*))$
Output: the kernel mean embedding $\hat{\mu}_{P_*(r)} = \sum_{i=1}^n \beta_i \ell(r_i, \cdot)$.
```

In the condition $g \in \mathrm{Range}(T^\alpha)$, the constant α can be considered as quantifying the smoothness of the function g: As α increases, g gets smoother.[6] Since g is the Radon-Nikodym derivative of $\mathbb{P}_{X_1}$ w.r.t. $\mathbb{P}_{X_0}$, α being large (or g being smooth) implies that the two distributions $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_1}$ are similar, and vice versa. Similarly, the constant β quantifies the smoothness of the function θ.

Based on the above interpretation of the assumptions, let us interpret the rate of Theorem 2. Assume that α is very close to 0, meaning that g may be non-smooth. Even in this case, if β is large, that is if θ is smooth, we can still guarantee a certain rate of convergence. A similar argument holds for the case when β is close to 0: If α is large, then our estimator converges at a reasonable rate. In other words, Theorem 2 states that our estimator is guaranteed to converge reasonably fast, if *either* the Radon-Nikodym derivative $g = \mathrm{d}\mathbb{P}_{X_1} / \mathrm{d}\mathbb{P}_{X_0}$ *or* the function θ is smooth. This property resembles that of *doubly robust estimators* [7, 10]. Therefore, even in the situation where the change from $\mathbb{P}_{X_1}$ to $\mathbb{P}_{X_0}$ is large, we may still expect a good performance with our estimator, given the relationship between X_0 and Y_0 is smooth. This will also be experimentally demonstrated.

3.3 Application: Off-policy evaluation

We demonstrate the effectiveness of our estimator in *off-policy evaluation* task. Let $\mathcal{U}$ be a space of user (or context) features, $\mathcal{A}$ be a space of treatments, and $\pi : \mathcal{U} \to \mathcal{A}$ be a stochastic policy which selects a treatment $\mathbf{a} \in \mathcal{A}$ given a user $\mathbf{u} \in \mathcal{U}$. Given a target policy π_*, off-policy evaluation generally aims to provide an unbiased estimate of its performance. It is assumed that we have access to logged data $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ obtained from an initial policy π_0 where $\mathbf{u}_i$ represent the user features, $\mathbf{a}_i \sim \pi_0(\mathbf{u}_i)$ are the selected treatments (*e.g.*, recommendations) given $\mathbf{u}_i$, and $r_i \sim \mathbb{P}_0(r \,|\, \mathbf{u}_i, \mathbf{a}_i)$ are the rewards, where $\mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a})$ is the conditional distribution of rewards given $\mathbf{u}$ and $\mathbf{a}$. The policy π_0 specifies how the set of recommendations, known as *slates*, are constructed given the user or context information. Finally, the reward r_i can simply be the number of clicks on the recommendation.

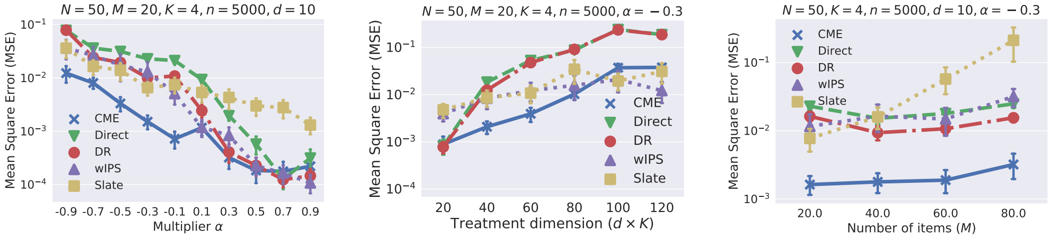
To adopt our framework, we assume that once the user feature $\mathbf{u}_i$ and the treatment $\mathbf{a}_i$ are specified, the reward distribution will be unchanged, *i.e.*, $\mathbb{P}_*(r \,|\, \mathbf{u}_i, \mathbf{a}_i) = \mathbb{P}_0(r \,|\, \mathbf{u}_i, \mathbf{a}_i)$, regardless of which policy produced them. (Here $\mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a})$ is the reward distribution when the policy is π^*.) Intuitively, we expect only the recommendations, but not the user behavior/response, to change as a result of a policy change. The covariate distribution may differ depending on the situation. Based on this assumption, the reward distribution $\mathbb{P}_*(r)$ under the target policy can be obtained as $\mathbb{P}_*(r) = \int \mathbb{P}_*(r \,|\, \mathbf{u}^*, \mathbf{a}^*) \,\mathrm{d}\mathbb{P}_*(\mathbf{u}^*, \mathbf{a}^*) = \int \mathbb{P}_0(r \,|\, \mathbf{u}^*, \mathbf{a}^*) \,\mathrm{d}\mathbb{P}_*(\mathbf{u}^*, \mathbf{a}^*)$. Since we have access to a sample $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ from π_0 and a sample $\{(\mathbf{u}_j^*, \mathbf{a}_j^*)\}_{j=1}^m$ from π_*, the embedding $\mu_{\mathbb{P}_*(r)}$ can be estimated directly using (4). That is, with the notation of Proposition 1, we let $\mathcal{X} := \mathcal{U} \times \mathcal{A}$, $\mathcal{Y} := \mathbb{R}$, $\mathbf{x}_i := (\mathbf{u}_i, \mathbf{a}_i)$, $\mathbf{y}_i := r_i$, $\mathbf{x}'_j := (\mathbf{u}_j^*, \mathbf{a}_j^*)$, $\mathbb{P}_{X_0} := \mathbb{P}_0$, $\mathbb{P}_{X_1} := \mathbb{P}_*$ and so on. By virtue of Lemma 1, we can interpret $\mu_{\mathbb{P}_*(r)}$ as the embedding of the actual counterfactual distribution. The resulting algorithm, which is very simple, is summarized in Algorithm 1.

4 Experiments

We compare our estimator to the following benchmark estimators in the off-policy evaluation task, using both simulated and real-world data. Below $\mathcal{D}_{\text{null}} := \{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ denotes logged data.

Direct Method (DM). The Direct method fits a regression model $\hat{\eta}(\mathbf{u}, \mathbf{a})$ for rewards r based on $\mathcal{D}_{\text{null}}$. The estimate is given by $\widehat{R}_{\text{DM}} = \frac{1}{n}\sum_{i=1}^n \sum_{\mathbf{a} \in \mathcal{A}} \pi_*(\mathbf{a}|\mathbf{u}_i)\hat{\eta}(\mathbf{u}_i, \mathbf{a})$ where $\pi_*(\mathbf{a}|\mathbf{u}_i)$ denotes the recommendation probabilities under the target policy. Note that $\hat{\eta}$ is typically biased toward the

[6] In fact, it is known that $\mathrm{Range}(T^\alpha)$ is norm-equivalent to a certain interpolation space between $L_2(\mathbb{P}_{X_0})$ and $\mathscr{H}$ for $0 < \alpha \leq 1/2$ [36, Thm. 4.6]; As α tends to 0 (resp. 1/2), $\mathrm{Range}(T^\alpha)$ tends to $L_2(\mathbb{P}_{X_0})$ (resp. $\mathscr{H}$); the situation $\alpha > 1/2$ is that g is smoother than least smooth functions in $\mathscr{H}$. A similar argument also applies to the interpretation of $\mathrm{Range}((T \otimes T)^\beta)$.



(a) Policy shift

(b) Treatment dimension

(c) Item size

Figure 1: Mean square error (MSE) of the expected reward estimated by different estimators as we vary the value of (a) the multiplier α, (b) the context dimension d while K is fixed, and (c) the number of available items. Each error bar represents a 95% confidence interval.

distribution $\mathbb{P}_0$ of $\mathcal{D}_{\text{null}}$. We used a 3-layer feedforward neural network as $\hat{\eta}$, for which the input feature vector is given by concatenating the vector of user $\mathbf{u}$ and all items in the recommendation $\mathbf{a}$.

Weighted Inverse Propensity Score (wIPS). The `wIPS` estimator obtains an unbiased estimate of the target reward by re-weighting each observation in the logged dataset by the ratio of *propensity scores* under the target and null policies [17]. The wIPS estimator is defined by $\widehat{R}_{\text{wIPS}} = (\sum_{i=1}^{n} w_i r_i)/(\sum_{i=1}^{n} w_i)$, where $w_i = \pi_*(\mathbf{a}_i|\mathbf{u}_i)/\pi_0(\mathbf{a}_i|\mathbf{u}_i)$ are the propensity weights.

Doubly Robust (DR). The `DR` estimator combines the two aforementioned estimators by exploiting both the regression model $\hat{\eta}(\mathbf{u}, \mathbf{a})$ and the propensity scores [7, 10]. The estimator is given by $\widehat{R}_{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{\mathbf{a}\in\mathcal{A}} \pi_*(\mathbf{a}|\mathbf{u}_i)\hat{\eta}(\mathbf{u}_i, \mathbf{a}) + w_i(r_i - \hat{\eta}(\mathbf{u}_i, \mathbf{a}_i))\right\}$.

Slate Estimator. The `Slate` estimator assumes that the reward value is linear w.r.t. a given recommendation [38]. It is defined as $\widehat{R}_{\text{slate}} = \frac{1}{n}\sum_{i=1}^{n} r_i(\mathbf{q}_{\mathbf{u}_i}^{\top}\Gamma_{\mathbf{u}_i}^{\dagger}\mathbf{1}_{\mathbf{a}_i})$, where $\mathbf{1}_{\mathbf{a}_i} \in \mathbb{R}^{KM}$ (K and M are the numbers of slots for recommendation and available items, respectively) is the indicator vector whose (k, m)-th element is 1 if the recommendation $\mathbf{a}_i$ contains the item m in the slot k, $\Gamma_{\mathbf{u}_i}^{\dagger}$ is the Moore-Penrose pseudoinverse of $\Gamma_{\mathbf{u}_i} := \mathbb{E}_{\pi_0}[\mathbf{1}_{\mathbf{a}}\mathbf{1}_{\mathbf{a}}^{\top}|\mathbf{u}_i]$, and $\mathbf{q}_{\mathbf{u}_i} := \mathbb{E}_{\pi_*}[\mathbf{1}_{\mathbf{a}}|\mathbf{u}_i]$. Because of the linearity assumption, the slate estimator has lower variance than `IPS` estimator; however, the estimator would suffer from bias if the assumption does not hold true.

For the `CME`, we used a kernel defined as $k((\mathbf{u}_i, \mathbf{a}_i), (\mathbf{u}_j, \mathbf{a}_j)) := k_1(\mathbf{u}_i, \mathbf{u}_\mathbf{j})k_2(\mathbf{a}_i, \mathbf{a}_j)$ where $k_1(\mathbf{u}_i, \mathbf{u}_j) := \exp\left(-\|\mathbf{u}_i - \mathbf{u}_j\|_2^2/2\sigma_u^2\right)$ and $k_2(\mathbf{a}_i, \mathbf{a}_j) := \exp\left(-\|\mathbf{a}_i - \mathbf{a}_j\|_2^2/2\sigma_a^2\right)$. To be able to compare to other estimators, we used $\ell(r_i, r_j) := \langle r_i, r_j\rangle$. The regularization parameter ε was selected by the cross validation procedure in Appendix B, while we determined σ_u and σ_a by the median heuristic, *i.e.*, $\sigma_u^2 = \text{median}\{\|\mathbf{u}_i - \mathbf{u}_j\|_2^2\}_{1\le i<j\le n}$ and $\sigma_a^2 = \text{median}\{\|\mathbf{a}_i - \mathbf{a}_j\|_2^2\}_{1\le i<j\le n}$.

4.1 Simulated data

As explained in §3.3, when a user visits a website, the system provides a recommendation as an ordered list of $K \in \mathbb{N}$ items out of $M \in \mathbb{N}$ available items to that user. Each item is represented by a feature vector $\mathbf{v}_m \sim \mathcal{N}(0, \sigma_v^2\mathbf{I}_d)$ for $m = 1, \ldots, M$. Hence, a recommendation is an order list $\mathbf{a}_i = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K)$, where $d \in \mathbb{N}$ is the dimensionality. Likewise, each user has a preference (or feature) vector $\mathbf{u}_j \sim \mathcal{N}(0, \sigma_u^2\mathbf{I}_d)$ for $j = 1, \ldots, N$ where N denotes the total number of users. The reward r_i is 1 if the user clicks any of the recommended items and 0 otherwise. Specifically, for each $(\mathbf{a}_i, \mathbf{u}_j)$ pair, let $\theta_{ij} = \mathbb{P}(\text{click}\,|\,\mathbf{a}_i, \mathbf{u}_j) = 1/(1 + \exp(-\bar{\mathbf{a}}_i^{\top}\mathbf{u}_j + \epsilon_{ij}))$ be the probability of a click, where $\bar{\mathbf{a}}_i$ is the mean vector of feature vectors for $\mathbf{a}_i$, and ϵ_{ij} is a Gaussian white noise. The reward of the recommendation $\mathbf{a}_i$ is defined as $r_i \sim \text{Bernoulli}(\theta_{ij})$.

For each user j a policy π generates the list of K recommended items by sampling without replacement with a multinomial distribution: The probability of item $\mathbf{v}_l$ being selected is $p_j(\mathbf{v}_l) := \exp(\mathbf{b}_j^{\top}\mathbf{v}_l)/\sum_{k=1}^{M}\exp(\mathbf{b}_j^{\top}\mathbf{v}_k)$, where $\mathbf{b}_j$ is the user preference vector of user j. For the target policy π_*, we set $\mathbf{b}_j^* = \mathbf{p}_j^{\top}\mathbf{u}_j$ for $j = 1, \ldots, N$ where $\mathbf{p}_j := (p_{jk})_{k=1}^{d}$ with $p_{jk} \sim \text{Bernoulli}(0.5)$. For the null policy π_0, we set $\mathbf{b}_j = \alpha\mathbf{b}_j^*$ where $\alpha \in [-1, 1]$.

After generating the item feature vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$, the datasets $\mathcal{D}_{\text{null}} = \{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^{n}$ and $\mathcal{D}_{\text{target}} = \{(\mathbf{u}_i^*, \mathbf{a}_i^*, r_i^*)\}_{i=1}^{n}$ were generated from π_0 and π_*, respectively. (Note that we only use r_i^*

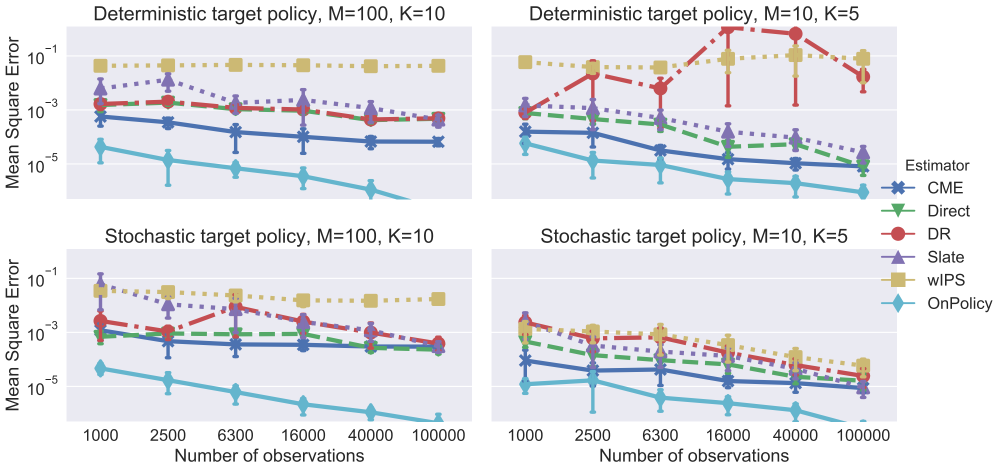


Figure 2: The performance of different estimators on the MSLR-WEB30K dataset.

for evaluation.) We set $N = 50$, $M = 20$, $K = 4$, $n = 5000$, and $d = 10$. We performed 5-fold CV over parameter grids, *i.e.*, the number of hidden units $n_h \in \{50, 100, 150, 200\}$ for the Direct and DR estimators, and the regularization parameter $\varepsilon \in \{10^{-8}, \dots, 10^{0}\}$ for our CME. We repeated experiments 30 times to obtain the mean squared error (MSE) for each estimator.

Figure 1 depicts the experimental results (note that vertical axis is in log scale). In brief, we found that (i) the performance of all estimators degrade as the difference between π_0 and π_* increases (*i.e.*, as α tends to -1), but the CME is least susceptible to this difference, (ii) the Slate estimator does not perform well in this setting because its assumptions do not hold, (iii) all estimators deteriorate as the context dimension increases, but the effect appears to be more pronounced for the Direct, DR, and CME estimators than for the IPS and Slate estimators as they do not rely directly on the context variables, (iv) the opposite effect is observed if we increase the number of available items M, as illustrated in Figure 1(c), and (v) the CME estimator achieves better performance than other estimators in most experiments. Supplementary results of this experiment can also be found in Appendix A.

4.2 Real data

For the real-world dataset, we use the data from the Microsoft Learning to Rank Challenge dataset (MSLR-WEB30K) [25] and treat them as an off-policy evaluation problem; the setup is similar to [38]. The data contains set of queries and corresponding URLs. Each query q and URL u pair is represented by a vector $f_{q,u}$ along with the relevant judgment $\rho(q,u) \in \{0, ..., 4\}$. For our reward function, we used the expected reciprocal rank (ERR) [8], which is defined by $\mathrm{ERR}(q,u) := \sum_{k=1}^{K} \frac{1}{k} \prod_{j=1}^{k-1}(1 - R(q,u_j))R(q,u_k)$, where $R(q,u) := \frac{2^{\rho(q,u)}-1}{2^{\mathrm{maxrel}}}$ with $\mathrm{maxrel} := 4$. For the null and target policies π_0 and π_*, the vector $f_{q,u}$ is split into URL feature f_{url} and body feature f_{body}, which are then used to train two regression models to fit $\rho(q,u)$: For π_0 the Lasso is used with f_{url} and denoted by $\mathrm{lasso}_{\text{url}}$, and for π_* the regression tree is used with f_{body} and denoted by $\mathrm{tree}_{\text{body}}$.

The logged data $\mathcal{D}_{\text{null}}$ is then generated as follows. We first sample query q uniformly from the dataset, and obtain top M candidate URLs based on the relevant scores predicted by $\mathrm{tree}_{\text{body}}$. The null policy π_0 then recommends K URLs out of these M candidates, according to the *Plackett-Luce model* parameterized by $p_\alpha(u|q) \propto 2^{-\alpha\lfloor \log_2 \mathrm{rank}(u,q) \rfloor}$, where $\mathrm{rank}(u,q)$ is the rank of the relevant score predicted by the $\mathrm{tree}_{\text{body}}$ model and $\alpha >= 0$ is an exploration rate. For $\mathcal{D}_{\text{target}}$, the target policy π_* employs the same setting as the null policy π_0, except that the predicted relevant scores are obtained from the $\mathrm{lasso}_{\text{url}}$ model. In this experiment, we set $\alpha = 1$ for the null policy, and consider both *deterministic* and *stochastic* target policies. For the stochastic policy, we set $\alpha = 2$, while the deterministic policy selects the top-K URLs directly from the predicted relevant scores.

In this experiment, we used for the Direct method the regression tree, instead of a neural network. In addition, we included the OnPolicy method as a baseline, which estimates rewards directly from the *target* policies (and thus, this baseline should always perform the best). To accelerate the computation of the CME, we used the Nyström approximation method [40].

Figure 2 depicts the results. In short, our CME dominates other estimators in most of experiment conditions. (Note also here that vertical axis is in log scale, so the margins are significantly large.) The wIPS clearly suffers from high variance, especially in the deterministic target policy, and DR also suffers from the same issue, except in the top left condition. This would be because, in the deterministic policy setup, the propensity score adjustment requires an exact match between logged and target treatments, but this almost never happens when the treatment space is large. The Slate, Direct and CME are relatively robust across different conditions. The Direct method and CME perform particularly well when sample size is small, regardless of the treatment space, while the Slate estimator requires larger samples, especially in the large treatment space.

5 Discussion

Our estimator of counterfactual distribution exhibits appealing theoretical properties, and also serves as a practical tool for causal inference. Ultimately, we hope that our work can be useful not only for researchers in disciplines such as social science, epidemiology, and econometrics that rely on the potential outcome framework, but also for the driving machine learning community to solve this challenging problem, because several open questions still remain, *e.g.*, the use of high-order moments of counterfactual distribution, and how to handle a hidden confounder and an instrumental variable.

Acknowledgments

We thank Ricardo Silva, Joris Mooij, Adith Swaminathan, Evan Robin, David Lopez-Paz, Wittawat Jitkrittum, Kenji Fukumizu, and Bernhard Schölkopf for fruitful discussions. Krikamol Muandet would like to acknowledge fundings from the Thailand Research Fund Grant No. MRG6080206 and additional funding from the Faculty of Science, Mahidol University. Motonobu Kanagawa has been partially supported by the European Research Council (StG project PANAMA).

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[2] C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:pp. 273–289, 1973.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

[4] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[5] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[6] L. Bottou, J. Peters, J. Q. nonero Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

[7] C. M. Cassel, C. E. Särndal, and J. H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

[8] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[9] V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[10] M. Dudík, J. Langford, and L. Li. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

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

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

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[14] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1414–1423. PMLR, 2017.

[15] J. J. Heckman and R. Robb. Alternative methods for evaluating the impact of interventions. *Journal of Econometrics*, 30(1):239–267, 1985.

[16] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[17] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[18] G. W. Imbens. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29, 2004.

[19] F. D. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML'16, pages 3020–3029. JMLR.org, 2016.

[20] J. Langford, A. Strehl, and J. Wortman. Exploration scavenging. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 528–535, New York, NY, USA, 2008. ACM.

[21] D. Liang, L. Charlin, and D. M. Blei. Causal inference for recommendation. 2016.

[22] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 1452–1461. JMLR, 2015.

[23] J. A. F. Machado and J. Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20(4):445–465, 2005.

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

[25] T. Qin and T. Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013.

[26] P. R. Rosenbaum. *Observational studies*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 2002.

[27] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[28] D. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[29] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1670–1679, New York, New York, USA, 2016. PMLR.

[30] U. Shalit, F. Johansson, and D. Sontag. Bounding and minimizing counterfactual error. arXiv:1606.03976 Preprint, 2016.

[31] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer-Verlag, 2007.

[32] L. Song. *Learning via Hilbert Space Embedding of Distributions*. PhD thesis, The University of Sydney, 2008.

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

[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 International Conference on Machine Learning (ICML)*, June 2009.

[35] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

[36] I. Steinwart and C. Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHS. *Constructive Approximation*, 35(363-417), 2012.

[37] A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 814–823. JMLR.org, 2015.

[38] A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems 30*, pages 3632–3642. Curran Associates, Inc., 2017.

[39] I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18:86:1–86:47, 2017.

[40] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

[41] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 819–827, Atlanta, Georgia, USA, 2013. PMLR.

[42] L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, volume 3120 of *Lecture Notes in Computer Science*, pages 594–608. Springer, 2004.

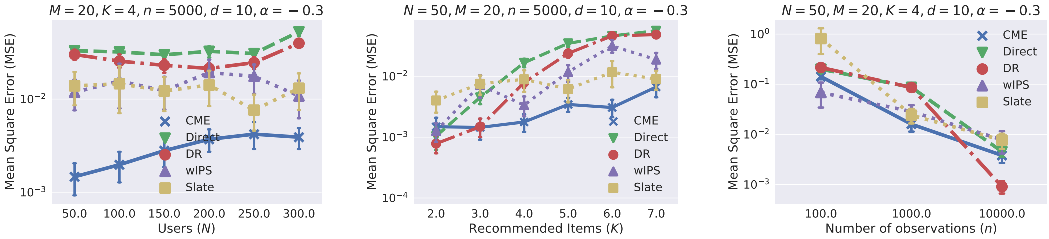
Appendix

A Experimental results

In this section, we provide additional results from extensive experimental studies presented in §4.

A.1 Simulated data

In this section, we investigate the behavior of different estimators as we vary the number of users N, the number of recommended items K, and the number of observations n. Figure 3 depicts the results.



(a) Users

(b) Recommended items

(c) Sample size

Figure 3: Mean square error (MSE) of the expected reward estimated by different estimators as we vary the value of (a) the number of users N, (b) the number of recommended items K, and (c) the number of observations n. Each error bar represents a 95% confidence interval.

B Cross validation procedure for counterfactual prediction

One of the key challenges in counterfactual prediction is to perform model selection. Unlike standard cross validation, performing cross validation directly on the logged data results in the wrong choice of parameters. That is, the estimate of the performance measure will always be biased. This is obviously due to the fundamental problem of causal inference. To this end, given a dataset $\mathcal{D}$ and a parameter grid $\mathcal{P}$, we propose the following general procedure for parameter selection.

1. Split $\mathcal{D}$ into K folds: $\mathcal{D}_k = \{(\mathbf{x}_j, \mathbf{s}_j, r_j)\}_{j=q(k-1)+1}^{qk}$ for $k = 1, \dots, K$ and $q = \lfloor n/K \rfloor$.
2. For each parameter $p = 1, 2, \dots, |\mathcal{P}|$:
   (a) For each fold $k = 1, 2, \dots, K$:
      i. Calculate $\{w_j\}_{j=1}^{q}$ using propensity scores or covariate matching.
      ii. Re-weight the validation reward $\hat{r}_k^* = \sum_{j=1}^{q} w_j r_{q(k-1)+j}$ (**bias correction**).
      iii. Use the remaining logged data $\mathcal{D}_{\neg k}$ and validation data $\{(\mathbf{x}_j^*, \mathbf{s}_j^*)\}_{j=q(k-1)+1}^{qk}$ to compute the estimated reward $\hat{r}_k$ and corresponding error $e_k = (\hat{r}_k - \hat{r}_k^*)^2$.
   (b) Calculate the mean CV error $\bar{e}_p = \frac{1}{K}\sum_{k=1}^{K} e_k$ (**variance reduction**).
3. Pick the p-th parameter setting whose $\bar{\varepsilon}_p$ is smallest.

The algorithm above follows the standard cross validation procedure, except the bias correction step on validation sets. In the bias correction step, we re-weight the sample in the validation set so that the performance estimate computed from this set is unbiased. Nevertheless, the estimate may have high variance, *e.g.*, when the propensity weights are used. This pitfall is alleviated by the variance reduction step.

C Proofs for theoretical results

In this section, we collect proofs for theoretical results presented so far. To this end, we need to introduce certain concepts such as kernel integral operators.

C.1 Preliminaries

Basic definitions and notation. Let $\mathcal{X}$ be a measurable space and P_{X_0} be a probability measure on $\mathcal{X}$, and denote by $L_2(P_{X_0})$ the Hilbert space of square-integrable functions with respect to P_{X_0}. Similarly, let $P_{X_0} \otimes P_{X_0}$ denote the product measure of P_{X_0} and P_{X_0} defined on the product space $\mathcal{X} \times \mathcal{X}$, and $L_2(P_{X_0} \otimes P_{X_0})$ be the Hilbert space of square integrable functions w.r.t. $P_{X_0} \otimes P_{X_0}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a measurable positive definite kernel, and $\mathscr{H}$ be the RKHS associated with k.

To be rigorous, we will use the following notation due to [36]: For a function $f : \mathcal{X} \to \mathbb{R}$, let $[f]_\sim$ denote the class of functions that are P_{X_0}-equivalent to f:

$$[f]_\sim := \{g \in L_2(P_{X_0}) : \; P_{X_0}(f \neq g) = 0\}\,.$$

We will assume the following property for k and P_{X_0}.

Assumption 3. *The kernel k and and probability measure P_{X_0} satisfy*

$$\int k(x,x)\,\mathrm{d}P_{X_0}(x) < \infty.$$

Integral operators. Define three integral operators $T : L_2(P_{X_0}) \to L_2(P_{X_0})$, $S : L_2(P_{X_0}) \to \mathscr{H}$ and $\mathcal{C}_{XX} : \mathscr{H} \to \mathscr{H}$ by

$$Tf := \int k(\cdot, x) f(x)\,\mathrm{d}P_{X_0}(x) \in L_2(P_{X_0}), \qquad f \in L_2(P_{X_0}), \tag{5}$$

$$Sf := \int k(\cdot, x) f(x)\,\mathrm{d}P_{X_0}(x) \in \mathscr{H}, \qquad f \in L_2(P_{X_0}), \tag{6}$$

$$\mathcal{C}_{XX} g := \int k(\cdot, x) f(x)\,\mathrm{d}P_{X_0}(x) \in \mathscr{H}, \qquad g \in \mathscr{H},. \tag{7}$$

Note that while these operators look similar, they are different in their domains and ranges. In particular, $\mathcal{C}_{XX}$ is the covariance operator based on which our estimator is defined. Under Assumption 3, [36, Lemma 2.3] implies that the operator $S^* : \mathscr{H} \to L_2(P_{X_0})$ defined by

$$S^* g = [g]_\sim, \quad g \in \mathscr{H}$$

is compact, and thus continuous. This operator S^* is the adjoint of the operator S defined in (6). Since S^* is continuous, by [36, Lemma 2.3], the operators T and $\mathcal{C}_{XX}$ can be written as

$$T = S^* S, \quad \mathcal{C}_{XX} = SS^*.$$

The following lemma summarizes conditions required for eigen-decompositions of (5), (6) and (7).

Lemma 2 (Spectral decomposition of integral operators)**.** *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. There exists a family $(e_i)_{i=1}^\infty \subset \mathscr{H}$ and $(\mu_i)_{i=1}^\infty \subset (0,\infty)$ such that $(\mu_i^{1/2} e_i)_{i=1}^\infty$ is an ONS in $\mathscr{H}$, $([e_i]_\sim)_{i=1}^\infty$ is an ONS in $L_2(P_{X_0})$, and*

$$Tf = \sum_{i=1}^\infty \mu_i \left\langle [e_i]_\sim, f \right\rangle_{L_2(P_{X_0})} [e_i]_\sim, \qquad f \in L_2(P_{X_0}), \tag{8}$$

$$Sf = \sum_{i=1}^\infty \mu_i \left\langle [e_i]_\sim, f \right\rangle_{L_2(P_{X_0})} e_i, \qquad f \in L_2(P_{X_0}), \tag{9}$$

$$\mathcal{C}_{XX} g = \sum_{i=1}^\infty \mu_i \left\langle \mu_i^{1/2} e_i, g \right\rangle_{\mathscr{H}} \mu_i^{1/2} e_i, \qquad g \in \mathscr{H}, \tag{10}$$

where the convergence is in $L_2(P_{X_0})$ for (8), and in $\mathscr{H}$ for (9) and (10).

Proof. Since k and P_{X_0} satisfy Assumption 3, it follows from [36, Lemma 2.3] that $\mathscr{H}$ is compactly embedded into $L_2(P_{X_0})$. As a result, [36, Lemma 2.12] implies that there exists a family $(e)_{i=1}^\infty \subset \mathscr{H}$ and $(\mu_i)_{i=1}^\infty \subset (0,\infty)$ such that $([e_i]_\sim)_{i=1}^\infty$ is an ONS in $L_2(P_{X_0})$, $(\mu_i^{1/2} e_i)_{i=1}^\infty$ is an ONS in $\mathscr{H}$, and (8) holds with convergence in $L_2(P_{X_0})$.

We next show (9). Since $([e_i]_\sim)_{i=1}^\infty$ is an ONS in $L_2(P_{X_0})$, any $f \in L_2(P_{X_0})$ can be written as

$$f = \sum_{i=1}^\infty \langle [e_i]_\sim, f \rangle_{L_2(P_{X_0})} [e_i]_\sim + f^\perp,$$

with convergence in $L_2(P_{X_0})$, where $f^\perp \in L_2(P_{X_0})$ is such that $\left\langle [e_i]_\sim, f^\perp \right\rangle_{L_2(P_{X_0})} = 0$ for all i. Since by [36, Lemma 2.12, Eq.15] we have $\mu_i e_i = S[e_i]_\sim$ for all i, it then holds that

$$Sf = \sum_{i=1}^\infty \mu_i \langle [e_i]_\sim, f \rangle_{L_2(P_{X_0})} e_i + Sf^\perp,$$

where the convergence is in $\mathscr{H}$ since S is continuous. Note that we have $Tf^\perp = 0$, since have (8) and $\left\langle [e_i]_\sim, f^\perp \right\rangle_{L_2(P_{X_0})} = 0$ for all i. This implies that $f^\perp$ is in the null space of T. Since the null spaces of S and T are equal [36, Lemma 2.12, Eq.16], it follows that $Sf^\perp = 0$, which implies (9).

Finally we show (10). First note that $\mathcal{C}_{XX} e_i = SS^* e_i = S[e_i]_\sim = \mu e_i$ for all i. Using this and (9), for any $g \in \mathscr{H}$ we have

$$\begin{aligned} \mathcal{C}_{XX} g &= SS^* g = \sum_{i=1}^\infty \mu_i \langle [e_i]_\sim, S^* g \rangle_{L_2(P_{X_0})} e_i = \sum_{i=1}^\infty \mu_i \langle SS^* e_i, g \rangle_{L_2(P_{X_0})} e_i \\ &= \sum_{i=1}^\infty \mu_i \langle \mathcal{C}_{XX} e_i, g \rangle_{\mathscr{H}} e_i = \sum_{i=1}^\infty \mu_i \langle \mu_i e_i, g \rangle_{\mathscr{H}} e_i, \end{aligned}$$

where the convergence is in $\mathscr{H}$, which implies (10).

□

Definition 2. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Let $(e_i)_{i=1}^\infty \subset \mathscr{H}$ and $(\mu_i)_{i=1}^\infty \subset (0,\infty)$ be as in Lemma 2. Then for a constant $\beta > 0$, the β-th powers of T, S and $\mathcal{C}_{XX}$ are respectively defined by*

$$\begin{aligned} T^\beta f &:= \sum_{i=1}^\infty \mu_i^\beta \langle [e_i]_\sim, f \rangle_{L_2(P_{X_0})} [e_i]_\sim, & f \in L_2(P_{X_0}), \\ S^\beta f &:= \sum_{i=1}^\infty \mu_i^\beta \langle [e_i]_\sim, f \rangle_{L_2(P_{X_0})} e_i, & f \in L_2(P_{X_0}), \\ \mathcal{C}_{XX}^\beta f &:= \sum_{i=1}^\infty \mu_i^\beta \left\langle \mu_i^{1/2} e_i, f \right\rangle_{\mathscr{H}} \mu_i^{1/2} e_i, & f \in \mathscr{H}. \end{aligned}$$

Lemma 3. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Let $(e_i)_{i=1}^\infty \subset \mathscr{H}$ and $(\mu_i)_{i=1}^\infty \subset (0,\infty)$ be as in Lemma 2. Assume that the mapping $S^* : \mathscr{H} \to L_2(P_{X_0})$ has a dense image in $L_2(P_{X_0})$. Then $([e]_\sim)_{i=1}^\infty$ forms an ONB of $L_2(P_{X_0})$.*

Proof. Since k and P_{X_0} satisfy Assumption 3, it follows from [36, Lemma 2.3] that $\mathscr{H}$ is compactly embedded into $L_2(P_{X_0})$. Then one can use [36, Theorem 3.1], which states that the assertion is equivalent to the assumption that the embedding $S^* : \mathscr{H} \to L_2(P_{X_0})$ has a dense image in $L_2(P_{X_0})$. □

Finally we define an integral operator $T \otimes T : L_2(P_{X_0} \otimes P_{X_0}) \to L_2(P_{X_0} \otimes P_{X_0})$ by, for any $\eta \in L_2(P_{X_0} \otimes P_{X_0})$,

$$(T \otimes T \eta)(x_1, x_2) := \int\int k(x_1, \tilde{x}_1) k(x_2, \tilde{x_2}) \eta(\tilde{x}_1, \tilde{x}_2) dP_{X_0}(\tilde{x}_1) dP_{X_0}(\tilde{x}_2), \quad x_1, x_2 \in \mathcal{X}.$$

This is an integral operator in $L_2(P_{X_0} \otimes P_{X_0})$ defined with the product measure $P_{X_0} \otimes P_{X_0}$ and the product kernel $k \otimes k : (\mathcal{X} \times \mathcal{X}) \times (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$ defined by

$$k \otimes k((x_1, x_2), (\tilde{x_1}, \tilde{x_2})) := k(x_1, \tilde{x_1})k(x_2, \tilde{x_2}), \quad (x_1, x_2), (\tilde{x}_1, \tilde{x}_2) \in \mathcal{X}. \times \mathcal{X}$$

For $\beta > 0$, let $(T \otimes T)^\beta$ be the β-th power of $T \otimes T$ for $\beta > 0$. This operator has the following property.

Lemma 4. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Then for any $\beta > 0$, we have*

$$(T \otimes T)^\beta (f \otimes g) = (T^\beta f) \otimes (T^\beta g), \quad f, g \in L_2(P_{X_0}).$$

Proof. Let $(e_i)_{i=1}^\infty \subset \mathscr{H}$ and $(\mu_i)_{i=1}^\infty \subset (0, \infty)$ be as in Lemma 2. By Assumption 3, we have

$$\int k \otimes k((x, \tilde{x}), (x, \tilde{x})) dP_{X_0} \otimes P_{X_0}(x, \tilde{x}) = \left(\int k(x, x) dP_{X_0}(x) \right)^2 < \infty.$$

Therefore $T \otimes T$ admits an eigen-decomposition from Lemma 2. It is easy to show that this eigen-decomposition is given by

$$(T \otimes T)\eta = \sum_{i=1}^\infty \sum_{j=1}^\infty \mu_i \mu_j \left\langle [e_i]_\sim \otimes [e_j]_\sim, \eta \right\rangle_{L_2(P_{X_0} \otimes P_{X_0})} [e_i]_\sim \otimes [e_j]_\sim, \quad \eta \in L_2(P_{X_0} \otimes P_{X_0}),$$

where the convergence is in $L_2(P_{X_0} \otimes P_{X_0})$. Thus, the β-th power of $T \otimes T$ can be written as

$$(T \otimes T)^\beta \eta = \sum_{i=1}^\infty \sum_{j=1}^\infty \mu_i^\beta \mu_j^\beta \left\langle [e_i]_\sim \otimes [e_j]_\sim, \eta \right\rangle_{L_2(P_{X_0} \otimes P_{X_0})} [e_i]_\sim \otimes [e_j]_\sim, \quad \eta \in L_2(P_{X_0} \otimes P_{X_0}).$$

Therefore, for any $f, g \in L_2(P_{X_0})$,

$$\begin{aligned} & (T \otimes T)^\beta (f \otimes g) \\ = \; & \sum_{i=1}^\infty \sum_{j=1}^\infty \mu_i^\beta \mu_j^\beta \left\langle [e_i]_\sim \otimes [e_j]_\sim, f \otimes g \right\rangle_{L_2(P_{X_0} \otimes P_{X_0})} [e_i]_\sim \otimes [e_j]_\sim, \\ = \; & \sum_{i=1}^\infty \sum_{j=1}^\infty \mu_i^\beta \mu_j^\beta \left\langle [e_i]_\sim, f \right\rangle_{L_2(P_{X_0})} \left\langle [e_j]_\sim, g \right\rangle_{L_2(P_{X_0})} [e_i]_\sim \otimes [e_j]_\sim, \\ = \; & \left(\sum_{i=1}^\infty \mu_i^\beta \left\langle [e_i]_\sim, f \right\rangle_{L_2(P_{X_0})} [e_i]_\sim \right) \otimes \left(\sum_{j=1}^\infty \mu_j^\beta \left\langle [e_j]_\sim, g \right\rangle_{L_2(P_{X_0})} [e_j]_\sim \right), \\ = \; & (T^\beta f) \otimes (T^\beta g). \end{aligned}$$

□

Motivated by Lemma 4, we will use the notation $T^\beta \otimes T^\beta := (T \otimes T)^\beta$ in Appendix C.3.

C.2 Proof of Theorem 1

Our proof relies on several lemmas, which are collected and proven in Appendix C.4.

Proof of Theorem 1. By the triangle inequality, we can bound the error of our estimator as

$$\begin{aligned} & \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon_n I)^{-1} \hat{\mu}_{X_1} - \mu_{Y\langle 0|1 \rangle}\|_{\mathcal{F}} \\ \leq \; & \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon_n I)^{-1} \hat{\mu}_{X_1} - \mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon_n I)^{-1} \mu_{X_1}\|_{\mathcal{F}} & (11) \\ + \; & \|\mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon_n I)^{-1} \mu_{X_1} - \mu_{Y\langle 0|1 \rangle}\|_{\mathcal{F}}, & (12) \end{aligned}$$

where Eq. (11) can be interpreted as the estimation (statistical) error and Eq. (12) as the approximation error. The estimation error (11) can be shown to converge to 0 in probability as the regularization constant ε_n decays to 0 sufficiently slowly, using the exactly same argument as in the proof of Theorem 8 in [12]. Therefore we omit the proof for the estimation error.

Here we aim to prove that the approximation error (12) goes to zero as $\varepsilon_n \to 0$. Note that to this end, we cannot apply the proof of Theorem 8 in [12], since it relies on stronger assumptions than ours. We do this by using Lemma 11, which shows that the approximation error can be written as

$$\begin{aligned} & \|\mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|^2_{\mathcal{F}} \\ = \; & \langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} & (13) \\ & -2 \left\langle g, (T + \varepsilon_n I)^{-1} T \, \mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0) | X_0 = \cdot] \right\rangle_{L_2(P_{X_0})} & (14) \\ & + \int\int \theta(x, \tilde{x}) dP_{X_1}(x) dP_{X_1}(\tilde{x}), \end{aligned}$$

where $g_{\varepsilon_n} = (T + \varepsilon_n I)^{-1} T g$. Below we show the convergence limits of (13) and (14) as $\varepsilon_n \to 0$, which conclude the proof.

Convergence of (13). We will show that

$$\langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} \to \int\int \theta(x, \tilde{x}) dP_{X_1}(x) dP_{X_1}(\tilde{x}) \quad (\varepsilon_n \to 0). \tag{15}$$

Note that we have

$$\begin{aligned} \langle g \otimes g, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} &= \int\int g(x) g(\tilde{x}) \theta(x, \tilde{x}) dP_{X_0}(x) dP_{X_0}(\tilde{x}) \\ &= \int\int \theta(x, \tilde{x}) dP_{X_1}(x) dP_{X_1}(\tilde{x}). \end{aligned}$$

Therefore it suffices to show that

$$\langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} \to \langle g \otimes g, \theta \rangle_{L_2(P_{X_0}} \quad (\varepsilon_n \to 0).$$

To this end, note that by the Cauchy-Schwartz inequality, we have

$$\begin{aligned} & \left| \langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} - \langle g \otimes g, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} \right| \\ = \; & \left| \langle g_{\varepsilon_n} \otimes g_{\varepsilon_n} - g \otimes g, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} \right| \\ \leq \; & \|g_{\varepsilon_n} \otimes g_{\varepsilon_n} - g \otimes g\|_{L_2(P_{X_0} \otimes P_{X_0})} \|\theta\|_{L_2(P_{X_0} \otimes P_{X_0})} \end{aligned}$$

Thus we focus on showing that

$$\|g_{\varepsilon_n} \otimes g_{\varepsilon_n} - g \otimes g\|_{L_2(P_{X_0} \otimes P_{X_0})} \to 0 \quad (\varepsilon_n \to 0). \tag{16}$$

First, by the triangle inequality, the left hand side of the above equation can be upper-bounded as

$$\begin{aligned} & \|g_{\varepsilon_n} \otimes g_{\varepsilon_n} - g \otimes g\|_{L_2(P_{X_0} \otimes P_{X_0})} \\ \leq \; & \|g_{\varepsilon_n} \otimes g_{\varepsilon_n} - g \otimes g_{\varepsilon_n}\|_{L_2(P_{X_0} \otimes P_{X_0})} + \|g \otimes g_{\varepsilon_n} - g \otimes g\|_{L_2(P_{X_0} \otimes P_{X_0})}. \end{aligned} \tag{17}$$

The first term of Eq. (17) can be written as

$$\begin{aligned} & \|g_{\varepsilon_n} \otimes g_{\varepsilon_n} - g \otimes g_{\varepsilon_n}\|_{L_2(P_{X_0} \otimes P_{X_0})} \\ = \; & \|(g_{\varepsilon_n} - g) \otimes g_{\varepsilon_n}\|_{L_2(P_{X_0} \otimes P_{X_0})} \\ = \; & \|g_{\varepsilon_n} - g\|_{L_2(P_{X_0})} \|g_{\varepsilon_n}\|_{L_2(P_{X_0})} \to 0 \quad (\varepsilon_n \to 0) \quad (\because \text{Lemma } 9), \end{aligned}$$

Similarly, the second term of Eq. (17) can be written as

$$\begin{aligned} & \|g \otimes g_{\varepsilon_n} - g \otimes g\|_{L_2(P_{X_0} \otimes P_{X_0})} \\ = \; & \|g \otimes (g_{\varepsilon_n} - g)\|_{L_2(P_{X_0} \otimes P_{X_0})} \\ = \; & \|g\|_{L_2(P_{X_0})} \|g_{\varepsilon_n} - g\|_{L_2(P_{X_0})} \to 0 \quad (\varepsilon_n \to 0) \quad (\because \text{Lemma } 9). \end{aligned}$$

We have shown (16), which concludes (15).

Convergence of (14). Next, we will show that

$$\left\langle g, (T+\varepsilon_n I)^{-1} T\, \mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot] \right\rangle_{L_2(P_{X_0})} \to \int\int \theta(x,\tilde{x}) dP_{X_1}(x) dP_{X_1}(\tilde{x}) \quad (\varepsilon_n \to 0). \tag{18}$$

From Lemma 9, as $\varepsilon_n \to 0$, the left hand side converges to

$$\begin{aligned} \left\langle g, \mathbb{E}[\mu_{Y\langle 0|1\rangle}(Y)|X=\cdot] \right\rangle_{L_2(P_{X_0})} &= \left\langle g, \int \theta(\cdot,\tilde{x}) dP_{X_1}(\tilde{x}) \right\rangle_{L_2(P_{X_0})} \quad (\because (36)) \\ &= \int\int \theta(x,\tilde{x}) dP_{X_1}(\tilde{x}) g(x) dP_{X_0}(x) \\ &= \int\int \theta(x,\tilde{x}) dP_{X_1}(x) dP_{X_1}(\tilde{x}). \end{aligned}$$

Thus we have shown (18). The proof completes by substituting (15) and (18) in (13) and (14) respectively.

□

C.3 Proof of Theorem 2

As in the previous section, the proof relies on lemmas collected in Appendix C.4. As mentioned in Appendix C.1, we will use the notation $T^\beta \otimes T^\beta := (T \otimes T)^\beta$, motivated by Lemma 4.

Proof of Theorem 2. By the triangle inequality we can bound the error of our estimator as

$$\begin{aligned} & \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}} \\ \leq\ & \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}} && (19) \\ +\ & \|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}}, && (20) \end{aligned}$$

where (19) is the estimation error, and (20) is the approximation error. We will derive convergence rates for these two types of error separately in the following, and then determine the optimal schedule for the decay of the regularization constant ε_n as $n \to \infty$.

Rate for the estimation error (19) We will show that the estimation error (19) decays at the rate

$$\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}} = O_p(n^{-1/2}\varepsilon^{\min(-1+\alpha,-1/2)}) \quad (n\to\infty,\ \varepsilon_n \to 0). \tag{21}$$

First, as in the proof of [12, Theorem 11], the left side can be upper-bounded as

$$\begin{aligned} & \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}} \\ \leq\ & \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}(\hat{\mu}_{X_1}-\mu_{X_1})\|_{\mathcal{F}} + \|(\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}} \\ & + \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}(\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}} && (22) \end{aligned}$$

Note that, by a classic result by Baker [2], $\widehat{\mathcal{C}}_{YX}$ can be decomposed as $\widehat{\mathcal{C}}_{YX} = \widehat{\mathcal{C}}_{YY}^{1/2}\widehat{\mathcal{W}}_{YX}\widehat{\mathcal{C}}_{XX}^{1/2}$ for a bounded linear operator $\widehat{\mathcal{W}}_{YX} : \mathscr{H} \to \mathcal{F}$ with $\|\widehat{\mathcal{W}}_{YX}\| \leq 1$. Therefore

$$\begin{aligned} \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\| &= \|\widehat{\mathcal{C}}_{YY}^{1/2}\widehat{\mathcal{W}}_{YX}\widehat{\mathcal{C}}_{XX}^{1/2}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\| \\ &\leq \|(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1/2}\| \leq \varepsilon_n^{-1/2}. && (23) \end{aligned}$$

Thus, the rate of the first term in (22) is

$$\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}(\hat{\mu}_{X_1}-\mu_{X_1})\|_{\mathcal{F}} \leq \varepsilon^{-1/2}\|\hat{\mu}_{X_1}-\mu_{X_1}\|_{\mathcal{H}} = O_p(\varepsilon^{-1/2}n^{-1/2}).$$

Next, the rate of the second term in (22) is given by

$$\begin{aligned} & \|(\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}} \\ \leq\ & \|\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX}\|\|(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{H}} \\ \leq\ & \|\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX}\| c_\alpha \varepsilon_n^{\min(-1/2+\alpha,0)} \quad (\because \text{ Lemma } 8) \\ =\ & O_p(n^{-1/2}\varepsilon_n^{\min(-1/2+\alpha,0)}), \end{aligned}$$

where c_α is a constant depending only on α and g. Finally, for the third term in (22), the rate is given as

$$\begin{aligned}
& \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}(\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}} \\
\leq\ & \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\|\|\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX}\|\|(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{H}} \\
\leq\ & \varepsilon_n^{-1/2}\|\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX}\|c_\alpha \varepsilon_n^{\min(-1/2+\alpha,0)} \quad (\because \text{ (23) and Lemma 8}) \\
=\ & O_p(n^{-1/2}\varepsilon_n^{\min(-1+\alpha,-1/2)}).
\end{aligned}$$

Since we will set ε_n so that $\varepsilon_n \to 0$ as $n \to \infty$, the rate of the third term is the slowest in the three terms in (22). Thus we have shown (21).

Rate for the approximation error (20) We will show that the approximation error decays at the rate

$$\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}-\mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}} = O(\varepsilon_n^{(\alpha+\beta)/2}) \quad (\varepsilon_n \to 0). \tag{24}$$

First note that, by the definitions of θ and g, we have

$$\int\int \theta(x,\tilde{x})dP_{X_1}(x)dP_{X_1}(\tilde{x}) = \left\langle g, \int \theta(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})} = \langle g\otimes g, \theta\rangle_{L_2(P_{X_0})\otimes L_2(P_{X_0})}. \tag{25}$$

Therefore, using Lemma 11, we can upper-bound the square of the approximation error (20) as

$$\begin{aligned}
& \|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}-\mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}}^2 \\
=\ & \langle g_{\varepsilon_n}\otimes g_{\varepsilon_n}, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})} \\
& -2\left\langle g, (T+\varepsilon_n I)^{-1}T\,\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot]\right\rangle_{L_2(P_{X_0})} \\
& +\int\int\theta(x,\tilde{x})dP_{X_1}(x)dP_{X_1}(\tilde{x}) \\
\leq\ & \left|\langle g_{\varepsilon_n}\otimes g_{\varepsilon_n}, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})} - \langle g\otimes g, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})}\right| \\
+\ & 2\left|\langle g\otimes g, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})} - \left\langle g, (T+\varepsilon_n I)^{-1}T\,\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot]\right\rangle_{L_2(P_{X_0})}\right|,
\end{aligned} \tag{26}$$

where $g_{\varepsilon_n} = (T+\varepsilon_n I)^{-1}Tg$.

Bound on the first term in (26). From the assumption $\theta \in \mathrm{Range}(T^\beta \otimes T^\beta)$, there exists a function $\eta \in L_2(P_{X_0}\otimes P_{X_0})$ such that $\theta = T^\beta \otimes T^\beta \eta$. We then have

$$\begin{aligned}
& \left|\langle g_{\varepsilon_n}\otimes g_{\varepsilon_n}, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})} - \langle g\otimes g, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})}\right| \\
=\ & \left|\langle g_{\varepsilon_n}\otimes g_{\varepsilon_n} - g\otimes g, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})}\right| \\
=\ & \left|\langle g_{\varepsilon_n}\otimes g_{\varepsilon_n} - g\otimes g, T^\beta\otimes T^\beta\eta\rangle_{L_2(P_{X_0}\otimes P_{X_0})}\right| \\
=\ & \left|\langle T^\beta g_{\varepsilon_n}\otimes T^\beta g_{\varepsilon_n} - T^\beta g\otimes T^\beta g, \eta\rangle_{L_2(P_{X_0}\otimes P_{X_0})}\right| \\
\leq\ & \left\|T^\beta g_{\varepsilon_n}\otimes T^\beta g_{\varepsilon_n} - T^\beta g\otimes T^\beta g\right\|_{L_2(P_{X_0}\otimes P_{X_0})}\|\eta\|_{L_2(P_{X_0}\otimes P_{X_0})}.
\end{aligned}$$

We thus focus on bounding $\left\|T^\beta g_{\varepsilon_n}\otimes T^\beta g_{\varepsilon_n} - T^\beta g\otimes T^\beta g\right\|_{L_2(P_{X_0}\otimes P_{X_0})}$. By the triangle inequality,

$$\begin{aligned}
& \|T^\beta g_{\varepsilon_n}\otimes T^\beta g_{\varepsilon_n} - T^\beta g\otimes T^\beta g\|_{L_2(P_{X_0}\otimes P_{X_0})} \\
\leq\ & \|T^\beta g_{\varepsilon_n}\otimes T^\beta g_{\varepsilon_n} - T^\beta g\otimes T^\beta g_{\varepsilon_n}\|_{L_2(P_{X_0}\otimes P_{X_0})} \\
+\ & \|T^\beta g\otimes T^\beta g_{\varepsilon_n} - T^\beta g\otimes T^\beta g\|_{L_2(P_{X_0}\otimes P_{X_0})}.
\end{aligned} \tag{27}$$

Before proceeding, we note that $T^\beta g \in \mathrm{Range}(T^{\alpha+\beta})$ holds because of the assumption $g \in \mathrm{Range}(T^\alpha)$. Therefore by Lemma 10, we have

$$\|T^\beta g_{\varepsilon_n} - T^\beta g\|_{L_2(P_{X_0})} \quad = \quad \|(T+\varepsilon_n I)^{-1}TT^\beta g - T^\beta g\|_{L_2(P_{X_0})} \leq c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta},$$

where $c_{\alpha+\beta}$ is a constant depending only on α, β and g. The first term of (27) can then be upper-bounded as

$$\begin{aligned}
& \|T^\beta g_{\varepsilon_n} \otimes T^\beta g_{\varepsilon_n} - T^\beta g \otimes T^\beta g_{\varepsilon_n}\|_{L_2(P_{X_0}\otimes P_{X_0})} \\
= \;& \|(T^\beta g_{\varepsilon_n} - T^\beta g) \otimes T^\beta g_{\varepsilon_n}\|_{L_2(P_{X_0}\otimes P_{X_0})} \\
= \;& \|T^\beta g_{\varepsilon_n} - T^\beta g\|_{L_2(P_{X_0})} \|T^\beta g_{\varepsilon_n}\|_{L_2(P_{X_0})} \\
\leq \;& \|T^\beta g_{\varepsilon_n} - T^\beta g\|_{L_2(P_{X_0})} (\|T^\beta g_{\varepsilon_n} - T^\beta g\|_{L_2(P_{X_0})} + \|T^\beta g\|_{L_2(P_{X_0})}) \\
\leq \;& c_{\alpha+\beta}^2 \varepsilon_n^{2(\alpha+\beta)} + c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta}\|g\|_{L_2(P_{X_0})} \quad (\because \text{Lemma } 10).
\end{aligned}$$

Similarly, the second term of (27) can be written as

$$\begin{aligned}
\|T^\beta g \otimes T^\beta g_{\varepsilon_n} - T^\beta g \otimes T^\beta g\|_{L_2(P_{X_0}\otimes P_{X_0})} &= \|T^\beta g \otimes (T^\beta g_{\varepsilon_n} - T^\beta g)\|_{L_2(P_{X_0}\otimes P_{X_0})} \\
&= \|T^\beta g\|_{L_2(P_{X_0})} \|T^\beta g_{\varepsilon_n} - T^\beta g\|_{L_2(P_{X_0})} \\
&\leq \|T^\beta g\|_{L_2(P_{X_0})} c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta} \quad (\because \text{Lemma } 10).
\end{aligned}$$

Therefore the first term in (26) is upper-bounded by

$$\left(c_{\alpha+\beta}^2 \varepsilon_n^{2(\alpha+\beta)} + 2\|T^\beta g\|_{L_2(P_{X_0})} c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta}\right) \|\eta\|_{L_2(P_{X_0}\otimes P_{X_0})}. \tag{28}$$

Bound on the second term in (26). First note that

$$\begin{aligned}
\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0 = \cdot] &= \mathbb{E}_{Y_0}\left[\int \mathbb{E}_{\tilde{Y}_0}[\ell(Y_0, \tilde{Y}_0)|\tilde{X}_0 = \tilde{x}]dP_{X_1}(\tilde{x})|X_0 = \cdot\right] \\
&= \int \mathbb{E}_{Y_0,\tilde{Y}_0}\left[\ell(Y_0, \tilde{Y}_0)|X_0 = \cdot, \tilde{X}_0 = \tilde{x}\right] dP_{X_1}(\tilde{x}) \\
&= \int \theta(\cdot, \tilde{x})dP_{X_1}(\tilde{x})
\end{aligned}$$

From this and the equivalence (25), (the half of) the second term in (26) can be written as

$$\left|\left\langle g, \int \theta(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})} - \left\langle g, (T+\varepsilon_n I)^{-1} T \int \theta(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})}\right|. \tag{29}$$

Note that we have $\theta = T^\beta \otimes T^\beta \eta$, which implies that

$$\int \theta(\cdot,\tilde{x})dP_{X_1}(\tilde{x}) = \int (T^\beta \otimes T^\beta h)(\cdot,\tilde{x})dP_{X_1}(\tilde{x}) = T^\beta \int (T_2^\beta \eta)(\cdot,\tilde{x})dP_{X_1}(\tilde{x}),$$

where T_2^β denotes the operator applied to the second argument of a function with two arguments.[7] Thus we have

$$\begin{aligned}
\left\langle g, \int \theta(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})} &= \left\langle g, T^\beta \int (T_2^\beta \eta)(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})} \\
&= \left\langle T^\beta g, \int (T_2^\beta \eta)(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})}.
\end{aligned}$$

Similarly, we have

$$\left\langle g, (T+\varepsilon_n I)^{-1} T \int \theta(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})} = \left\langle (T+\varepsilon_n I)^{-1} T T^\beta g, \int (T_2^\beta \eta)(\cdot,\tilde{x})dP_{X_1}(\tilde{x})\right\rangle_{L_2(P_{X_0})}.$$

[7] Note that $\int (T_2^\beta \eta)(\cdot,\tilde{x})dP_{X_1}(\tilde{x})$ is a function with only one argument, so the expression $T^\beta \int (T_2^\beta \eta)(\cdot,\tilde{x})dP_{X_1}(\tilde{x})$ is justified.

Note that we have $Tg \in \mathrm{Range}(T^{\alpha+\beta})$, which follows from the assumption $g \in \mathrm{Range}(T^{\alpha})$. Thus it follows that

$$\begin{aligned}
(29) \quad &= \left| \left\langle (T+\varepsilon_n I)^{-1} T T^{\beta} g - T^{\beta} g, \int (T_2^{\beta}\eta)(\cdot,\tilde{x}) dP_{X_1}(\tilde{x}) \right\rangle_{L_2(P_{X_0})} \right| \\
&\leq \|(T+\varepsilon_n)^{-1} T T^{\beta} g - T^{\beta} g\|_{L_2(P_{X_0})} \left\| \int (T_2^{\beta}\eta)(\cdot,\tilde{x}) dP_{X_1}(\tilde{x}) \right\|_{L_2(P_{X_0})} \\
&\leq c_{\alpha+\beta}\, \varepsilon_n^{\alpha+\beta} \left\| \int (T_2^{\beta}\eta)(\cdot,\tilde{x}) dP_{X_1}(\tilde{x}) \right\|_{L_2(P_{X_0})} \quad (\because \text{ Lemma } 10), \qquad (30)
\end{aligned}$$

where $c_{\alpha+\beta} > 0$ is a constant depending only on α, β and g.

Resulting approximation error rate. Using (28) and (30) in (26), we now obtain a bound on the approximation error:

$$\begin{aligned}
& \|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}}^2 \\
\leq\ & \left(c_{\alpha+\beta}^2 \varepsilon_n^{2(\alpha+\beta)} + 2\|T^{\beta} g\|_{L_2(P_{X_0})} c_{\alpha+\beta} \varepsilon_n^{\alpha+\beta} \right) \|\eta\|_{L_2(P_{X_0}\otimes P_{X_0})} \\
+\ & 2 c_{\alpha+\beta}\, \varepsilon_n^{\alpha+\beta} \left\| \int (T_2^{\beta}\eta)(\cdot,\tilde{x}) dP_{X_1}(\tilde{x}) \right\|_{L_2(P_{X_0})}.
\end{aligned}$$

Since we will set ε_n to decay to 0, the rate is dominated by the terms involving $\varepsilon_n^{\alpha+\beta}$. Noting that the above bound is for the squared approximation error, we therefore have the rate (24) for the approximation error.

Balancing the estimation and approximation error rates. Let $\varepsilon_n = n^{-b}$ for some constant $b > 0$, which is determined by balancing the two rates (21) and (24). This yields $b = 1/(2-\alpha+\beta)$ for $\alpha \leq 1/2$, and $b = 1/(1+\alpha+\beta)$ for $\alpha \geq 1/2$; equivalently, $b = 1/(1+\beta+\max(1-\alpha,\alpha))$ for $0 < \alpha \leq 1$. The proof completes by substituting the resulting $\varepsilon_n = n^{-b}$ in (21) and (24).

□

C.4 Lemmas

We collect lemmas that are needed for proving the main results.

Lemma 5. *Assume that P_1 is absolutely continuous with respect to P_{X_0}, and let $g := dP_{X_1}/dP_{X_0}$ be the Radon-Nikodym derivative. If $g \in L_2(P_{X_0})$, we have $\mu_{X_1} = Sg$.*

Proof. By the definitions of the kernel mean μ_{X_1} and the Radon-Nikodym derivative g, we have

$$\mu_{X_1} = \int k(\cdot,x) dP_{X_1}(x) = \int k(\cdot,x) g(x) dP_{X_0}(x) = Sg \in \mathscr{H}.$$

□

Lemma 6. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Then for any $f \in L_2(P_{X_0})$ and $\varepsilon > 0$, we have*

$$S^*(\mathcal{C}_{XX} + \varepsilon I)^{-1} S f = (T+\varepsilon I)^{-1} T f.$$

Proof. Let $(e_i)_{i=1}^\infty \subset \mathscr{H}$ and $(\mu_i)_{i=1}^\infty \subset (0,\infty)$ as in Lemma 2. Then by Lemma 2, which is applicable from our assumption on k and P_{X_0}, we have

$$\begin{aligned}
S^*(\mathcal{C}_{XX}+\varepsilon I)^{-1}Sf &= S^*(\mathcal{C}_{XX}+\varepsilon I)^{-1}\sum_{j=1}^\infty \mu_j \left\langle f,[e_j]_\sim \right\rangle_{L_2(P_{X_0})} e_j \\
&= S^*\sum_{i=1}^\infty (\mu_i+\varepsilon)^{-1}\left\langle \mu_i^{1/2}e_i, \sum_{j=1}^\infty \mu_j \left\langle f,[e_j]_\sim \right\rangle_{L_2(P_{X_0})} e_j \right\rangle_{\mathscr{H}} \mu_i^{1/2}e_i \\
&= S^*\sum_{i=1}^\infty (\mu_i+\varepsilon)^{-1}\mu_i \left\langle f,[e_i]_\sim \right\rangle_{L_2(P_{X_0})} e_i \\
&= \sum_{i=1}^\infty (\mu_i+\varepsilon)^{-1}\mu_i \left\langle f,[e_i]_\sim \right\rangle_{L_2(P_{X_0})} [e_i]_\sim \\
&= (T+\varepsilon I)^{-1}Tf.
\end{aligned}$$

□

Lemma 7. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Then for any $f \in L_2(P_{X_0})$ and $\alpha > 0$, we have*

$$ST^\alpha f = \mathcal{C}_{XX}^{1/2+\alpha}S^{1/2}f.$$

Proof. Let $(e_i)_{i=1}^\infty \subset \mathscr{H}$ and $(\mu_i)_{i=1}^\infty \subset (0,\infty)$ as in Lemma 2. Then we have by Lemma 2 and Definition 2

$$\begin{aligned}
\mathcal{C}_{XX}^{1/2+\alpha}S^{1/2}f &= \mathcal{C}_{XX}^{1/2+\alpha}\sum_{i=1}^\infty \mu_i^{1/2}\left\langle [e_i]_\sim, f\right\rangle_{L_2(P_{X_0})} e_i \\
&= \sum_{\ell=1}^\infty \mu_\ell^{1/2+\alpha}\left\langle \mu_\ell^{1/2}e_\ell, \sum_{i=1}^\infty \mu_i^{1/2}\left\langle [e_i]_\sim, f\right\rangle_{L_2(P_{X_0})} e_i\right\rangle_{\mathscr{H}} \mu_\ell^{1/2}e_\ell \\
&= \sum_{\ell=1}^\infty \mu_\ell^{1/2+\alpha}\left\langle [e_\ell]_\sim, f\right\rangle_{L_2(P_{X_0})} \mu_\ell^{1/2}e_\ell \\
&= \sum_{\ell=1}^\infty \mu_\ell^{\alpha}\left\langle [e_\ell]_\sim, f\right\rangle_{L_2(P_{X_0})} \mu_\ell e_\ell \\
&= \sum_{\ell=1}^\infty \mu_\ell^{\alpha}\left\langle [e_\ell]_\sim, f\right\rangle_{L_2(P_{X_0})} S[e_\ell]_\sim \\
&= S\sum_{\ell=1}^\infty \mu_\ell^{\alpha}\left\langle [e_\ell]_\sim, f\right\rangle_{L_2(P_{X_0})} [e_\ell]_\sim \\
&= ST^\alpha f.
\end{aligned}$$

□

Lemma 8. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Assume that the Radon-Nikodym derivative $g := dP_{X_1}/dP_{X_0}$ satisfies $g \in \mathrm{Range}(T^\alpha)$ for a constant $\alpha > 0$. Then for any $\varepsilon > 0$, we have*

$$\left\|(\mathcal{C}_{XX}+\varepsilon I)^{-1}\mu_{X_1}\right\|_{\mathscr{H}} \le \begin{cases} c_\alpha \varepsilon^{-1/2+\alpha}, & (\text{if } \alpha \le 1/2) \\ c_\alpha \left\|\mathcal{C}_{XX}^{\alpha-1/2}\right\|, & (\text{if } \alpha > 1/2), \end{cases}$$

where $c_\alpha := \|S^{1/2}h\|_{\mathscr{H}}$ is a constant with $h \in L_2(P_{X_0})$ being a function such that $g = T^\alpha h$ (which exists from the assumption $g \in \mathrm{Range}(T^\alpha)$).

Proof. As in the assertion, write $g = T^{\alpha} h$ for $h \in L_2(P_{X_0})$. By Lemmas 5 and 7, we can then write μ_{X_1} as

$$\mu_{X_1} = Sg = ST^{\alpha}h = \mathcal{C}_{XX}^{1/2+\alpha} S^{1/2} h.$$

Therefore we have

$$\begin{aligned} \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1} \mu_{X_1}\right\|_{\mathscr{H}} &= \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}^{1/2+\alpha} S^{1/2} h\right\|_{\mathscr{H}} \\ &\leq \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}^{1/2+\alpha}\right\| \left\|S^{1/2} h\right\|_{\mathscr{H}}. \end{aligned}$$

Below we focus on bounding the first term in the above bound. If $\alpha \leq 1/2$,

$$\begin{aligned} \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}^{1/2+\alpha}\right\| &\leq \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1/2+\alpha}\right\| \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1/2-\alpha} \mathcal{C}_{XX}^{1/2+\alpha}\right\| \\ &\leq \varepsilon^{-1/2+\alpha}. \end{aligned}$$

On the other hand, if $\alpha > 1/2$,

$$\begin{aligned} \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}^{1/2+\alpha}\right\| &\leq \left\|(\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}\right\| \left\|\mathcal{C}_{XX}^{\alpha-1/2}\right\| \\ &\leq \left\|\mathcal{C}_{XX}^{\alpha-1/2}\right\|. \end{aligned}$$

□

Lemma 9. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Assume that the mapping $S^* : \mathscr{H} \to L_2(P_{X_0})$ has a dense image in $L_2(P_{X_0})$. Then any $g \in L_2(P_{X_0})$, we have*

$$\lim_{\varepsilon \to 0} \|(T + \varepsilon I)^{-1} Tg - g\|_{L_2(P_{X_0})} = 0.$$

Proof. Let $(e_i)_{i=1}^{\infty} \subset \mathscr{H}$ and $(\mu_i)_{i=1}^{\infty} \subset (0, \infty)$ be as in Lemma 2. By Lemma 3 and our assumption on S^*, $([e_i]_\sim)_{i=1}^{\infty}$ is an ONB of $L_2(P_{X_0})$, which implies that g can be expanded using $([e_i]_\sim)_{i=1}^{\infty}$. From this and Lemma 2, we then have

$$\begin{aligned} (T + \varepsilon I)^{-1} Tg - g &= \sum_{i=1}^{\infty} (\mu_i + \varepsilon)^{-1} \mu_i \left\langle g, [e_i]_\sim \right\rangle_{L_2(P_{X_0})} [e_i]_\sim - \sum_{i=1}^{\infty} \left\langle g, [e_i]_\sim \right\rangle_{L_2(P_{X_0})} [e_i]_\sim \\ &= \sum_{i=1}^{\infty} -\varepsilon (\mu_i + \varepsilon)^{-1} \left\langle g, [e_i]_\sim \right\rangle_{L_2(P_{X_0})} [e_i]_\sim. \end{aligned}$$

Thus, by Parseval's identity,

$$\|(T + \varepsilon I)^{-1} Tg - g\|_{L_2(P_{X_0})}^2 = \sum_{i=1}^{\infty} \left|\varepsilon (\mu_i + \varepsilon)^{-1}\right|^2 | \left\langle g, [e_i]_\sim \right\rangle_{L_2(P_{X_0})} |^2.$$

Note that $\left|\varepsilon (\mu_i + \varepsilon)^{-1}\right|^2 \leq 1$ for all i, that $\sum_{i=1}^{\infty} | \left\langle g, [e_i]_\sim \right\rangle_{L_2(P_{X_0})} |^2 = \|g\|_{L_2(P_{X_0})}^2 < \infty$, and that $\lim_{\varepsilon \to 0} \left|\varepsilon (\mu_i + \varepsilon)^{-1}\right|^2 = 0$ (which follows from $\mu_i > 0$ for all $i = 1, 2, \dots$). These facts enable the use of the dominated convergence theorem, from which we have

$$\begin{aligned} \lim_{\varepsilon \to 0} \|(T + \varepsilon I)^{-1} Tg - g\|_{L_2(P_{X_0})}^2 &= \lim_{\varepsilon \to 0} \sum_{i=1}^{\infty} \left|\varepsilon (\mu_i + \varepsilon)^{-1}\right|^2 | \left\langle g, [e_i]_\sim \right\rangle_{L_2(P_{X_0})} |^2 \\ &= \sum_{i=1}^{\infty} \lim_{\varepsilon \to 0} \left|\varepsilon (\mu_i + \varepsilon)^{-1}\right|^2 | \left\langle g, [e_i]_\sim \right\rangle_{L_2(P_{X_0})} |^2 = 0. \end{aligned}$$

□

Lemma 10. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Let $g \in L_2(P_{X_0})$, and assume that $g \in \mathrm{Range}(T^{\alpha})$ for a constant $0 < \alpha \leq 1$. Then, for all $\varepsilon > 0$, we have*

$$\|(T + \varepsilon I)^{-1} Tg - g\|_{L_2(P_{X_0})} \leq c_{\alpha} \varepsilon^{\alpha},$$

where $c_{\alpha} := \|h\|_{L_2(P_{X_0})}$ with $h \in L_2(P_{X_0})$ being such that $g = T^{\alpha} h$.

Proof. Let $(e_i)_{i=1}^{\infty} \subset \mathscr{H}$ and $(\mu_i)_{i=1}^{\infty} \subset (0, \infty)$ be as in Lemma 2. As in the assertion, from the assumption $g \in \mathrm{Range}(T^{\alpha})$ there exists $h \in L_2(P_{X_0})$ such that $g = T^{\alpha}h$. Therefore g can be written as

$$g = T^{\alpha}h = \sum_{i=1}^{\infty} \mu_i^{\alpha} b_i [e_i]_{\sim}, \tag{31}$$

where the convergence is in $L_2(P_{X_0})$, and $b_i := \langle h, [e_i]_{\sim} \rangle_{L_2(P_{X_0})}$. It then follows that

$$\begin{aligned}(T + \varepsilon I)^{-1}Tg - g &= \sum_{i=1}^{\infty} (\mu_i + \varepsilon)^{-1} \mu_i \mu_i^{\alpha} b_i [e_i]_{\sim} - \sum_{i=1}^{\infty} \mu_i^{\alpha} b_i [e_i]_{\sim} \\ &= \sum_{i=1}^{\infty} -\varepsilon (\mu_i + \varepsilon)^{-1} \mu_i^{\alpha} b_i [e_i]_{\sim}.\end{aligned}$$

Therefore, by Parseval's identity, we have

$$\|(T + \varepsilon I)^{-1}Tg - g\|_{L_2(P_{X_0})}^2 = \sum_{i=1}^{\infty} \varepsilon^2 (\mu_i + \varepsilon)^{-2} \mu_i^{2\alpha} b_i^2.$$

The right side of the above equation can be upper-bounded as

$$\begin{aligned}\varepsilon^2 (\mu_i + \varepsilon)^{-2} \mu_i^{2\alpha} b_i^2 &= \varepsilon^2 (\mu_i + \varepsilon)^{-2+2\alpha} (\mu_i + \varepsilon)^{-2\alpha} \mu_i^{2\alpha} b_i^2 \\ &\leq \varepsilon^2 (\mu_i + \varepsilon)^{-2+2\alpha} b_i^2 \\ &= \varepsilon^{2\alpha} \varepsilon^{2-2\alpha} (\mu_i + \varepsilon)^{-2+2\alpha} b_i^2 \\ &\leq \varepsilon^{2\alpha} b_i^2,\end{aligned}$$

where the above two inequalities follow from $\varepsilon > 0$ and $\mu_i > 0$, and the last inequality uses $\alpha \leq 1$. Thus, we have

$$\begin{aligned}\|(T + \varepsilon I)^{-1}Tg - g\|_{L_2(P_{X_0})}^2 \leq \varepsilon^{2\alpha} \sum_{i=1}^{\infty} b_i^2 &= \varepsilon^{2\alpha} \sum_{i=1}^{\infty} \left(\langle h, [e_i]_{\sim} \rangle_{L_2(P_{X_0})} \right)^2, \\ &\leq \varepsilon^{2\alpha} \|h\|_{L_2(P_{X_0})}^2,\end{aligned}$$

where the last inequality follows from $([e_i]_{\sim})_{i=1}^{\infty}$ is an ONS in $L_2(P_{X_0})$ and the Parseval's identity.

□

Remark 1. *Different from Lemma 9, Lemma 10 does not require the condition that $S*$ has a dense image in $L_2(P_{X_0})$. In Lemma 9, this condition is required to guarantee that $([e_i]_{\sim})_{i=1}^{\infty}$ is an ONB in $L_2(P_{X_0})$, so that g can be expanded by this ONB. On the other hand, in Lemma 10, g can be written as (31), thanks to the assumption $g \in \mathrm{Range}(T^{\alpha})$. Therefore Lemma 10 does not need the condition on S^*.*

The following is the key lemma, based on which we show the consistency and convergence rates of our estimator.

Lemma 11. *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and P_{X_0} be a probability measure on $\mathcal{X}$ such that Assumption 3 is satisfied. Assume P_{X_1} is absolutely continuous with respect to P_{X_0} with the Radon-Nikodym derivative $g = dP_{X_1}/dP_{X_0}$ such that $g \in L_2(P_{X_0})$. Define a function $\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by $\theta(x, x') := \mathbb{E}_{Y_0, Y_0'}[\ell(Y_0, Y_0')|X_0 = x, X_0' = x']$. Then for any $\varepsilon_n > 0$, we have*

$$\begin{aligned}&\|\mathcal{C}_{YX}(\mathcal{C}_{XX} + \varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}}^2 \\ = \;& \langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(P_{X_0} \otimes P_{X_0})} \\ &-2 \left\langle g, (T + \varepsilon_n I)^{-1} T \, \mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0 = \cdot] \right\rangle_{L_2(P_{X_0})} \\ &+ \int\int \theta(x, \tilde{x}) dP_{X_1}(x) dP_{X_1}(\tilde{x})\end{aligned}$$

where $g_{\varepsilon_n} := (T + \varepsilon_n I)^{-1}Tg$. In the second term of the right hand side, the inner-product is well defined, since we have $(T + \varepsilon_n I)^{-1}T \, \mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0 = \cdot] \in L_2(P_{X_0})$.

Proof. First note that

$$\begin{aligned} & \|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}}^2 & (32)\\ = & \|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathcal{F}}^2 - 2\left\langle \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}, \mu_{Y\langle 0|1\rangle}\right\rangle_{\mathcal{F}} + \|\mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}}^2. \end{aligned}$$

As shown in the proof of Theorem 8 in [12], the third term in (32) can be written as

$$\|\mu_{Y\langle 0|1\rangle}\|_{\mathcal{F}}^2 = \int\int \theta(x,\tilde{x})dP_{X_1}(x)dP_{X_1}(\tilde{x}). \tag{33}$$

We thus derive the expressions for the first two terms in (32) below.

The first term in (32): Let $f \in \mathscr{H}$ be arbitrary, and let $(\tilde{X}_0, \tilde{Y}_0)$ denote an independent copy of (X_0, Y_0). By the definitions of $\mathcal{C}_{YX}$ and θ, we have

$$\begin{aligned} \|\mathcal{C}_{YX}f\|_{\mathcal{F}}^2 &= \langle \mathcal{C}_{YX}f, \mathcal{C}_{YX}f\rangle_{\mathcal{F}} \\ &= \mathbb{E}_{X_0,Y_0}[f(X_0)(\mathcal{C}_{YX}f)(Y_0))] \\ &= \mathbb{E}_{X_0,Y_0}[f(X_0)\mathbb{E}_{\tilde{X}_0,\tilde{Y}_0}[\ell(Y_0,\tilde{Y}_0)f(\tilde{X}_0)]] \\ &= \mathbb{E}_{X_0,\tilde{X}_0}[f(X_0)f(\tilde{X}_0)\mathbb{E}_{Y_0,\tilde{Y}_0}[\ell(Y_0,\tilde{Y}_0)|X_0,\tilde{X}_0]] \\ &= \mathbb{E}_{X_0,\tilde{X}_0}[f(X_0)f(\tilde{X}_0)\theta(X_0,\tilde{X}_0)] \qquad (34) \end{aligned}$$

Now define $f := (\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1} \in \mathscr{H}$. With this choice of f, the quantity $\|\mathcal{C}_{YX}f\|_{\mathcal{F}}^2$ is equal to the first term in (32). From (34), it follows that

$$\begin{aligned} \|\mathcal{C}_{YX}f\|_{\mathcal{F}}^2 &= \mathbb{E}_{X_0,\tilde{X}_0}[f(X_0)f(\tilde{X}_0)\theta(X_0,\tilde{X}_0)] \\ &= \int\int f(x)f(\tilde{x})\theta(x,\tilde{x})dP_{X_0}(x)dP_{X_0}(\tilde{x}) \\ &= \langle S^*f\otimes S^*f, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})} \\ &= \left\langle S^*(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\otimes S^*(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}, \theta\right\rangle_{L_2(P_{X_0}\otimes P_{X_0})} \\ &= \left\langle S^*(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}Sg\otimes S^*(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}Sg, \theta\right\rangle_{L_2(P_{X_0}\otimes P_{X_0})} \quad (\because \text{Lemma } 5) \\ &= \left\langle (T+\varepsilon_n I)^{-1}Tg\otimes (T+\varepsilon_n I)^{-1}Tg, \theta\right\rangle_{L_2(P_{X_0}\otimes P_{X_0})} \quad (\because \text{Lemma } 6) \\ &= \langle g_{\varepsilon_n}\otimes g_{\varepsilon_n}, \theta\rangle_{L_2(P_{X_0}\otimes P_{X_0})}, \qquad (35) \end{aligned}$$

where $g_{\varepsilon_n} := (T+\varepsilon_n I)^{-1}Tg$.

The second term in (32) : First we have

$$\begin{aligned} \mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot] &= \mathbb{E}_{Y_0}\left[\int \mathbb{E}_{\tilde{Y}_0}[\ell(Y_0,\tilde{Y}_0)|\tilde{X}_0=\tilde{x}]dP_{X_1}(\tilde{x})|X_0=\cdot\right] \\ &= \int \mathbb{E}_{Y_0,\tilde{Y}_0}\left[\ell(Y_0,\tilde{Y}_0)|X_0=\cdot,\tilde{X}_0=\tilde{x}\right]dP_{X_1}(\tilde{x}) \\ &= \int \theta(\cdot,\tilde{x})dP_{X_1}(\tilde{x}) \qquad (36) \end{aligned}$$

where $(\tilde{X}_0, \tilde{Y}_0)$ is an independent copy of (X_0, Y_0). For the first expression in Eq. (36), we can also show that $\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot] \in L_2(P_{X_0})$ as follows.

$$\begin{aligned} & \int \left(\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=x]\right)^2 dP_{X_0}(x) \\ = & \int\left(\int \theta(x,\tilde{x})dP_{X_1}(\tilde{x})\right)^2 dP_{X_0}(x) \\ = & \int\left(\int \theta(x,\tilde{x})g(\tilde{x})dP_{X_0}(\tilde{x})\right)^2 dP_{X_0}(x) \\ \leq & \int\int \theta^2(x,\tilde{x})dP_{X_0}(\tilde{x})\int g^2(\tilde{x})dP_{X_0}(\tilde{x})dP_{X_0}(x) \quad (\because \text{Cauchy}-\text{Schwartz}) \\ = & \|g\|_{L_2(P_{X_0})}\int\int \theta^2(x,\tilde{x})dP_{X_0}(\tilde{x})dP_{X_0}(x) < \ +\infty. \end{aligned}$$

We also have

$$\begin{aligned}\mathcal{C}_{XY}\mu_{Y\langle 0|1\rangle} &= \mathbb{E}_{X_0,Y_0}[k(\cdot,X_0)\mu_{Y\langle 0|1\rangle}(Y_0)] \\ &= \mathbb{E}_{X_0}\left[k(\cdot,X_0)\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0]\right] \\ &= S\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot]. \qquad (37)\end{aligned}$$

Note that $S\mathbb{E}[\mu_{Y\langle 0|1\rangle}(Y)|X=\cdot]$ is well defined, since $\mathbb{E}[\mu_{Y\langle 0|1\rangle}(Y)|X=\cdot] \in L_2(P_{X_0})$. Now for the second term in (32), we have

$$\begin{aligned}&\left\langle \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}, \mu_{Y\langle 0|1\rangle}\right\rangle_{\mathcal{F}} \\ =\ &\left\langle \mu_{X_1}, (\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mathcal{C}_{XY}\mu_{Y\langle 0|1\rangle}\right\rangle_{\mathscr{H}} \\ =\ &\left\langle Sg, (\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mathcal{C}_{XY}\mu_{Y\langle 0|1\rangle}\right\rangle_{\mathscr{H}} \quad (\because \text{ Lemma } 5) \\ =\ &\left\langle Sg, (\mathcal{C}_{XX}+\varepsilon_n I)^{-1}S\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot]\right\rangle_{\mathscr{H}} \quad (\because (37)) \\ =\ &\left\langle g, S^*(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}S\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot]\right\rangle_{L_2(P_{X_0})} \\ =\ &\left\langle g, (T+\varepsilon_n I)^{-1}T\,\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0=\cdot]\right\rangle_{L_2(P_{X_0})} \quad (\because \text{Lemma } 6).\end{aligned}$$

□