Journal of Machine Learning Research 22 (2021) 1-71 Submitted 2/20; Published 6/21

Counterfactual Mean Embeddings

Krikamol Muandet[*] KRIKAMOL@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Motonobu Kanagawa[†] MOTONOBU.KANAGAWA@EURECOM.FR
Data Science Department, EURECOM
Sophia Antipolis, France

Sorawit Saengkyongam SS@MATH.KU.DK
University of Copenhagen
Copenhagen, Denmark

Sanparith Marukatat SANPARITH.MARUKATAT@NECTEC.OR.TH
National Electronics and Computer Technology Center
National Science and Technology Development Agency
Pathumthani, Thailand

Editor: David Sontag

Abstract

Counterfactual inference has become a ubiquitous tool in online advertisement, recommendation systems, medical diagnosis, and econometrics. Accurate modelling of outcome distributions associated with different interventions—known as counterfactual distributions—is crucial for the success of these applications. In this work, we propose to model counterfactual distributions using a novel Hilbert space representation called counterfactual mean embedding (CME). The CME embeds the associated counterfactual distribution into a reproducing kernel Hilbert space (RKHS) endowed with a positive definite kernel, which allows us to perform causal inference over the entire landscape of the counterfactual distribution. Based on this representation, we propose a distributional treatment effect (DTE) which can quantify the causal effect over entire outcome distributions. Our approach is nonparametric as the CME can be estimated under the unconfoundedness assumption from observational data without requiring any parametric assumption about the underlying distributions. We also establish a rate of convergence of the proposed estimator which depends on the smoothness of the conditional mean and the Radon-Nikodym derivative of the underlying marginal distributions. Furthermore, our framework allows for more complex outcomes such as images, sequences, and graphs. Our experimental results on synthetic data and off-policy evaluation tasks demonstrate the advantages of the proposed estimator.

Keywords: counterfactual inference, kernel mean embedding, potential outcome framework, reproducing kernel Hilbert space, causality

∗. A part of this work was done when KM was affiliated with the Department of Mathematics, Mahidol University, Thailand.

†. A part of this work was done when MK was affiliated with the Institute of Statistical Mathematics, Japan, and with the University of Tübingen and Max Planck Institute for Intelligent Systems, Germany.

1. Introduction

To make a rational decision, a decision maker must be able to anticipate the effects of a decision to the outcomes of interest, before committing to that decision. For instance, before building a certain facility in a city, *e.g.*, a dam, policymakers and citizens must seek to understand its environmental effects. In medicine, a doctor has some prior knowledge about the effects a certain drug will have on a patient's health, before actually prescribing it. In business, a company needs to understand the effects of a certain strategy of advertisement to its revenue. One approach to addressing these questions is *counterfactual inference*.

Counterfactual inference we consider in this work consists of the following three main ingredients. Suppose that there exists a hypothetical subject (*e.g.*, a patient in medical treatment), and let X be *covariates* representing the features of the subject (*e.g.*, age, weight, medical record, etc.), T be a *treatment indicator* representing the treatment assigned to the subject (a drug of interest or a placebo), and Y be the *observed outcome* representing the post-treatment quantity of interest (*e.g.*, whether the patient is recovered or not). Given certain realizations of these variables $\{(\mathbf{x}_i, t_i, \mathbf{y}_i)\}_{i=1}^n$, in which each index i represents the identity of a subject, an analyst wishes to know how the treatment affects the outcome.

This problem is called counterfactual since for each subject i, we only observe the outcome $\mathbf{y}_i$ resulting from the assigned treatment t_i and can never observe the outcome (say $\mathbf{y}_i'$) that would have been realized under an alternative treatment $t_i' \neq t_i$. For example, if a patient receives an active treatment (*e.g.*, a drug of interest), we can never observe the outcome from the same patient under a control treatment (*e.g.*, a placebo). This is known as the fundamental problem of causal inference (Holland, 1986) and also as *bandit feedback* in the bandit literature (Dudík et al., 2011). One way to partially address this issue is a randomized experiment (Fisher, 1935), in which treatments are randomly assigned to subjects. Although considered a gold standard, in practice randomization can be too expensive, time-consuming, or unethical. In most cases, therefore, analysis about treatment effects needs to be done on the basis of observational data $\{(\mathbf{x}_i, t_i, \mathbf{y}_i)\}_{i=1}^n$ in which the treatment assignment t_i may depend on covariates $\mathbf{x}_i$ and possibly on some hidden confounders; this setting is commonly known as *observational studies* (Rosenbaum, 2002; Rubin, 2005).

A fundamental framework for observational studies is the *potential outcome framework* (Neyman, 1923; Rubin, 1974). It provides a clear notation for *potential outcomes*, *i.e.*, the outcomes that would have been observed under different treatments, and elucidates the conditions required for making a valid inference about treatment effects; see Section 3.1. The framework has been studied extensively in statistics, and has a wide range of applications in biomedical and social sciences; see, *e.g.*, Imbens and Rubin (2015). Moreover, important applications of machine learning such as off-policy evaluation for online advertisement and recommendation systems can be reformulated under this framework (Schnabel et al., 2016; Kallus and Zhou, 2018). We argue, however, that there exist the following challenges:

Average treatment effects. Many of existing works focus on estimating the *average treatment effect* (ATE), which is the difference between the means of the outcome distributions; see Section 3.1 for details. However, the ATE does not inform changes in higher-order moments, even when they exist. For instance, if a treatment of interest has an effect only in the *variance* of the distribution of outcomes, then the analysis of average treatment effects cannot capture such effects. Suppose that the treatment is whether to provide a certain drug,

and the outcome is the blood pressure of a patient; just analyzing the average treatment effects may lead to an incorrect conclusion, if the drug increases/decreases the blood pressure of a patient whose blood pressure was already high/low. This highlights the importance of analyzing the outcome distribution as a whole.

In this work, we focus on the *distributional treatment effect* (DTE), which involves the entire outcome distributions. This scenario often arises in several real-world socioeconomic applications; see, *e.g.*, Rothe (2010); Chernozhukov et al. (2013).

Parametric models. Many of the classical approaches in causal inference make parametric assumptions about relationships between covariates X, treatment assignment T, and observed (or potential) outcomes. However, if the imposed parametric assumption is incorrect, *i.e.*, model misspecification, then the conclusion about treatment effects can be wrong or misleading. To overcome this limitation, there is a recent surge in applying nonparametric machine learning models to causal inference problems, e.g., Shalit et al. (2016) and Alaa and van der Schaar (2017) among others. This paper also contributes to this endeavour.

Overparameterized models. Deep learning has become the first choice in many applied fields due to its excellent empirical performance, and thus has also been applied to counterfactual inference, *e.g.*, Johansson et al. (2016); Hartford et al. (2017). Unfortunately, such approaches based on deep learning lack theoretical guarantees, because arguably deep learning itself lacks an established theory as a learning method (at least until now). This is problematic when consequential decisions are based on the analysis of treatment effects (*e.g.*, political decisions and medical treatments). Having better theoretical grounding, kernel methods have recently become popular tools for causal inference (Alaa and van der Schaar, 2017; Singh et al., 2019; Muandet et al., 2020a,b).

Multivariate and structured outputs. Existing works often deal with outcomes that are discrete or real-valued. However, depending on the application, outcome variables may be multivariate (possibly high-dimensional) or structured, such as images and graphs. For example, in medical data analysis, outcomes may be fMRI data taken from a subject after receiving a certain treatment. Thus, it is not straightforward to apply existing approaches.

In this work, we propose a novel approach to counterfactual inference that addresses the above challenges, which we term *counterfactual mean embedding* (CME). Our approach is built on kernel mean embedding (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2017), a framework for representing probability distributions as elements in a reproducing kernel Hilbert space (RKHS), so that each element representing a distribution maintains all of its information (cf. Section 2.2 and 2.3). We define an element representing a counterfactual distribution, for which we propose a nonparametric estimator. Notable advantages of the proposed approach are summarized as follows:

1. The proposed estimator can be computed based only on linear algebraic operations involving kernel matrices. Being a kernel method, it can be applied to not only standard domains (such as the Euclidean space), but also more complex and structured covariates and/or outcomes such as images, sequences, and graphs, by using off-the-shelf kernels designed for such data (Gärtner, 2003); this widens possible applications of counterfactual inference in general (cf. Section 3.4). Thus our work offers more flexibility than the existing approaches by Rothe (2010) and Chernozhukov et al. (2013),

who focused on estimating the cumulative distribution functions of counterfactual distributions by assuming real-valued outcomes.

2. The proposed estimator can be used for computing a distance between the counterfactual and controlled distributions, thereby providing a way of quantifying the effect of a treatment to the distribution of outcomes; we define this distance as the maximum mean discrepancy (MMD) (Borgwardt et al., 2006; Gretton et al., 2012) between the counterfactual and controlled distributions. It also provides a way to sample points from a counterfactual distribution based on kernel herding (Chen et al., 2010), a kernel-based deterministic sampling method (cf. Section 3.5).

3. The proposed estimator is nonparametric, and has theoretical guarantees. Specifically, we prove the consistency of the proposed estimator under a very mild condition (cf. Theorem 8), and derive its convergence rates under certain regularity assumptions involving kernels and underlying distributions (cf. Theorem 13). Both results hold without assuming any parametric assumption.

The rest of the paper is organized as follows. After summarizing related work in Section 1.1, we review in Section 2 the potential outcome framework as well as kernel mean embedding of distributions. Section 3 introduces counterfactual learning and then provides a generalization of Hilbert space embedding to counterfactual distributions. This section also presents how we can quantify and estimate distributional treatment effects (DTEs) with our approach. We subsequently provide the detailed convergence analysis in Section 4, followed by examples of the important applications in Section 5 (sampling and testing) and Section 6 (off-policy evaluation). Finally, we demonstrate the effectiveness of the proposed estimator on simulated data as well as real-world policy evaluation tasks in Section 7.

1.1 Related Work

We summarize below related works on counterfactual inference.

Treatment effect estimation. Estimating treatment effects is one of the most fundamental tasks in counterfactual inference (Rubin, 1974; Shalit et al., 2017). This task is hindered by the fact that one cannot observe all potential outcomes at the same time for each subject. Moreover, the data is usually biased by a non-randomized treatment assignment. Modern approaches attempt to resolve these problems by usingo state-of-the-art ML algorithms. For example, Hill (2011) develops a nonparametric method for estimating the ITE based on Bayesian additive regression tree (BART). Athey and Imbens (2016) and Wager and Athey (2018) adapt tree-based methods to treatment effect estimation. Shalit et al. (2016) and Johansson et al. (2016) formulate the problem as a domain adaptation problem and propose to balance the covatiates using representation learning. Hartford et al. (2017) develop a two-step regression method based on deep neural networks for instrumental variable regression. Adversarial training of neural networks for causal inference have also been considered in Yoon et al. (2018), for example.

Off-policy evaluation and learning from observational data. In many circumstances, evaluating and learning a policy by interacting directly with an environment may not be possible due to practical constraints (*e.g.*, monetary costs, safety and ethics). As

a result, several works have attempted to leverage historical data collected using a logging policy in off-policy evaluation and learning, *e.g.*, Langford et al. (2008); Atan et al. (2018). Most methods rely on importance weighting (Langford et al., 2008; Bottou et al., 2013; Swaminathan and Joachims, 2015). Dudík et al. (2011) uses a doubly robust estimator to reduce the variance of off-policy evaluation. Swaminathan and Joachims (2015) presents a framework for policy learning called counterfactual risk minimization (CRM) based on empirical variance regularization. In this work, we also demonstrate the application of our estimator in off-policy evaluation.

Causal inference with kernel mean embeddings. Hilbert space embedding of distributions has been applied extensively in causal inference. For instance, in causal discovery, Fukumizu et al. (2008); Zhang et al. (2011); Doran et al. (2014) develop powerful kernel-based tests of conditional independence which allow for the recovery of causal graphs up to the Markov equivalence class. See Muandet et al. (2017, Section 4.8) for a review of many other applications. In treatment effect estimation, kernel methods have become a popular approach to covariate balancing between treatment and control groups (Shalit et al., 2016; Johansson et al., 2016; Wong and Chan, 2017; Kallus, 2017). Our work, on the contrary, focuses on characterizing the representation of counterfactual distribution of outcomes using the kernel mean embedding and provides nonparametric inference tools.

2. Preliminaries

The counterfactual mean embedding relies on the potential outcome framework as well as the concepts of kernels, reproducing kernel Hilbert spaces (RKHSs), and kernel mean embedding of distributions. We review these concepts in this section.

2.1 Kernels and Reproducing Kernel Hilbert Spaces (RKHSs)

We first review kernels and RKHSs, details of which can be found in, *e.g.*, Schölkopf and Smola (2002), Berlinet and Thomas-Agnan (2004), and Smola et al. (2007).

Let $\mathcal{X}$ be a nonempty set. Let $\mathscr{H}$ be a Hilbert space consisting of functions on $\mathcal{X}$ with $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ and $\|\cdot\|_{\mathscr{H}}$ being its inner-product and norm, respectively. The Hilbert space $\mathscr{H}$ is called a *reproducing kernel Hilbert space* (RKHS), if there exists a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, called the *reproducing kernel* of $\mathscr{H}$, satisfying the following properties:

1. For all $\mathbf{x} \in \mathcal{X}$, we have $k(\cdot, \mathbf{x}) \in \mathscr{H}$. Here $k(\cdot, \mathbf{x})$ is the function of the first argument with $\mathbf{x}$ being fixed, such that $\mathbf{x}' \mapsto k(\mathbf{x}', \mathbf{x})$.

2. For all $f \in \mathscr{H}$ and $\mathbf{x} \in \mathcal{X}$, we have $f(\mathbf{x}) = \langle k(\cdot, \mathbf{x}), f \rangle_{\mathscr{H}}$. This is called the *reproducing property* of $\mathscr{H}$ (or of k).

It is known that the linear span of functions $k(\cdot, \mathbf{x})$, denoted by $\mathrm{span}(k(\cdot, \mathbf{x}) \mid \mathbf{x} \in \mathcal{X})$, is dense in $\mathscr{H}$, *i.e.*,

$$\mathscr{H} = \overline{\mathrm{span}(k(\cdot, \mathbf{x}) \mid \mathbf{x} \in \mathcal{X})},$$

where the closure on the right hand side is taken with respect to the norm of $\mathscr{H}$. In other words, any $f \in \mathscr{H}$ can be written as $f = \sum_{i=1}^{\infty} \alpha_i k(\cdot, \mathbf{x}_i)$ for some $(\alpha_i)_{i=1}^{\infty} \subset \mathbb{R}$ and $(\mathbf{x}_i)_{i=1}^{\infty} \subset \mathcal{X}$ such that $\|\sum_{i=1}^{\infty} \alpha_i k(\cdot, \mathbf{x}_i)\|_{\mathscr{H}}^2 = \sum_{i,j=1}^{\infty} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) < \infty$.

Any RKHS is uniquely associated with its reproducing kernel k, which is *positive definite*: a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite, if for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, we have $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$. On the other hand, for *any* positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists an RKHS $\mathscr{H}$ for which k is the reproducing kernel (Aronszajn, 1950). Therefore, by defining a positive definite kernel, one always implicitly defines its RKHS.

As indicated from the definition of positive definiteness, kernels can be defined on *any* nonempty set $\mathcal{X}$. Therefore, they have been defined not only for the real vector space $\mathbb{R}^d$, but also for non-standard domains such as those of images and graphs. Popular kernels on $\mathcal{X} \subset \mathbb{R}^d$ include linear kernels $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$, polynomial kernels $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + c)^p, c > 0, p \in \mathbb{N}_+$, Gaussian kernels $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\sigma^2), \sigma > 0$, and Laplace (or more generally Matérn) kernels $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2 / 2\sigma^2), \sigma > 0$. More examples of positive definite kernels can be found in Genton (2002) and Hofmann et al. (2008).

2.2 Kernel Mean Embedding of Distributions

In this work, we use kernels and RKHSs to represent, compare, and estimate *probability distributions*. This is enabled by the approach known as *kernel mean embedding* of distributions (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2017), which we review here. In what follows, we assume that $\mathcal{X}$ is a measurable space with some sigma algebra $\mathcal{B}_\mathcal{X}$.

Definition 1 (Kernel mean embedding (KME)) *Let $\mathscr{P}$ be the set of all probability measures on a measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a measurable positive definite kernel with associated RKHS $\mathscr{H}$, such that $\sup_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}) < \infty$. Then, the kernel mean embedding (KME) of $\mathbb{P} \in \mathscr{P}$ is defined as the Bochner integral*[1] *of $k(\cdot, \mathbf{x})$ with respect to $\mathbb{P}$:*

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

The element $\mu_\mathbb{P}$ may be alternatively called the kernel mean of $\mathbb{P}$. For a random variable $X \sim \mathbb{P}$, the kernel mean may also be written as μ_X.

The kernel mean $\mu_\mathbb{P}$ serves as a representation of $\mathbb{P} \in \mathscr{P}$ in the RKHS $\mathscr{H}$. This is justified if $\mathscr{H}$ is *characteristic* (Fukumizu et al., 2004): the RKHS $\mathscr{H}$ (and the associated kernel k) is defined to be characteristic, if the mapping $\mu : \mathscr{P} \to \mathscr{H}$ in (1) is injective. In other words, $\mathscr{H}$ is characteristic, if for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$, we have $\mu_\mathbb{P} = \mu_\mathbb{Q}$ *if and only if* $\mathbb{P} = \mathbb{Q}$. That is, $\mu_\mathbb{P}$ is uniquely associated with $\mathbb{P} \in \mathscr{P}$, and thus $\mu_\mathbb{P}$ becomes a unique representation of $\mathbb{P}$ in $\mathscr{H}$, maintaining all information about $\mathbb{P}$. Examples of characteristic kernels on $\mathcal{X} = \mathbb{R}^d$ include Gaussian, Matérn and Laplace kernels (Sriperumbudur et al., 2010). On the other hand, linear and polynomial kernels are not characteristic, since their RKHSs are finite dimensional and only provide unique representations of distributions up to certain moments.

The kernel mean embedding (1) is the key ingredient of a well-known metric on probability measures called maximum mean discrepancy (MMD) (Borgwardt et al., 2006; Gretton

1. See, *e.g.*, Diestel and Uhl (1977, Chapter 2) and Dinculeanu (2000, Chapter 1) for the definition of Bochner integral.

et al., 2012). For two distributions $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$, their MMD is given as the RKHS distance between the corresponding kernel means $\mu_{\mathbb{P}}$, $\mu_{\mathbb{Q}}$:

$$\mathrm{MMD}[\mathscr{H}, \mathbb{P}, \mathbb{Q}] := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}} = \sup_{f \in \mathscr{H}, \|f\|_{\mathscr{H}} \leq 1} \left| \int f(\mathbf{x}) \, \mathrm{d}\mathbb{P}(\mathbf{x}) - \int f(\mathbf{x}) \, \mathrm{d}\mathbb{Q}(\mathbf{x}) \right|, \quad (2)$$

where the second identity follows from the reproducing property and $\mathscr{H}$ being a vector space (Gretton et al., 2012, Lemma 4). The right expression is the maximum discrepancy between the means of functions from the unit ball of the RKHS $\mathscr{H}$, and is the original definition of MMD. Being defined via the RKHS distance, MMD is a pseudo-metric on $\mathscr{P}$. Moreover, if $\mathscr{H}$ is characteristic, $\mathrm{MMD}[\mathscr{H}, \mathbb{P}, \mathbb{Q}] = 0$ holds if and only if $\mathbb{P} = \mathbb{Q}$, and thus MMD becomes a proper metric on probability measures. See Sriperumbudur et al. (2010); Simon-Gabriel and Schölkopf (2018) for details and relationships to other popular metrics on probability measures.

Given an i.i.d. (identically and independently distributed) sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathbb{P}$, the kernel mean $\mu_{\mathbb{P}}$ can be estimated simply by the empirical average

$$\hat{\mu}_{\mathbb{P}} := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, \mathbf{x}_i). \quad (3)$$

The $\sqrt{n}$-consistency of (3), that is $\|\mu_{\mathbb{P}} - \hat{\mu}_{\mathbb{P}}\|_{\mathscr{H}} = O_p(n^{-1/2})$ as $n \to \infty$, has been established in Song (2008, Theorem 27) and also in Gretton et al. (2012); Lopez-Paz et al. (2015); Tolstikhin et al. (2017). Importantly, this holds without any parametric assumption about the underlying distribution $\mathbb{P}$.

Given another i.i.d. sample $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$ from $\mathbb{Q}$, and defining $\hat{\mu}_{\mathbb{Q}} := \frac{1}{m} \sum_{j=1}^{m} k(\cdot, \mathbf{x}'_j)$ as an estimate of the kernel mean $\mu_{\mathbb{Q}}$, the (squared) MMD (2) can be estimated as

$$\begin{aligned} \widehat{\mathrm{MMD}}^2[\mathscr{H}, \mathbb{P}, \mathbb{Q}] &= \|\hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}}\|^2_{\mathscr{H}} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(\mathbf{x}_i, \mathbf{x}'_j) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(\mathbf{x}'_i, \mathbf{x}'_j), \end{aligned}$$

where the right expression follows from the reproducing property (Gretton et al., 2012, Eq. 5). Applying the triangle inequality, it follows that $|\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathscr{H}} - \|\hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}}\|_{\mathscr{H}}| \leq \|\mu_{\mathbb{P}} - \hat{\mu}_{\mathbb{P}}\|_{\mathscr{H}} + \|\mu_{\mathbb{Q}} - \hat{\mu}_{\mathbb{Q}}\|_{\mathscr{H}} = O_p(n^{-1/2}) + O_p(m^{-1/2})$ as $n, m \to \infty$, implying the consistency of the above estimator of MMD with a parametric convergence rate. This estimator only requires evaluations of the kernel, and therefore is easy to implement in practice. We note that the above MMD estimator is biased, while being consistent; an unbiased estimator is also available for MMD (Gretton et al., 2012, Eq. 3).

2.3 Kernel Mean Embedding of Conditional Distributions

Finally, the notion of KME can be extended to conditional distributions (Song et al., 2009; Grünewälder et al., 2012; Song et al., 2013; Fukumizu et al., 2013). To describe this, let (X, Y) be a random variable taking values in the product space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces. We define a measurable kernel k on $\mathcal{X}$ and let $\mathscr{H}$ be the associated RKHS. Similarly, we define a measurable kernel ℓ on $\mathcal{Y}$ and let $\mathscr{F}$ be the associated RKHS.

Let $\mathbb{P}_{XY}$ be the joint distribution of (X, Y), and $\mathbb{P}_{Y|X=\mathbf{x}}$ be the conditional distribution of Y given $X = \mathbf{x}$.

The KME of the conditional distribution $\mathbb{P}_{Y|X=\mathbf{x}}$ is then defined as the conditional expectation of $\ell(\cdot, \mathbf{y})$ with respect to $\mathbb{P}_{Y|X=\mathbf{x}}$:

$$\mu_{Y|X=\mathbf{x}} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y|X=\mathbf{x}}(\mathbf{y}) \in \mathscr{F} \quad (\mathbf{x} \in \mathcal{X}). \tag{4}$$

Again, if $\mathscr{F}$ is characteristic, this kernel mean maintains all information about $\mathbb{P}_{Y|X=\mathbf{x}}$, thus being qualified as its representation. It is instructive to note that $\mu_{Y|X=\mathbf{x}}$ is defined for each $\mathbf{x} \in \mathcal{X}$ individually.

Given an i.i.d. sample $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ from the joint distribution $\mathbb{P}_{XY}$, the conditional mean embedding (4) can be estimated as

$$\hat{\mu}_{Y|X=\mathbf{x}} := \sum_{i=1}^{n} w_i(\mathbf{x}) \ell(\cdot, \mathbf{y}_i), \tag{5}$$

where

$$\begin{aligned} (w_1(\mathbf{x}), \ldots, w_n(\mathbf{x}))^\top &:= (\mathbf{K} + n\varepsilon \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}) \in \mathbb{R}^n, \\ \mathbf{k}(\mathbf{x}) &:= (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n))^\top \in \mathbb{R}^n. \end{aligned}$$

Here, $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix such that $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, and $\varepsilon > 0$ is a regularization constant. As pointed out by Grünewälder et al. (2012), this estimator can be interpreted as that of *function-valued kernel ridge regression*, where the task is to estimate the mapping $\mathbf{x} \mapsto \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y|X=\mathbf{x}}(\mathbf{y})$ from training data $(\mathbf{x}_1, \ell(\cdot, \mathbf{y}_1)), \ldots, (\mathbf{x}_n, \ell(\cdot, \mathbf{y}_n)) \in \mathcal{X} \times \mathscr{F}$. In fact, the weights $w_1(\mathbf{x}), \ldots, w_n(\mathbf{x})$ in (5) are identical to those of kernel ridge regression (or Gaussian process regression). As such, the regularization constant ε should decay to 0 at an appropriate speed as $n \to \infty$, in order to ensure a good convergence rate of the estimator (5), see, *e.g.*, Caponnetto and Vito (2007).

3. Counterfactual Mean Embeddings

In this section, we formulate our problem of estimating distributional treatment effects and describe our approach. In Section 3.1, we review the potential outcome framework and, based on it, we define distributional treatment effects. The key concepts here are counterfactual distributions on outcomes. In Section 3.2, we describe our approach, *counterfactual mean embeddings*, as the kernel mean embeddings of counterfactual distributions. Section 3.3 provides details of the distributional effects of covariate distributions, which are essential for applications in off-policy evaluation. We then define their empirical estimators in Section 3.4. Finally, we introduce the kernel treatment effect (KTE) as a way to evaluate the distributional treatment effect in Section 3.5.

3.1 Potential Outcome Framework and Distributional Causal Effects

We pose our problem based on the potential outcome framework, also known as the Neyman-Rubin causal model, which is a classic and widely used approach to estimating causal effects of treatments from observational data (Neyman, 1923; Rubin, 1974, 2005).

We consider a hypothetical subject (*e.g.*, a patient) in a population. Let $X \in \mathcal{X}$ be a *covariate* random variable representing the subject's features (*e.g.*, age, weight, blood pressure, etc.), where $\mathcal{X}$ is a measurable space. Let $T \in \mathcal{T}$ a random variable that indicates the *treatment* assigned to the subject, where $\mathcal{T}$ denotes the set of treatments of interest. We call T *treatment indicator* or *treatment assignment*. In this work, we focus on binary treatments $\mathcal{T} := \{0, 1\}$ for simplicity, but an extension to multiple treatments is straightforward. For instance, $T = 1$ may represent that the subject is assigned an active treatment (e.g., a drug of interest), and $T = 0$ a control treatment (e.g., placebo).

Let $Y_0^*, Y_1^* \in \mathcal{Y}$ be random variables representing *potential* outcomes, where $\mathcal{Y}$ is a measurable space. That is, Y_1^* represents the outcome of interest after the subject is exposed to treatment 1, and Y_0^* the outcome after the subject is exposed to treatment 0. For instance, Y_1^* may be the blood pressure of the patient measured after the patient had the drug, and Y_0^* be that after having nothing. The problem here, known as the fundamental problem of causal inference, is that one can only observe either Y_1^* or Y_0^*, but not both. For instance, if one gave the drug to the patient and measured the resulting blood pressure, it is no longer possible to measure the blood pressure of the same patient without the drug. Thus, the *observed* outcome $Y \in \mathcal{Y}$ can be defined as

$$Y := \mathbb{1}(T = 0)Y_0^* + \mathbb{1}(T = 1)Y_1^*,$$

where $\mathbb{1}(T = j) := 1$ if $T = j$, and zero otherwise. Note that in observational studies, the treatment assignment may not be completely random, *i.e.*, T depends on Y_0^*, Y_1^* and X.

Assume that there are N subjects, and that each subject $i = 1, \ldots, N$ is associated with random variables $(\mathbf{x}_i, t_i, \mathbf{y}_{0i}^*, \mathbf{y}_{1i}^*)$ that are distributed as (X, T, Y_0^*, Y_1^*) independently to the other subjects,[2] *i.e.*,

$$(\mathbf{x}_i, t_i, \mathbf{y}_{0i}^*, \mathbf{y}_{1i}^*)_{i=1}^N \sim (X, T, Y_0^*, Y_1^*), \quad \text{i.i.d.} \tag{6}$$

Note that for each subject i, only one of $\mathbf{y}_{0i}^*$ or $\mathbf{y}_{1i}^*$ can be observed. Thus, observational data given to the analyst are

$$(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^N, \quad \mathbf{y}_i := \mathbb{1}(t_i = 0)\mathbf{y}_{0i}^* + \mathbb{1}(t_i = 1)\mathbf{y}_{1i}^*, \tag{7}$$

which are i.i.d. with (X, T, Y). We write $n := \sum_{i=1}^N \mathbb{1}(t_i = 0)$ the number of subjects receiving treatment $T = 0$, and $m := \sum_{i=1}^N \mathbb{1}(t_i = 1)$ that of treatment $T = 1$.

We consider three kinds of distributional causal effect, as described below. For ease of understanding, we also present the corresponding expressions based on the sample (6). Nevertheless, these sample expressions are also counterfactual quantities due to the fundamental problem of causal inference.

3.1.1 Distributional Treatment Effect (DTE)

Let $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$ be the distributions of the potential outcomes Y_0^* and Y_1^*, respectively. Then we define the distributional treatment effect (DTE) as the difference between these two distributions:

$$\mathbb{P}_{Y_0^*}(\cdot) - \mathbb{P}_{Y_1^*}(\cdot). \tag{8}$$

2. This independence assumption may be seen as a version of the Stable Unit Treatment Value Assumption (SUTVA), which requires that the potential outcomes of any subject i are independent of the treatments t_j assigned to the other subjects $j \neq i$.

The corresponding sample expression is given by

$$\frac{1}{N}\sum_{i=1}^{N}\delta(\cdot-\mathbf{y}_{0i}^*)-\frac{1}{N}\sum_{i=1}^{N}\delta(\cdot-\mathbf{y}_{1i}^*),$$

where δ is the Dirac distribution. As mentioned, this sample expression cannot be obtained from observational data (7), since for each subject i, we only have either $\mathbf{y}_{0i}^*$ or $\mathbf{y}_{1i}^*$.

The DTE (8) can capture the treatment effects on the potential outcomes that may not be identified only by the average treatment effect (ATE) (Imbens, 2004), the difference between the expectations of Y_0^* and Y_1^*:

$$\mathrm{ATE} := \mathbb{E}[Y_0^*]-\mathbb{E}[Y_1^*] \tag{9}$$

or its corresponding sample version

$$\mathrm{ATE}_N := \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_{0i}^*-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_{1i}^*.$$

For instance, even when the ATE is 0, the higher order moments of $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$, such as their variances, may differ. The DTE can capture such a difference, while the ATE cannot.

3.1.2 Distributional Treatment Effects on the Treated

This is defined as the difference in two conditional distributions as

$$\mathbb{P}_{Y_1^*|T}(\cdot\,|\,t)-\mathbb{P}_{Y_0^*|T}(\cdot\,|\,t),\quad t\in\{0,1\}. \tag{10}$$

For $t=1$, this can be understood as the distributional treatment effect for the treated, and the corresponding sample expression is given by

$$\frac{1}{m}\sum_{i=1}^{N}\mathbb{1}(t_i=1)\delta(\cdot-\mathbf{y}_{1i}^*)-\frac{1}{m}\sum_{i=1}^{N}\mathbb{1}(t_i=1)\delta(\cdot-\mathbf{y}_{0i}^*),$$

where the second term is counterfactual. The details of the conditional treatment effect (10) can be found, for example, in Chernozhukov et al. (2013, p.2214).

3.1.3 Distributional Effects of the Covariate Distributions

This is defined as the difference between the conditional distribution of Y_0^* given $T=0$ and that of Y_0^* given $T=1$:

$$\mathbb{P}_{Y_0^*|T}(\cdot\,|\,0)-\mathbb{P}_{Y_0^*|T}(\cdot\,|\,1) \tag{11}$$

where $\mathbb{P}_{Y_0^*|T}$ is the conditional distribution of Y_0^* given T. A similar definition can be given for Y_1^*. The corresponding sample expression is given by

$$\frac{1}{n}\sum_{i=1}^{N}\mathbb{1}(t_i=0)\delta(\cdot-\mathbf{y}_{0i}^*)-\frac{1}{m}\sum_{i=1}^{N}\mathbb{1}(t_i=1)\delta(\cdot-\mathbf{y}_{0i}^*). \tag{12}$$

Note that the second term in (11) and (12) are counterfactual in the sense that the potential outcome $\mathbf{y}_{0i}^*$ of subject i with $t_i = 1$ is not observable.

The above distributional differences capture the effects caused by the difference in the characteristics (*i.e.*, *covariates*) of subjects exposed to different treatments, e.g., selection bias, rather than the effects caused by the treatment itself. For instance, let us assume that a drug was assigned to subjects whose blood pressures were already high ($t_i = 1$) and not assigned to subjects with low blood pressures ($t_i = 0$). Then the counterfactual blood pressures $\mathbf{y}_{0i}^*$ of the subjects with $t_i = 1$, which would have been observed if they had not taken the drug, would be higher than those $\mathbf{y}_{0i}^*$ with $t_i = 0$.

This kind of "selection bias" can be captured in the above distributional difference, and this helps understand how the difference in *observed* outcome distributions arises. To explain this more precisely, however, we need the notation, definitions and assumptions introduced in the next subsection. Thus, we defer further explanations to Section 3.3. There, we also explain that this distributional difference is useful in studying counterfactual effects of a policy defined as a specification of a covariate distribution. In fact, this is how we formulate the problem of off-policy evaluation in Section 6.

3.2 Counterfactual Distributions

To deal with distributional treatment effects discussed in the previous subsection, we need to introduce the notion of counterfactual distributions (Chernozhukov et al., 2013). We first summarize the notation defined above and introduce new ones, which we follow Chernozhukov et al. (2013, Appendix C).

Definition 2 *Let Y_0^* and Y_1^* be random variables taking values in $\mathcal{Y}$, and X and T be random variables taking values in $\mathcal{X}$ and $\mathcal{T} = \{0, 1\}$, respectively. The random variables Y, Y_t and X_t $(t = 0, 1)$ are defined as*

$$\begin{aligned} Y &:= \mathbb{1}(T = 0)Y_0^* + \mathbb{1}(T = 1)Y_1^*, \\ Y_t &:= Y \mid T = t \quad (t = 0, 1), \\ X_t &:= X \mid T = t \quad (t = 0, 1). \end{aligned}$$

In Definition 2, Y is the observed outcome variable. Thus, Y_t is Y given that the treatment assignment is $T = t$ $(t = 0, 1)$. By the definition of Y, this implies that $Y_t = Y_t^* \,|\, (T = t)$, that is, Y_t is the potential outcome conditional on $T = t$. Note that, since Y_t^* and T may be dependent, $Y_t^* \,|\, (T = t)$ may differ from Y_t^* as a random variable. The variable X_t is the covariate variable X conditional on $T = t$. The pair of variables (X_t, Y_t) can thus be seen as observed random variables conditional on the treatment assignment $T = t$.

The following is a key assumption, which is needed in general for counterfactual inference with observational data.

Assumption 1 ***(A1)*** **Conditional exogeneity***: $Y_0^*, Y_1^* \perp\!\!\!\perp T \,|\, X$ almost surely for X.*

(A2) **Support condition***: $\mathcal{X}_0 = \mathcal{X}_1$, where $\mathcal{X}_j$ is the support of the distribution $\mathbb{P}_{X_j}$ of X_j for $j = 0, 1$.*

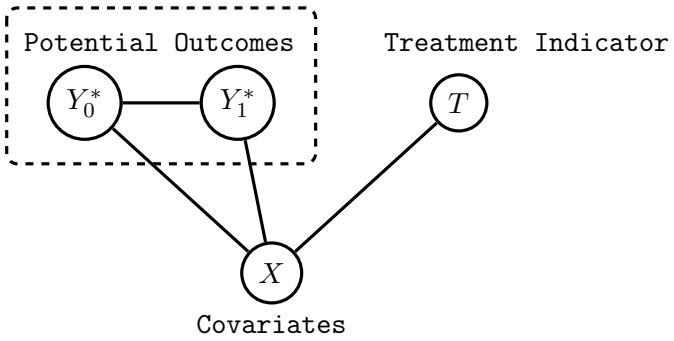


Figure 1: A graphical representation of the conditional exogeneity assumption. An edge between two random variables indicates that they are dependent. The conditional exogeneity assumption states that given the covariates X, the potential outcomes Y_0^*, Y_1^* and the treatment assignment T are conditionally independent. The assumption does not hold if there exists an edge between Y_0^*, Y_1^* and T, which is the case, for instance, when there exists a hidden confounder Z that is dependent to both Y_0^*, Y_1^* and T.

The conditional exogeneity (A1), also known as the *unconfoundedness* or *ignorability*, is a common assumption in observational studies to guarantee the identifiability of causal effects from observational data (Rosenbaum and Rubin, 1983; Imbens, 2004; Rubin, 2005). It requires that there is no hidden confounder, say Z, that affects both the treatment assignment T and potential outcomes Y_0^*, Y_1^*. In other words, the covariates X include all important characteristics regarding the potential outcomes. This assumption is described further in Figure 1, where the graphical model represents the conditional independence structure between the random variables. The support condition (A2) is needed to make the counterfactual distribution (introduced in (13) below) well-defined, and is also made in Chernozhukov et al. (2013, Eq. 2.3). It is analogous to the overlap assumption required for propensity score methods (*e.g.* Imbens, 2004, Assumption 2.2).

We now define counterfactual distributions. Let $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_1}$ be the probability distributions of X_0 and X_1, respectively. Denote by $\mathbb{P}_{Y\langle 0|0\rangle}$ and $\mathbb{P}_{Y\langle 1|1\rangle}$ the corresponding marginal distributions of outcomes defined by

$$\begin{aligned}\mathbb{P}_{Y\langle 0|0\rangle}(\mathbf{y}) &:= \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) = \mathbb{P}_{Y_0}(\mathbf{y}) \\ \mathbb{P}_{Y\langle 1|1\rangle}(\mathbf{y}) &:= \int \mathbb{P}_{Y_1|X_1}(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) = \mathbb{P}_{Y_1}(\mathbf{y})\end{aligned}$$

where $\mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})$ is the conditional distribution of Y_0 given X_0, and $\mathbb{P}_{Y_1|X_1}(\mathbf{y}|\mathbf{x})$ is that of Y_1 given X_1. Following Chernozhukov et al. (2013), *counterfactual distributions* are then defined as

$$\mathbb{P}_{Y\langle 0|1\rangle}(\mathbf{y}) := \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}), \tag{13}$$

$$\mathbb{P}_{Y\langle 1|0\rangle}(\mathbf{y}) \quad := \quad \int \mathbb{P}_{Y_1|X_1}(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}), \tag{14}$$

which are well-defined as long as the support condition in Assumption 1 is satisfied.

The distributions introduced above are defined in terms of the observed random variables $(X_t, Y_t)_{t=0,1}$. We now see how these distributions are related to the distributions on potential outcomes that appear in distributional causal effects (10) and (11). First, as summarized in the following lemma, $\mathbb{P}_{Y\langle 0|0\rangle}$ and $\mathbb{P}_{Y\langle 1|1\rangle}$ are nothing but $\mathbb{P}_{Y_0^*|T}(\mathbf{y}|0)$ and $\mathbb{P}_{Y_1^*|T}(\mathbf{y}|1)$, respectively. For completeness, we include the proof in Appendix C.1.

Lemma 3 *We have* $\mathbb{P}_{Y\langle 0|0\rangle}(\mathbf{y}) = \mathbb{P}_{Y_0^*|T}(\mathbf{y}|0)$ *and* $\mathbb{P}_{Y\langle 1|1\rangle}(\mathbf{y}) = \mathbb{P}_{Y_1^*|T}(\mathbf{y}|1)$.

On the other hand, the counterfactual distributions $\mathbb{P}_{Y\langle 0|1\rangle}$ and $\mathbb{P}_{Y\langle 1|0\rangle}$ are respectively equal to distributions $\mathbb{P}_{Y_0^*|T}(\mathbf{y}|1)$ and $\mathbb{P}_{Y_1^*|T}(\mathbf{y}|0)$ appearing in (11) and (10), provided that Assumption 1 holds (Chernozhukov et al., 2013, Lemma 2.1); we provide a proof for completeness in Appendix C.2.

Lemma 4 (Causal interpretation) *Suppose that Assumption 1 is satisfied. Then we have* $\mathbb{P}_{Y\langle 0|1\rangle} = \mathbb{P}_{Y_0^*|T=1}$ *and* $\mathbb{P}_{Y\langle 1|0\rangle} = \mathbb{P}_{Y_1^*|T=0}$.

Lemma 4 shows that the distributions $\mathbb{P}_{Y_0^*|T=1}$ and $\mathbb{P}_{Y_1^*|T=0}$, which play the key role in analyzing distributional treatment effects (10) (11), can be obtained by estimating the corresponding counterfactual distributions $\mathbb{P}_{Y\langle 0|1\rangle}$ and $\mathbb{P}_{Y\langle 1|0\rangle}$ defined in terms of observed random variables $(X_t, Y_t)_{t=0,1}$. The key assumption in this regard is the conditional exogeneity in Assumption 1.

3.3 Further Explanation on the Distributional Effects of Covariate Distributions

We are now in a position to provide further explanation on the distributional difference introduced in Section 3.1.3. To this end, let us assume that the conditional exogeneity in Assumption 1 is satisfied. Then, by Lemmas 3 and 4, the distributional difference in (11) can be written as

$$\begin{aligned}\mathbb{P}_{Y_0^*|T}(\mathbf{y}\,|\,0) - \mathbb{P}_{Y_0^*|T}(\mathbf{y}\,|\,1) &= \mathbb{P}_{Y\langle 0|0\rangle}(\mathbf{y}) - \mathbb{P}_{Y\langle 0|1\rangle}(\mathbf{y}) \\ &= \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) - \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}). \end{aligned} \tag{15}$$

The rhs of (15) shows that this distributional difference (if it exists) is due to the difference between the covariate distributions $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_1}$. In what follows, we provide two distinct interpretations. First, it quantifies a selection bias that affects the difference in observed outcome distributions. Second, it quantifies the causal effect for a policy implemented as a specification of a covariate distribution.

3.3.1 Quantifying a Selection Bias

We can decompose the difference in the *observed* outcome distributions $\mathbb{P}_{Y_0}$ and $\mathbb{P}_{Y_1}$ as

$$\mathbb{P}_{Y_0}(\mathbf{y}) - \mathbb{P}_{Y_1}(\mathbf{y}) \quad = \quad \mathbb{P}_{Y_0^*|T}(\mathbf{y}\,|\,0) - \mathbb{P}_{Y_1^*|T}(\mathbf{y}\,|\,1)$$

$$= \underbrace{\mathbb{P}_{Y_0^*|T}(\mathbf{y}\,|\,0) - \mathbb{P}_{Y_0^*|T}(\mathbf{y}\,|\,1)}_{(A)} + \underbrace{\mathbb{P}_{Y_0^*|T}(\mathbf{y}\,|\,1) - \mathbb{P}_{Y_1^*|T}(\mathbf{y}\,|\,1)}_{(B)},$$

where the first term (A) is the distributional effect of covariate distributions (15), and the second term (B) is the distributional treatment effect on the treated. Thus, the difference in the observed outcome distributions can arise from (A) and/or (B), and the estimation of (A) and (B) is useful in studying the origin of the difference in observed outcome distributions. For instance, if we find that (A) is zero, the difference between the observed outcome distributions is originated from the distributional difference on the treated (B). On the other hand, if (B) is zero, then the difference between the observed outcome distributions is due to (A), i.e., by the selection bias, and is not due to the effects of the treatment.

Note that this difference is different from the difference between the *potential* outcome distributions $\mathbb{P}_{Y_0^*}$, $\mathbb{P}_{Y_1^*}$, which accounts for the effects of the treatments 0 and 1 and thus is of primary interest. The observed outcome distributions $\mathbb{P}_{Y_0}$, $\mathbb{P}_{Y_1}$ are biased approximations to the potential outcome distributions, if the treatment assignment is not randomized (i.e., if X and T are not independent).

3.3.2 Policy as a Specification of a Covariate Distribution

The distributional difference (15) can also be used to quantify the effects of a policy that specifies a covariate distribution. Recall that we introduced the random variables X_0, X_1 as the covariate random variable X conditioned on $T = t$, $t \in \{0, 1\}$, i.e., $X_t := X \,|\, (T = t)$. We can instead *directly define* two random variables X_0, X_1 by specifying their probability distributions $\mathbb{P}_{X_0} = \mathbb{P}_{X|T}(\cdot\,|\,0)$ and $\mathbb{P}_{X_1} = \mathbb{P}_{X|T}(\cdot\,|\,1)$, respectively. In this case, the conditioning $T = t$ for $t \in \{0, 1\}$ may be regarded as specifying the covariate distribution $\mathbb{P}_{X_t} = \mathbb{P}_{X|T}(\cdot\,|\,t)$ on the space of covariates $\mathcal{X}$. This specification of the covariate distribution $\mathbb{P}_{X_t} = \mathbb{P}_{X|T}(\cdot\,|\,t)$ *itself* can be regarded as a certain *policy*.[3]

For instance, Rothe (2010, Section 5.2) used this formulation to study the effects of smoking of a pregnant mother on the birth weight of the baby. There, the observed outcome $Y > 0$ is the birth weight of the baby, and covariates $X := (X^1, X^2, X^3, X^4) \in \mathbb{R}^4$ are relevant features of the mother: X^1 is the number of cigarettes per day, X^2 is the age, X^3 is the weight gain and X^4 is the marital status. The distribution $\mathbb{P}_{X_0}$ is the covariate distribution of available data of smoking mothers, while $\mathbb{P}_{X_1}$ is a transformation of $\mathbb{P}_{X_0}$ so that the number of cigarettes per day, X_0^1, is reduced to 75%. Thus, $T = 1$ or $\mathbb{P}_{X_1}$ may be regarded as a hypothetical policy that reduces the amount of cigarettes of smoking pregnant women. Then, $\mathbb{P}_{\langle 0|1\rangle}(\mathbf{y}) = \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}\,|\,\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x})$ is the counterfactual distribution of the birth weights of babies that would have been observed if the mothers had smoked 75 % less amount of cigarettes than they actually did. The comparison to the observed outcome distribution $\mathbb{P}_{\langle 0|0\rangle}(\mathbf{y}) = \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}\,|\,\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_0}(\mathbf{x})$ then enables studying the effects of the amount of cigarettes on birth weights.

Another important instance is the off-policy evaluation task, which will be discussed further in Section 6.

3. Here we use the terminology "policy" instead of "treatment" not to confuse the two notions. In our paper, a "treatment" $t \in \{0, 1\}$ specifies the corresponding *potential* outcome Y_t^* and its distribution $\mathbb{P}_{Y_t^*}$; thus, the difference between $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$ characterizes the treatment effects. On the other hand, a "policy" here $t \in \{0, 1\}$ specifies the corresponding covariate random variable X_t and its distribution $\mathbb{P}_{X_t}$.

3.4 Kernel Mean Embeddings for Counterfactual Distributions

We now define counterfactual mean embeddings. Let ℓ be a positive definite kernel on $\mathcal{Y}$ with RKHS $\mathscr{F}$, and assume that the support condition in Assumption 1 is satisfied. We then refer to the kernel mean embeddings of the counterfactual distributions (13) and (14)

$$\mu_{Y\langle 0|1\rangle} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y\langle 0|1\rangle}(\mathbf{y}) \in \mathscr{F}, \tag{16}$$

$$\mu_{Y\langle 1|0\rangle} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y\langle 1|0\rangle}(\mathbf{y}) \in \mathscr{F}, \tag{17}$$

as *counterfactual mean embeddings (CME)*. Lemma 4 implies that, under Assumption 1, these CMEs are respectively identical to the kernel mean embeddings of $\mathbb{P}_{Y_0^*|T}(\mathbf{y}|1)$ and $\mathbb{P}_{Y_1^*|T}(\mathbf{y}|0)$ defined as

$$\mu_{Y_0^*|T=1} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0^*|T}(\mathbf{y}|1), \quad \mu_{Y_1^*|T=0} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_1^*|T}(\mathbf{y}|0).$$

Therefore, by defining an empirical estimator of the CME (16), one can hope to estimate the distributional treatment effects in (10) and (11), which will be done below.

Estimating counterfactual mean embeddings. In what follows, we introduce our estimator of the CME $\mu_{Y\langle 0|1\rangle}$ defined in (16); one can define an estimator of (17) in a similar manner. In practice, it is not possible to obtain a sample from $\mathbb{P}_{Y\langle 0|1\rangle}$, and therefore the counterfactual mean embedding $\mu_{Y\langle 0|1\rangle}$ cannot be estimated directly. Instead, we propose an estimator that uses samples from $\mathbb{P}_{X_0Y_0}$ and $\mathbb{P}_{X_1}$ to estimate $\mu_{Y\langle 0|1\rangle}$. To this end, first note that $\mu_{Y\langle 0|1\rangle}$ in (16) can be written in terms of the conditional mean embedding (4) of $\mathbb{P}_{Y_0|X_0=\mathbf{x}}$:

$$\mu_{Y\langle 0|1\rangle} = \int \mu_{Y_0|X_0=\mathbf{x}} \, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) \in \mathscr{F},$$

where $\mu_{Y_0|X_0=\mathbf{x}} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0|X_0=\mathbf{x}}(\mathbf{y}) \in \mathscr{F}$. This formulation suggests that $\mu_{Y\langle 0|1\rangle}$ can be estimated by i) constructing an estimator of the conditional mean embedding $\mu_{Y_0|X_0=\mathbf{x}}$ and then ii) taking its average over $\mathbb{P}_{X_1}(\mathbf{x})$. This is how our estimator is derived below.

Suppose that we are given independent samples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ from $\mathbb{P}_{Y_0X_0}(\mathbf{x}, \mathbf{y})$ and $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$ from $\mathbb{P}_{X_1}(\mathbf{x})$. For $\mathbf{x} \in \mathcal{X}$, let $\hat{\mu}_{Y_0|X_0=\mathbf{x}}$ denote the estimate (5) of the conditional mean embedding $\mu_{Y_0|X_0=\mathbf{x}}$ based on $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$. Then, an empirical estimator of $\mu_{Y\langle 0|1\rangle}$ is defined and expressed as

$$\hat{\mu}_{Y\langle 0|1\rangle} := \frac{1}{m} \sum_{j=1}^{m} \hat{\mu}_{Y_0|X_0=\mathbf{x}'_j} = \sum_{i=1}^{n} \beta_i \ell(\cdot, \mathbf{y}_i) \quad \text{with} \quad (\beta_1, \ldots, \beta_n)^\top = (\mathbf{K} + n\varepsilon \mathbf{I})^{-1} \widetilde{\mathbf{K}} \mathbf{1}_m, \tag{18}$$

where $\varepsilon > 0$ is a regularization constant, $\mathbf{1}_m = (1/m, \ldots, 1/m)^\top \in \mathbb{R}^m$, $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and $\widetilde{\mathbf{K}} \in \mathbb{R}^{n \times m}$ with $\widetilde{\mathbf{K}}_{ij} = k(\mathbf{x}_i, \mathbf{x}'_j)$.

The proposed estimator (18) is nonparametric, and can be implemented without knowledge about parametric forms of the conditional $\mathbb{P}_{Y_0|X_0}$ and marginal $\mathbb{P}_{X_1}$. Thus, the estimator is useful when such knowledge is not available. In Section 4, we theoretically analyze the asymptotic behavior of the estimator, proving its consistency and deriving convergence

rates. In doing so, we elucidate conditions required for the consistency of the proposed estimator.

The computational complexity of our estimator (18) is $\mathcal{O}(n^3)$ because of the matrix inversion, which may be expensive when the sample size n is huge. To reduce the complexity, one can adopt existing approximation methods such as Nyström method and random Fourier features (Williams and Seeger, 2001; Rahimi and Recht, 2008).

We note that the form of the estimator is identical to the *kernel sum rule* (Song et al., 2013, Section 4.1), a mean embedding approach to computing forward probabilities in Bayesian inference. The way we use the estimator is different from this previous approach, however. That is, we use our estimator to estimate the counterfactual distribution and distributional causal effects (11), and this requires Assumption 1 to hold for data (or for the population random variables), as shown in Lemma 4.

3.5 Kernel Treatment Effects

We quantify distributional treatment effects by using the RKHS distance between the mean embeddings of potential outcome distributions under consideration. We call this approach *Kernel Treatment Effects (KTE)*. We show below how KTEs can be defined for the different distributional treatment effects discussed in Section 3.1.

3.5.1 KTE for Distributional Treatment Effects

As before, let ℓ be a kernel on the output space $\mathcal{Y}$ and $\mathscr{F}$ be its RKHS. For the distributional treatment effect (8) discussed in Section 3.1.1, the corresponding KTE is defined as

$$\mathrm{KTE}(Y_0^*, Y_1^*, \mathscr{F}) := \|\mu_{Y_0^*} - \mu_{Y_1^*}\|_{\mathscr{F}}, \tag{19}$$

where $\mu_{Y_0^*}$ and $\mu_{Y_1^*}$ are the kernel mean embeddings of the distributions of potential outcomes $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$, respectively, i.e.,

$$\mu_{Y_0^*} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0^*}(\mathbf{y}), \qquad \mu_{Y_1^*} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_1^*}(\mathbf{y}). \tag{20}$$

The KTE (19) may be regarded as a generalization of the ATE (9) in the sense that, if ℓ is the linear kernel $\ell(\mathbf{y}, \mathbf{y}') = \langle \mathbf{y}, \mathbf{y}' \rangle$ on $\mathcal{Y} = \mathbb{R}^d$, then the KTE only distinguishes the means of the two outcome distributions. By using a different kernel ℓ, the KTE may capture the differences between higher-order statistics of the outcome distributions $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$. For instance, if ℓ is a polynomial kernel $\ell(\mathbf{y}, \mathbf{y}') = (\langle \mathbf{y}, \mathbf{y}' \rangle + c)^m$ of degree $m \in \mathbb{N}$ with $c > 0$, then the KTE (19) is equal to 0 if and only if $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$ have the same moments up to degree m (see, *e.g.*, Muandet et al. 2017, Chapter 3).

If ℓ is a characteristic kernel, such as Gaussian and Matérn kernels, then the KTE (19) is equal to 0 if and only if the two distributions $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$ are the same. In this case, the KTE takes a positive value if and only if there is a difference between $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$. This means that the KTE informs the existence of any difference in the potential outcome distributions, quantifying the distributional treatment effect.

The question is how to estimate the KTE (19) from data. As in (7), let $(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^N$ be observational data, which are i.i.d. with the random variables (X, T, Y). Recall that Y is the

observed outcome and thus given by $Y = \mathbb{1}(T=0)Y_0^* + \mathbb{1}(T=1)Y_1^*$ where Y_0^* and Y_1^* are the potential outcomes. In observational studies, it is common to use the *propensity score* $e(\mathbf{x}) := \mathbb{E}[T \,|\, X = \mathbf{x}]$, the conditional probability of the treatment assignment T being made given that the covariates are $X = \mathbf{x}$, to define an unbiased estimator of the average treatment effect $\mathbb{E}[Y_1^*] - \mathbb{E}[Y_0^*]$ (Rosenbaum and Rubin, 1983). We show here that the same strategy of *inverse propensity weighting* (Imbens, 2004, Section III-C) can be straightforwardly used to define unbiased estimators of the mean embeddings $\mu_{Y_1^*}$ and $\mu_{Y_0^*}$ of potential outcome distributions $\mathbb{P}_{Y_1^*}$ and $\mathbb{P}_{Y_0^*}$, respectively, thus providing a way of estimating the KTE. That is, assuming that the propensity $e(\mathbf{x})$ is available, we define

$$\hat{\mu}_{Y_1^*} := \frac{1}{m}\sum_{i=1}^{N} \frac{t_i \ell(\cdot, \mathbf{y}_i)}{e(\mathbf{x}_i)}, \qquad \hat{\mu}_{Y_0^*} := \frac{1}{n}\sum_{j=1}^{N} \frac{(1-t_j)\ell(\cdot, \mathbf{y}_j)}{1-e(\mathbf{x}_j)}, \tag{21}$$

where $m := \sum_{i=1}^N t_i$ and $n := \sum_{j=1}^N (1-t_j)$ are the populations of treated and control groups, respectively.

In the special case of a completely randomized experiment where X and T are independent and thus the propensity is $e(\mathbf{x}) = 1/2$ for all $\mathbf{x} \in \mathcal{X}$, the above estimators reduce to the standard empirical estimators of mean embeddings: $\hat{\mu}_{Y_1^*} := \frac{2}{m}\sum_{i=1}^N t_i \ell(\cdot, \mathbf{y}_i)$ and $\hat{\mu}_{Y_0^*} := \frac{2}{n}\sum_{j=1}^N (1-t_j)\ell(\cdot, \mathbf{y}_j)$. Note that these uniformly-weighted empirical estimators are *biased* if the experiment is *not* completely randomized, *i.e.*, in observational studies. This is because, for instance, the sample $\mathbf{y}_i$ contributing to $\hat{\mu}_{Y_1^*}$ follows the distribution of $Y_1^*|T=1$, which is different from the unconditional Y_1^*. Thus, we need the inverse propensity weighting to obtain unbiased estimators in the case of observational studies.

The following result shows that the estimators (21) are indeed unbiased estimators of the corresponding mean embeddings $\mu_{Y_1^*}$ and $\mu_{Y_0^*}$ of potential outcome distributions. The proof is presented in Appendix C.3.

Theorem 5 *Suppose that $0 < e(\mathbf{x}) < 1$ for all $\mathbf{x} \in \mathcal{X}$ and that the conditional exogeneity in Assumption 1 is satisfied. Let $(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^N$ be i.i.d. with (X, T, Y), and let $\hat{\mu}_{Y_1^*}$ and $\hat{\mu}_{Y_0^*}$ be the estimators (21) of the mean embeddings $\mu_{Y_1^*}$ and $\mu_{Y_0^*}$ of the potential outcome distributions $\mathbb{P}_{Y_1^*}$ and $\mathbb{P}_{Y_0^*}$ in (20). Then, we have*

$$\mathbb{E}[\hat{\mu}_{Y_1^*}] = \mu_{Y_1^*}, \qquad \mathbb{E}[\hat{\mu}_{Y_0^*}] = \mu_{Y_0^*}.$$

Theorem 5 shows that the estimators (21) are unbiased, but does not say anything about their convergence rates as the sample size goes to infinity. The following result provides this; it essentially shows that the estimators (21) converge to the mean embeddings $\mu_{Y_1^*}$ and $\mu_{Y_0^*}$ at the same rates as the standard kernel mean estimators, which are minimax optimal (Tolstikhin et al., 2017). The key assumption here is that the propensity $e(\mathbf{x})$ is uniformly lower- and upper-bounded away from 0 and 1. The proof is presented in Appendix C.4.

Theorem 6 *Suppose the propensity score $e(\mathbf{x})$ satisfies $\inf_{\mathbf{x}\in\mathcal{X}} e(\mathbf{x}) > 0$ and $\sup_{\mathbf{x}\in\mathcal{X}} e(\mathbf{x}) < 1$, that $\sup_{\mathbf{y}\in\mathcal{Y}} \ell(\mathbf{y}, \mathbf{y}) < \infty$, and that the conditional exogeneity in Assumption 1 is satisfied. Let $(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^N$ be i.i.d. with (X, T, Y), and let $\hat{\mu}_{Y_1^*}$ and $\hat{\mu}_{Y_0^*}$ be the estimators (21) of the*

mean embeddings $\mu_{Y_1^*}$ *and* $\mu_{Y_0^*}$ *of the potential outcome distributions* $\mathbb{P}_{Y_1^*}$ *and* $\mathbb{P}_{Y_0^*}$ *in* (20). *Then, we have*

$$\mathbb{E}\left[\left\|\hat{\mu}_{Y_1^*} - \mu_{Y_1^*}\right\|_{\mathscr{F}}^2\right] = O(m^{-1}), \qquad \mathbb{E}\left[\left\|\hat{\mu}_{Y_0^*} - \mu_{Y_0^*}\right\|_{\mathscr{F}}^2\right] = O(n^{-1}), \qquad (N \to \infty),$$

where $m := \sum_{i=1}^N t_i$ *and* $n := \sum_{i=1}^N (1 - t_i)$.

Based on the estimators (21), we can define a consistent estimator of (19) as

$$\begin{aligned}\widehat{\mathrm{KTE}}_b^2(Y_0^*, Y_1^*, \mathscr{F}) &:= \left\|\hat{\mu}_{Y_1^*} - \hat{\mu}_{Y_0^*}\right\|_{\mathscr{F}}^2 = \left\|\hat{\mu}_{Y_1^*}\right\|_{\mathscr{F}}^2 - 2\left\langle \hat{\mu}_{Y_1^*}, \hat{\mu}_{Y_0^*}\right\rangle_{\mathscr{F}} + \left\|\hat{\mu}_{Y_0^*}\right\|_{\mathscr{F}}^2 \\ &= \frac{1}{m^2}\sum_{i,j=1}^N \frac{t_i t_j \ell(\mathbf{y}_i, \mathbf{y}_j)}{e(\mathbf{x}_i)e(\mathbf{x}_j)} - \frac{2}{mn}\sum_{i,j=1}^N \frac{t_i(1-t_j)\ell(\mathbf{y}_i, \mathbf{y}_j)}{e(\mathbf{x}_i)(1-e(\mathbf{x}_j))} + \frac{1}{n^2}\sum_{i,j=1}^N \frac{(1-t_i)(1-t_j)\ell(\mathbf{y}_i, \mathbf{y}_j)}{(1-e(\mathbf{x}_i))(1-e(\mathbf{x}_j))},\end{aligned} \tag{22}$$

where the last equality follows from the reproducing property of the kernel ℓ. By the triangle inequality, we have $|\widehat{\mathrm{KTE}}_b(Y_0^*, Y_1^*, \mathscr{F}) - \mathrm{KTE}(Y_0^*, Y_1^*, \mathscr{F})| = |\left\|\hat{\mu}_{Y_1^*} - \hat{\mu}_{Y_0^*}\right\|_{\mathscr{F}} - \|\mu_{Y_0^*} - \mu_{Y_1^*}\|_{\mathscr{F}}| \leq |\left\|\hat{\mu}_{Y_1^*} - \mu_{Y_1^*}\right\|_{\mathscr{F}} + \left\|\hat{\mu}_{Y_0^*} - \mu_{Y_0^*}\right\|_{\mathscr{F}}| = O_p(m^{-1/2} + n^{-1/2})$ as $n, m \to \infty$, which shows that the estimator (22) is asymptotically unbiased.

Note that (22) is a biased estimator, while being asymptotically unbiased. This bias is caused by the terms with identical indices ($i = j$) in the first and third summations of (22). Thus, by subtracting these terms, an unbiased estimator of the KTE can be defined as

$$\begin{aligned}\widehat{\mathrm{KTE}}_u^2(Y_0^*, Y_1^*, \mathscr{F}) &:= \frac{1}{m(m-1)}\sum_{i \neq j} \frac{t_i t_j \ell(\mathbf{y}_i, \mathbf{y}_j)}{e(\mathbf{x}_i)e(\mathbf{x}_j)} \\ &- \frac{2}{mn}\sum_{i,j=1}^N \frac{t_i(1-t_j)\ell(\mathbf{y}_i, \mathbf{y}_j)}{e(\mathbf{x}_i)(1-e(\mathbf{x}_j))} + \frac{1}{n(n-1)}\sum_{i \neq j} \frac{(1-t_i)(1-t_j)\ell(\mathbf{y}_i, \mathbf{y}_j)}{(1-e(\mathbf{x}_i))(1-e(\mathbf{x}_j))}.\end{aligned} \tag{23}$$

By similar arguments as in the proof of Theorem 6, it can be shown that this is indeed an unbiased estimator of (the square of) KTE (19). Moreover, since it can be shown that

$$\left|\widehat{\mathrm{KTE}}_u^2(Y_0^*, Y_1^*, \mathscr{F}) - \widehat{\mathrm{KTE}}_b^2(Y_0^*, Y_1^*, \mathscr{F})\right| = O_p(m^{-1} + n^{-1}) \quad (n, m \to \infty)$$

given that the assumptions in Theorem 6 hold, this unbiased estimator (23) enjoys the same convergence rate as the biased one (22): $|\widehat{\mathrm{KTE}}_u(Y_0^*, Y_1^*, \mathscr{F}) - \mathrm{KTE}(Y_0^*, Y_1^*, \mathscr{F})| = O_p(m^{-1/2} + n^{-1/2})$ as $n, m \to \infty$.

3.5.2 KTE for Distributional Treatment Effects on the Treated

We define KTE for the distributional effect $\mathbb{P}_{Y_1^*|T}(\cdot \,|\, t) - \mathbb{P}_{Y_0^*|T}(\cdot \,|\, t)$ introduced in Section 3.1.2, where $t \in \{0, 1\}$. We only consider the case $t = 1$ here, which is interpreted as the distributional treatment effect for the treated; the case $t = 0$ can be defined similarly. The definition is

$$\mathrm{KTE}(Y_1^* | (T = 1), Y_0^* | (T = 1), \mathscr{F}) := \left\|\mu_{Y_1^*|T=1} - \mu_{Y_0^*|T=1}\right\|_{\mathscr{F}}, \tag{24}$$

where $\mu_{Y_1^*|T=1}$ and $\mu_{Y_0^*|T=1}$ are the kernel mean embeddings of $\mathbb{P}_{Y_1^*|T}(\cdot \,|\, 1)$ and $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$, respectively:

$$\mu_{Y_1^*|T=1} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_1^*|T}(\mathbf{y} \,|\, 1), \qquad \mu_{Y_0^*|T=1} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0^*|T}(\mathbf{y} \,|\, 1).$$

Lemma 3 shows that $\mathbb{P}_{Y\langle 1|1\rangle}(\mathbf{y}) = \mathbb{P}_{Y_1^*|T}(\mathbf{y} \,|\, 1)$, while Lemma 4 implies that $\mathbb{P}_{Y\langle 0|1\rangle} = \mathbb{P}_{Y_0^*|T=1}$ under Assumption 1. Thus, we can define an estimator of the above KTE (24) as follows. Let $\hat{\mu}_{Y\langle 0|1\rangle} = \sum_{i=j}^{n} \beta_i \ell(\cdot, \mathbf{y}_j)$ be the estimator (18) of the CME, and let $\hat{\mu}_{Y\langle 1|1\rangle} := \frac{1}{m} \sum_{i=1}^{m} \ell(\check{\mathbf{y}}_i, \cdot)$ where $\check{\mathbf{y}}_1, \ldots, \check{\mathbf{y}}_m$ is a sample from $\mathbb{P}_{Y\langle 1|1\rangle}$. Note that such $\check{\mathbf{y}}_1, \ldots, \check{\mathbf{y}}_m$ can be obtained in practice, as $\mathbb{P}_{Y\langle 1|1\rangle}$ is the distribution of the observed outcome Y given that the treatment assignment is $T = 1$. Then, we can define an empirical estimator of the KTE in (24) as follows:

$$\begin{aligned}
&\widehat{\mathrm{KTE}}^2(Y_1^*|(T=1), Y_0^*|(T=1), \mathscr{F}) \\
&\quad := \left\| \hat{\mu}_{Y\langle 1|1\rangle} - \hat{\mu}_{Y\langle 0|1\rangle} \right\|_{\mathscr{F}}^2 \\
&\quad = \frac{1}{m^2} \sum_{i,j=1}^{m} \ell(\check{\mathbf{y}}_i, \check{\mathbf{y}}_j) - \frac{2}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} \beta_j \ell(\check{\mathbf{y}}_i, \mathbf{y}_j) + \sum_{i,j=1}^{n} \beta_i \beta_j \ell(\mathbf{y}_i, \mathbf{y}_j).
\end{aligned} \tag{25}$$

3.5.3 KTE for Distributional Effects of the Covariate Distributions

Similarly, we can define a KTE for the distributional effect $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0) - \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$ defined in (11) by

$$\mathrm{KTE}(Y_0^*|(T=0), Y_0^*|(T=1), \mathscr{F}) := \left\| \mu_{Y_0^*|T=0} - \mu_{Y_0^*|T=1} \right\|_{\mathscr{F}}, \tag{26}$$

where $\mu_{Y_0^*|T=0}$ and $\mu_{Y_0^*|T=1}$ are the kernel mean embeddings of $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0)$ and $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$, respectively, i.e.,

$$\mu_{Y_0^*|T=0} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0^*|T}(\mathbf{y} \,|\, 0), \qquad \mu_{Y_0^*|T=1} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0^*|T}(\mathbf{y} \,|\, 1).$$

Lemma 3 shows that $\mathbb{P}_{Y\langle 0|0\rangle}(\mathbf{y}) = \mathbb{P}_{Y_0^*|T}(\mathbf{y} \,|\, 0)$, and Lemma 4, under Assumption 1, implies that $\mathbb{P}_{Y\langle 0|1\rangle} = \mathbb{P}_{Y_0^*|T=1}$. Therefore, we define an estimator of (26) in the following way. Let $\hat{\mu}_{Y\langle 0|1\rangle}$ be the estimator (18) of the CME, and let $\hat{\mu}_{Y\langle 0|0\rangle} := \frac{1}{n} \sum_{i=1}^{n} \ell(\cdot, \tilde{\mathbf{y}}_i)$ where $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n$ is a sample from $\mathbb{P}_{Y\langle 0|0\rangle}$. Note that such $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n$ can be obtained in practice, as $\mathbb{P}_{Y\langle 0|0\rangle}$ is the distribution of the observed outcome Y given that the treatment assignment is $T = 0$. Then, we can define an empirical estimator of the KTE in (26) as follows:

$$\begin{aligned}
&\widehat{\mathrm{KTE}}^2(Y_0^*|(T=0), Y_0^*|(T=1), \mathscr{F}) \\
&\quad := \left\| \hat{\mu}_{Y\langle 0|0\rangle} - \hat{\mu}_{Y\langle 0|1\rangle} \right\|_{\mathscr{F}}^2 \\
&\quad = \frac{1}{n^2} \sum_{i,j=1}^{n} \ell(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j) - \frac{2}{n} \sum_{i,j=1}^{n} \beta_j \ell(\tilde{\mathbf{y}}_i, \mathbf{y}_j) + \sum_{i,j=1}^{n} \beta_i \beta_j \ell(\mathbf{y}_i, \mathbf{y}_j).
\end{aligned} \tag{27}$$

In the next section, we analyze the convergence behavior of the proposed CME estimator in (18) as the sample size n goes to infinity. Readers who are interested in applications may skip the next section and jump to Section 5 and Section 6 directly.

4. Convergence Analysis of the CME Estimator

We provide here a convergence analysis of the CME estimator introduced in Section 3.4. In Section 4.1, we first establish its consistency under mild assumptions. In Section 4.2, we derive its convergence rates by making a quantitative assumption on the smoothness of certain functions involved.

We first introduce necessary notation and definitions. We assume that the covariate space $\mathcal{X}$ and the outcome space $\mathcal{Y}$ are measurable spaces, and that the kernels k and ℓ are measurable on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with $\mathscr{H}$ and $\mathscr{F}$ being their respective RKHSs. Let $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_1}$ be the probability distributions of the random variables $X_0 \in \mathcal{X}$ and $X_1 \in \mathcal{X}$, respectively (see Definition 2 for the definition of these random variables).

Let $L_2(\mathbb{P}_{X_0})$ be the Hilbert space of square-integral functions[4] with respect to $\mathbb{P}_{X_0}$:

$$L_2(\mathbb{P}_{X_0}) := \left\{ f : \mathcal{X} \to \mathbb{R} \mid \int f^2(x) \, \mathrm{d}\mathbb{P}_{X_0}(x) < \infty \right\},$$

which is equipped with the inner product $\langle f, g \rangle_{L_2(\mathbb{P}_{X_0})} := \int f(x) g(x) \, \mathrm{d}\mathbb{P}_{X_0}(x)$ and the resulting norm $\|f\|_{L_2(\mathbb{P}_{X_0})} := \sqrt{\langle f, f \rangle_{L_2(\mathbb{P}_{X_0})}}$. Let $\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}$ be the product measure of $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_0}$ on the product space $\mathcal{X} \times \mathcal{X}$.

4.1 Consistency

To establish the consistency of the CME estimator, we require the following conditions.

Assumption 2 *Assume that the following conditions are satisfied:*

- **(i)** *The kernels k and ℓ are bounded on $\mathcal{X}$ and $\mathcal{Y}$, respectively, that is,* $\sup_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}) < \infty$ *and* $\sup_{\mathbf{y} \in \mathcal{Y}} \ell(\mathbf{y}, \mathbf{y}) < \infty$.
- **(ii)** *The RKHS $\mathscr{H}$ of k is dense in $L_2(\mathbb{P}_{X_0})$.*
- **(iii)** *The distribution $\mathbb{P}_{X_1}$ is absolutely continuous with respect to $\mathbb{P}_{X_0}$ 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)** *$(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ are i.i.d. observations of the random variables (X_0, Y_0), and $\mathbf{x}'_1, \ldots, \mathbf{x}'_m$ are i.i.d. observations of the random variable X_1, with $n = m$.*

Remark 7 We make the following comments on Assumption 2.

- The boundedness condition **(i)** is satisfied, for instance, if k and ℓ are shift-invariant kernels, such as Gaussian and Matérn kernels.
- The condition **(ii)** requires that the RKHS be rich enough to approximate square-integrable functions with respect to $\mathbb{P}_{X_0}$. For instance, this is satisfied by the Gaussian kernel (Steinwart and Christmann, 2008, Theorem 4.63), and therefore by any kernel whose RKHS is larger than that of the Gaussian kernel, such as Laplace and Matèrn kernels (Steinwart and Christmann, 2008, Theorem 4.48).

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

- The condition **(iii)** requires the support of $\mathbb{P}_{X_1}$ be included in that of $\mathbb{P}_{X_0}$, and thus is related to the common support assumption in Assumption 1. If both $\mathbb{P}_{X_1}$ and $\mathbb{P}_{X_0}$ have density functions p_{X_1} and p_{X_0}, respectively, with respect to a common reference measure (e.g., the Lebesgue measure in the case of $\mathcal{X} \subset \mathbb{R}^d$), then the Radon-Nikodym derivative becomes the density ratio or the importance weight function $g(x) = p_{X_1}(x)/p_{X_0}(x)$. Thus, the square-integrability of g requires, intuitively, that p_{X_1} should not be very different from p_{X_0}.
- In the condition **(iv)**, we assume $n = m$ for simplicity of presentation.

Before presenting the result, we introduce a function $\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$\theta(\mathbf{x}, \tilde{\mathbf{x}}) := \iint \ell(\mathbf{y}, \tilde{\mathbf{y}}) \, \mathrm{d}\mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbb{P}_{Y_0|X_0}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}). \tag{28}$$

This function appears in the proof of consistency, and also is needed to derive convergence rates in Section 4.2. Note that the assumption in **(i)** that ℓ being bounded implies that $\theta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$; this property is used in the proof of consistency.

Theorem 8 below shows the consistency of the CME estimator (18). The proof can be found in Appendix E.2.

Theorem 8 (Consistency) *Suppose that Assumption 2 is satisfied. Let $\hat{\mu}_{Y\langle 0|1\rangle}$ be the estimator defined in (18) with a regularization constant $\varepsilon_n > 0$. Then if $\varepsilon_n \to 0$ and $n^{1/2}\varepsilon_n \to \infty$ as $n \to \infty$, we have*

$$\left\| \hat{\mu}_{Y\langle 0|1\rangle} - \mu_{Y\langle 0|1\rangle} \right\|_{\mathscr{F}} \to 0$$

in probability as $n \to \infty$.

Remark 9 As discussed, the form of the CME estimator (18) is the same as that of the kernel sum rule, and Fukumizu et al. (2013, Theorem 8) proves its consistency. Unlike ours, however, Fukumizu et al. (2013) assume that the function θ in (28) belongs to the tensor-product RKHS $\mathscr{H} \otimes \mathscr{H}$, which is a rather strong assumption for proving just the consistency. For instance, if $\mathscr{H}$ is the RKHS of the Gaussian kernel, then this assumption requires that θ be infinitely differentiable. The theoretical contribution of our analysis is in removing this condition.

Recall that by Lemma 4 we have $\mu_{Y\langle 0|1\rangle} = \mu_{Y_0^*|T=1}$ under the conditional exogeneity condition (Assumption 1). Thus, Theorem 8 implies that the CME estimator is consistent in estimating $\mu_{Y_0^*|T=1}$, as summarized in the following corollary. This justifies the use of the CME estimator in dealing with counterfactual questions, as will be described in Section 5.

Corollary 10 *Suppose that Assumptions 1 and 2 are satisfied. Let $\hat{\mu}_{Y\langle 0|1\rangle}$ be the estimator defined in (18) with a regularization constant $\varepsilon_n > 0$. Then, if $\varepsilon_n \to 0$ and $n^{1/2}\varepsilon_n \to \infty$ as $n \to \infty$, we have*

$$\left\| \hat{\mu}_{Y\langle 0|1\rangle} - \mu_{Y_0^*|T=1} \right\|_{\mathscr{F}} \to 0$$

in probability as $n \to \infty$.

4.2 Convergence Rates

Next, we present a result on the convergence rate of the CME estimator (18). This result is obtained based on certain smoothness assumptions on the Radon-Nikodym derivative $g = \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$ and the function θ defined in (28). To state these assumptions, we need to introduce the following concepts, details of which can be found in Appendix D.1.

In the sequel, $I \subset \mathbb{N}$ denotes a set of indices, which is a finite set or an infinite set depending on whether the RKHS $\mathscr{H}$ is finite dimensional (e.g., if k is a linear or polynomial kernel) or infinite dimensional (e.g., if k is a Gaussian or Matérn kernel). We define an integral operator $T : L_2(\mathbb{P}_{X_0}) \to L_2(\mathbb{P}_{X_0})$ by

$$Tf := \int k(\cdot, \mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}), \quad f \in L_2(\mathbb{P}_{X_0}).$$

Intuitively, the output function Tf is a smoother version of the input function f, as Tf can be seen as a convolution between f and the kernel k.

Under Assumption 2 **(i)** and **(ii)**, there exist at most countable families of functions $(e_i)_{i\in I} \subset \mathscr{H}$ and the associated positive constants $(\mu_i)_{i\in I} \subset (0, \infty)$ such that i) $\mu_1 \geq \mu_2 \geq \cdots > 0$, that ii) $(\mu_i^{1/2} e_i)_{i\in I}$ is an orthonormal basis (ONB) in $\mathscr{H}$, that iii) $(e_i)_{i\in I}$ is an ONB in $L_2(\mathbb{P}_{X_0})$, and that iv) the integral operator can be written as

$$Tf = \sum_{i\in I} \mu_i \langle f, e_i \rangle_{L_2(\mathbb{P}_{X_0})} e_i,$$

with convergence in $L_2(\mathbb{P}_{X_0})$; see Lemmas 16 and 18 in Appendix D.1. In other words, the pairs $(\mu_i, e_i)_{i\in I}$ are eigenvalues and eigenfunctions of the integral operator: $Te_i = \mu_i e_i$ for $i \in I$. Based on this eigendecomposition, one can define a *power* of the integral operator T: for a constant $\alpha \geq 0$, the α-th power of T is defined as

$$T^\alpha f := \sum_{i\in I} \mu_i^\alpha \langle f, e_i \rangle_{L_2(\mathbb{P}_{X_0})} e_i, \quad f \in L_2(\mathbb{P}_{X_0}).$$

We now make the following assumption about the smoothness of the Radon-Nikodym derivative $g = \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$, where $\mathrm{Range}(T^\alpha)$ denotes the range or image of T^α. This way of stating a smoothness condition is common in learning theory for kernel methods, e.g., Caponnetto and Vito (2007); Smale and Zhou (2007); Fukumizu et al. (2013).

Assumption 3 *There exists a constant* $0 \leq \alpha \leq 1$ *such that the Radon-Nikodym derivative* $g = \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$ *satisfies* $g \in \mathrm{Range}(T^\alpha)$.

Remark 11

- Assumption 3 quantifies the smoothness of the Radon-Nikodym derivative $g = \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$ by the constant $0 \leq \alpha \leq 1$. That is, g is smoother if α is close to 1 and less smooth if α is close to 0. Since we have $\mathrm{d}\mathbb{P}_{X_1}(x) = g(x)\,\mathrm{d}\mathbb{P}_{X_0}(x)$, a larger α may therefore be understood as that $\mathbb{P}_{X_0}$ and $\mathbb{P}_{X_1}$ are more similar. This interpretation can be obtained as follows.

- The assumption implies that there exists a square-integrable function $f \in L_2(\mathbb{P}_{X_0})$ that $g = T^\alpha f$. As mentioned, T acts as a smoother, outputting a smoothed version

Tf of an input function f. Similarly, its power T^α acts as a smoother, but now α determines the degree of smoothness of the output function. For $\alpha = 0$, T^α is just the identity map, and there is no effect of smoothing. As α increases, the degree of smoothness increases. In fact, Steinwart and Scovel (2012, Theorem 4.6) shows that $\mathrm{Range}(T^\alpha)$ for $0 < \alpha \leq 1/2$ is equal to an interpolation space between $L_2(\mathbb{P}_{X_0})$ and the RKHS $\mathscr{H}$ as a set of functions. In particular, we have $\mathrm{Range}(T^\alpha) = \mathscr{H}$ for $\alpha = 1/2$, and thus the assumption implies $g \in \mathscr{H}$. Thus, the case $\alpha > 1/2$ is that g is smoother than the least smooth functions in $\mathscr{H}$.

We next state a smoothness assumption about the function $\theta(\mathbf{x}, \mathbf{x}')$ defined in (28). To simplify the presentation, let $\mathcal{X}^2 := \mathcal{X} \times \mathcal{X}$. Define a kernel on $\mathcal{X}^2$ as the product kernel $k_{\mathrm{prod}} : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ such that

$$k_{\mathrm{prod}}\left((\mathbf{x}_1, \mathbf{x}_2), (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2)\right) := k(\mathbf{x}_1, \tilde{\mathbf{x}}_1) k(\mathbf{x}_2, \tilde{\mathbf{x}}_2), \quad (\mathbf{x}_1, \mathbf{x}_2), (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2) \in \mathcal{X}^2.$$

We then define an integral operator $T_{\mathrm{prod}} : L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}) \to L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$ by

$$T_{\mathrm{prod}}\eta := \int k_{\mathrm{prod}}\left(\cdot, (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2)\right) \eta\left((\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2)\right) \mathrm{d}\left(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}\right)\left((\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2)\right), \quad \eta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}).$$

By Assumption 2 **(i)** and **(ii)**, this can be written in terms of the eigensystem $(\mu_i, e_i)_{i \in I}$ as

$$T_{\mathrm{prod}}\eta = \sum_{i,j \in I} \mu_i \mu_j \left\langle \eta, e_i \otimes e_j \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} e_i \otimes e_j,$$

where $e_i \otimes e_j : \mathcal{X}^2 \to \mathbb{R}$ denotes the tensor product of e_i and e_j, and the convergence is in $L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$; see Lemma 19 in Appendix D. That is, each $e_i \otimes e_j$ is an eigenfunction of T_{prod} with the corresponding eigenvalue $\mu_i \mu_j$. The β-th power of T_{prod} for $0 \leq \beta \leq 1$ is then defined as

$$T_{\mathrm{prod}}^\beta \eta = \sum_{i,j \in I} (\mu_i \mu_j)^\beta \left\langle \eta, e_i \otimes e_j \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} e_i \otimes e_j. \tag{29}$$

Similar to Assumption 3, we make the following smoothness assumption for the function $\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined in (28), based on the range of the power T_{prod}^β.

Assumption 4 *There exists a constant $0 \leq \beta \leq 1$ such that the function θ defined in (28) satisfies $\theta \in \mathrm{Range}(T_{\mathrm{prod}}^\beta)$.*

Remark 12 As for Assumption 3, we can interpret Assumption 4 as quantifying the smoothness of θ by the constant β. That is, larger β implies that θ is smoother. Note that θ can be written as $\theta(\mathbf{x}, \mathbf{x}') = \left\langle \mu_{Y_0|X_0=\mathbf{x}}, \mu_{Y_0|X_0=\mathbf{x}'} \right\rangle_{\mathscr{H}}$, where $\mu_{Y_0|X_0=\mathbf{x}} := \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})$ is the kernel mean of $\mathbb{P}_{Y_0|X_0}(\cdot|\mathbf{x})$. Therefore, θ is smooth if the mapping $\mathbf{x} \to \mu_{Y_0|X_0=\mathbf{x}}$ is smooth. Thus, β may be interpreted as quantifying the smoothness of this mapping.

We are now ready to state Theorem 13 below, which establishes the convergence rate of the CME estimator. The rate is given in terms of the constants α and β introduced in the above assumptions. The proof is given in Appendix E.3.

Theorem 13 (Convergence rates) *Suppose that Assumptions 2, 3 and 4 hold with* $\alpha + \beta \leq 1$. *Let* $\hat{\mu}_{Y\langle 0|1\rangle}$ *be the estimator defined in* (18) *with a regularization constant* $\varepsilon_n > 0$. *Let* $c > 0$ *be an arbitrary constant, and set* $\varepsilon_n = cn^{-1/(1+\beta+\max(1-\alpha,\alpha))}$. *Then we have*

$$\left\|\hat{\mu}_{Y\langle 0|1\rangle} - \mu_{Y\langle 0|1\rangle}\right\|_{\mathscr{F}} = O_p\left(n^{-(\alpha+\beta)/2(1+\beta+\max(1-\alpha,\alpha))}\right) \quad (n \to \infty).$$

Remark 14 Let us interpret the rate of Theorem 13.

- The exponent $(\alpha+\beta)/2(1+\beta+\max(1-\alpha,\alpha))$ in the rate is smaller than $1/2$ for any α and β, and thus the rate is always slower than the parametric rate $n^{-1/2}$. For instance, the rate becomes $n^{-1/4}$ if $\alpha = \beta = 1/2$. This is due to the CME estimator being nonparametric, as for other nonparametric statistical estimators in general (Tsybakov, 2008). We are not aware of, however, whether the obtained rate is minimax optimal. We leave this question for future research.

- An important interpretation of the rate is as follows: 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, then the CME estimator converges reasonably fast. For instance, the rate becomes $n^{-1/6}$ if $\alpha = 0$ and $\beta = 1$, and $n^{-1/4}$ if $\alpha = 1$ and $\beta = 0$ (recall that α and β quantify the smoothness of g and θ, respectively). 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 for the CME estimator if the relationship between X_0 and Y_0 is smooth (and vice versa).

As for Corollary 10, we obtain the following corollary from Theorem 13.

Corollary 15 *Suppose that Assumptions 1, 2, 3 and 4 hold with* $\alpha + \beta \leq 1$. *Let* $\hat{\mu}_{Y\langle 0|1\rangle}$ *be the estimator defined in* (18) *with a regularization constant* $\varepsilon_n > 0$. *Let* $c > 0$ *be an arbitrary constant, and set* $\varepsilon_n = cn^{-1/(1+\beta+\max(1-\alpha,\alpha))}$. *Then we have*

$$\left\|\hat{\mu}_{Y\langle 0|1\rangle} - \mu_{Y_0^*|T=1}\right\|_{\mathscr{F}} = O_p\left(n^{-(\alpha+\beta)/2(1+\beta+\max(1-\alpha,\alpha))}\right) \quad (n \to \infty).$$

5. Applications to Sampling and Testing

In this section, we discuss important applications of the proposed framework.

5.1 Sampling from Counterfactual Distributions

While a CME estimate can be seen as a weighted sample $(\beta_i, \mathbf{y}_i)_{i=1}^n$, the coefficients $\beta_1, \ldots, \beta_n$ may in general include negative values; thus it is not straightforward to interpret them as importance weights. If one can generate sample points $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ from the CME estimate, then these unweighted points may be more useful for an analyst. For instance, we might use them for the purpose of visualization (*e.g.*, scatter plot and histogram). Moreover, as will be described below, such unweighted points can be straightforwardly used for testing hypotheses regarding distributional treatment effects.

We propose a method for sampling from the counterfactual distribution based on the CME estimator and the kernel herding algorithm (Chen et al., 2010; Kanagawa et al., 2016; Kajihara et al., 2018). The method is summarized in Algorithm 1, which generates sample

Algorithm 1 Sampling from a counterfactual mean embedding estimate

1: **Input**: A CME estimate $\hat{\mu}_{Y\langle 0|1\rangle} = \sum_{i=1}^{n} \beta_i \ell(\mathbf{y}_i, \cdot)$ with $(\beta_i, \mathbf{y}_i)_{i=1}^{n} \subset \mathbb{R} \times \mathcal{Y}$ and kernel $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$; the number $m \in \mathbb{N}$ of sample points to generate.
2: Compute $\tilde{\mathbf{y}}_1 := \arg\max_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^{n} \beta_i \ell(\mathbf{y}_i, \mathbf{y})$.
3: **for** $t = 2$ to m **do**
4: Compute $\tilde{\mathbf{y}}_t := \arg\max_{\mathbf{y} \in \mathcal{Y}} \sum_{i=1}^{n} \beta_i \ell(\mathbf{y}_i, \mathbf{y}) - \frac{1}{t} \sum_{i=1}^{t-1} \ell(\tilde{\mathbf{y}}_i, \mathbf{y})$.
5: **end for**
6: **Output**: $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$.

points $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ from $\hat{\mu}_{Y\langle 0|1\rangle}$ in (18). When the kernel ℓ is shift-invariant (*e.g.*, Gaussian), the procedure in Algorithm 1 to generate $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_t$ for $t = 1, \ldots, m \in \mathbb{N}$ is equivalent to the greedy minimization of the RKHS distance between the CME estimate $\hat{\mu}_{Y\langle 0|1\rangle}$ and the empirical kernel mean $\frac{1}{t} \sum_{i=1}^{t} \ell(\tilde{\mathbf{y}}_i, \cdot)$:

$$\left\| \hat{\mu}_{Y\langle 0|1\rangle} - \frac{1}{t} \sum_{i=1}^{t} \ell(\tilde{\mathbf{y}}_i, \cdot) \right\|_{\mathscr{F}} = \sup_{\|f\|_{\mathscr{F}} \leq 1} \left| \sum_{i=1}^{n} \beta_i f(\mathbf{y}_i) - \frac{1}{t} \sum_{j=1}^{t} f(\tilde{\mathbf{y}}_j) \right|. \tag{30}$$

See Chen et al. (2010) for details. In other words, the points $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ are those greedily minimizing the worst case error to the weighted points $(\beta_i, \mathbf{y}_i)_{i=1}^{n}$ in the unit ball of the RKHS $\mathscr{F}$; thus, this algorithm is a greedy variant of Quasi Monte Carlo methods (Dick et al., 2013). Notice that therefore these points are, of course, not independent to each other. The convergence rate $O(n^{-1/2})$ is guaranteed for $\frac{1}{t} \sum_{i=1}^{t} \ell(\cdot, \tilde{\mathbf{y}}_i)$ (Bach et al., 2012), which may hold even when the optimization problem (30) is solved approximately (Lacoste-Julien et al., 2015; Kanagawa et al., 2016).

Lastly, to obtain high dimensional samples, *e.g.*, images, from the counterfactual distribution, one can train deep generative models using MMD-GAN (Li et al., 2015; Dziugaite et al., 2015; Sutherland et al., 2017; Li et al., 2017) based the CME estimate. We defer this promising application to future work.

5.2 Counterfactual Inference as Two-sample Testing

One can also identify distributional treatment effects by formulating the problem as that of hypothesis testing, or more specifically, two-sample testing. To describe this, we assume that we are given data $(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^{N}$, which are i.i.d. with random variables (X, T, Y) with $Y = Y_0^* \mathbb{1}(T = 0) + Y_1^* \mathbb{1}(T = 1)$ being the observed outcome and Y_0^*, Y_1^* being the potential outcomes. Let $n := \sum_{i=1}^{N} (1 - t_i)$ and $m := \sum_{i=1}^{N} t_i$. See Section 3.5 for the notation.

Distributional treatment effects. Here we are interested in testing whether $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$ are equal or not (see Section 3.1.1). The null hypothesis H_0 and the alternative hypothesis H_1 are thus defined as

$$H_0 : \mathbb{P}_{Y_0^*} = \mathbb{P}_{Y_1^*}, \quad H_1 : \mathbb{P}_{Y_0^*} \neq \mathbb{P}_{Y_1^*},$$

As a test statistic, we propose to use an estimate $\widehat{\mathrm{KTE}}(Y_0^*, Y_1^*, \mathscr{F})$ of the kernel treatment effect, $\mathrm{KTE}(Y_0^*, Y_1^*, \mathscr{F}) = \|\mu_{Y_0^*} - \mu_{Y_1^*}\|_{\mathscr{F}}^2$, introduced in Section 3.5.1. This esti-

mate $\widehat{\mathrm{KTE}}(Y_0^*, Y_1^*, \mathscr{F})$, computed from the data $(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^N$, can either be the biased one, $\widehat{\mathrm{KTE}}_b(Y_0^*, Y_1^*, \mathscr{F})$, defined in (22), or the unbiased one, $\widehat{\mathrm{KTE}}_u(Y_0^*, Y_1^*, \mathscr{F})$, in (23).

To decide a critical region, we need the distribution of the test statistic under the null hypothesis H_0. One way to approximate this distribution is to use a bootstrap procedure (Efron and Tibshirani, 1993), as follows. Let $B \in \mathbb{N}$ be the number of bootstrap samples. For each $b = 1, \ldots, B$, we randomly permute the indices $1, \ldots, N$ to, say, $\pi_b(1), \ldots, \pi_b(N) \subset \{1, \ldots, N\}$. Then we compute the test statistic $\eta_b := \widehat{\mathrm{KTE}}(Y_0^*, Y_1^*, \mathscr{F})$ based on the permuted data $(\mathbf{x}_i, t_{\pi_b(i)}, \mathbf{y}_i)_{i=1}^N$. We then approximate the null distribution by the histogram of $\eta_1, \ldots, \eta_B$, and determine a critical tail region for rejecting the null hypothesis (*e.g.*, with significance level $\alpha = 0.05$).

Distributional effects of the covariate distributions. Here, we are interested in whether the two distributions $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0)$ and $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$ are equal or not (see Section 3.1.3). If they are different, there is a distributional effect on the outcomes arising from the difference in covariate distributions. The identification of such an effect can be phrased as a hypothesis test with the null and alternative hypotheses being

$$H_0 : \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0) = \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1), \quad H_1 : \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0) \neq \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1), \tag{31}$$

To describe the approach, we rearrange the data $(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^N$ so that $t_i = 0$ for $i = 1, \ldots, n$ and $t_i = 1$ for $i = n+1, \ldots, n+m = N$. Note that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are a sample from $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0)$, while the distribution $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$ is counterfactual and we do not have a sample from it. However, we can estimate the kernel mean of $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$ by the CME estimator (18) using the data $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ and $(\tilde{\mathbf{x}}_j)_{j=1}^m := (\mathbf{x}_{n+j})_{j=1}^m$ under the conditional exogeneity assumption, and let $\hat{\mu}_{\langle 0|1 \rangle}$ be the resulting estimate. We then apply kernel herding to $\hat{\mu}_{\langle 0|1 \rangle}$ for obtaining sample points $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ approximating $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$, as described in Algorithm 1.

We can then apply any method for two-sample test, *e.g.*, those in Gretton et al. (2012), to the two samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ to test the hypotheses (31). We note that this is a rather heuristic approach, since $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ are not drawn from $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$, but generated deterministically so as to approximate $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$. We leave a further theoretical study regarding the validity of this approach for future research.

Finally, one can develop a testing procedure for identifying distributional treatment effects on the treated (see Section 3.1.2) in the same way as described here, and thus we omit the explanation.

5.3 Discussion

Here we discuss how the above sampling and testing procedures may be used in practice. Suppose that an analyst is interested in the distributional effects of the difference in covariate distributions, *i.e.*, the hypotheses in (31), and that the null hypothesis has been rejected as a result of applying the above testing procedure. This suggests the existence of a difference in the two distributions, $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0)$ and $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$. This is usually *not* the end of the analysis, but is the *starting point* of a further exploratory analysis. There are several ways to proceed, to understand how the two outcome distributions differ.

One way is to use the samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_m$ (the latter being a counterfactual sample generated from Algorithm 1) approximating the distributions $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0)$ and

$\mathbb{P}_{Y_0^*|T}(\cdot \mid 1)$, respectively. The analyst can use any available statistical method for finding the source of the difference in the two distributions. For instance, she may compute summary statistics of both samples (e.g., mean, variance, etc.) and compare them. It is also possible to just plot both samples, or to estimate the densities, to visualize the difference (as we demonstrate in Section 7.1). Another useful method in this context is the approach of Jitkrittum et al. (2016) and their follow-up works, which returns interpretable features for explaining the difference in the two samples, such as the sample locations on which (smoothed version) of the two density values differ substantially.

Another important point for discussion is the use of non-characteristic kernels. If the kernel ℓ is not characteristic, such as polynomial kernels, rejecting the null hypothesis implies that there exist a *certain kind* of difference in the two distributions. For instance, if the kernel ℓ is a polynomial kernel of order 2, then rejecting the null hypothesis implies that there exists a difference in the mean or in the variance of the two outcome distributions. In this sense, if one is interested in the existence of a specific difference in the outcome distributions (such as the mean and variance), non-characteristic kernels may be more useful.

6. Application to Off-Policy Evaluation (OPE)

We describe here how our approach can be applied to the *off-policy evaluation* task (OPE), *e.g.*, Dudík et al. (2011), which aims at evaluating the performance of a given target policy of deciding a certain action given a context. The performance is measured in terms of the resulting rewards. For instance, consider a recommendation system, where an action is a list of items to be recommended to a user, and a policy determines which action to take, given the features of the user. There will be a positive reward if the user clicks or buys one of the recommended items, and no reward otherwise. The goal of OPE is to estimate how a given policy would work, without actually implementing the policy. Instead, the evaluation is to be done relying only on logged (or historical) data obtained from a possibly unknown initial policy, which is different from the target policy. This task is important when actually implementing a new policy is expensive or difficult, with wide applications including ad placement, recommendation systems, and health care.

We first describe the OPE problem more formally in Section 6.1. We then interpret it with the potential outcome framework and formulate the OPE problem as an estimation of a counterfactual distribution in Section 6.2. Finally, we present a concrete algorithm in Section 6.3.

6.1 Problem Description

Formally, the OPE task may be defined as follows. Let $\mathcal{U}$ be a space of context features, $\mathcal{A}$ be a space of actions and $\mathcal{R}$ be a space of rewards. For instance, in the case of recommendation systems, each $\mathbf{u} \in \mathcal{U}$ represents a user's features, $\mathbf{a} \in \mathcal{A}$ a recommendation (*e.g.*, a list of items), and $r \in \mathcal{R}$ the number of clicks on the recommendation. A policy $\pi(\mathbf{a}|\mathbf{u})$ is a conditional distribution on the action space $\mathcal{A}$ given context features $\mathbf{u} \in \mathcal{U}$. In a recommendation system, $\pi(\mathbf{a}|\mathbf{u})$ determines the probability of providing a recommendation $\mathbf{a}$ to a user having features $\mathbf{u}$.

Assume that tuples $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n \subset \mathcal{U} \times \mathcal{A} \times \mathcal{R}$ of context features $\mathbf{u}_i \in \mathcal{U}$, action $\mathbf{a}_i \in \mathcal{A}$ and reward $r_i \in \mathcal{R}$ are available as *logged (or historical) data*. We assume that they

were independently generated from a joint distribution $\mathbb{P}_0(\mathbf{u}, \mathbf{a}, r) := q_0(\mathbf{u})\pi_0(\mathbf{a}|\mathbf{u})\mathbb{P}_0(r|\mathbf{u}, \mathbf{a})$ in the data collection phase, where $q_0(\mathbf{u})$ is a marginal distribution on $\mathcal{U}$, $\pi_0(\mathbf{a}|\mathbf{u})$ is an *initial (or logging/behavior) policy*, and $\mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a})$ is a conditional distribution of a reward $r \in \mathcal{R}$ given $(\mathbf{u}, \mathbf{a}) \in \mathcal{U} \times \mathcal{A}$. In a recommendation system, for instance, $\mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a})$ describes whether a user with features $\mathbf{u}$ who has been recommended a list of items $\mathbf{a}$ would choose one of the items. As such, it is typically unknown a priori. Similarly, $q_0(\mathbf{u})$ and $\pi_0(\mathbf{a}|\mathbf{u})$ may be unknown in practice, if $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ are given as historical data.

Let $\pi_*(\mathbf{a}|\mathbf{u})$ be another conditional distribution of actions $\mathbf{a} \in \mathcal{A}$ given context features $\mathbf{u} \in \mathcal{U}$, which represents the *target policy* that one wants to evaluate. By design, the target policy is known and sampling from it is possible. Let $q_*(\mathbf{u})$ be a probability distribution on $\mathcal{U}$, which represents the distribution of context features under the target environment (*e.g.*, the distribution of user features when a recommendation system is deployed). In the standard OPE setting, it is typically assumed that $q_0(\mathbf{u}) = q_*(\mathbf{u})$, i.e., the historical and target environments are the same; but in general these can be different, $q_0(\mathbf{u}) \neq q_*(\mathbf{u})$. The latter situation is not uncommon in practice and has been recently studied by Uehara et al. (2020). Finally, let $\mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a})$ be the conditional distribution of a reward r given context features $\mathbf{u}$ and action $\mathbf{a}$ under the target environment. We assume that this remains the same as in the data collection phase, *i.e.*,

$$\mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a}) = \mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a}).$$

This assumption may be understood as the *policy invariance* assumption commonly made in the econometric policy evaluation literature, *e.g.*, Heckman and Vytlacil (2007).[5]

The task of off-policy evaluation is then to estimate the expected reward under the target environment:

$$R_* := \int_{\mathcal{U}\times\mathcal{A}} \int_{\mathcal{R}} r \, \mathrm{d}\mathbb{P}_*(r|\mathbf{u}, \mathbf{a}) \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a}) = \int_{\mathcal{U}\times\mathcal{A}} \int_{\mathcal{R}} r \, \mathrm{d}\mathbb{P}_0(r|\mathbf{u}, \mathbf{a}) \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a}), \tag{32}$$

where the identity follows from the assumption $\mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a}) = \mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a})$, and $\pi_*(\mathbf{u}, \mathbf{a}) := \pi_*(\mathbf{a}|\mathbf{u})q_*(\mathbf{u})$ is the joint distribution on $\mathcal{U} \times \mathcal{A}$ given by $\pi_*(\mathbf{a}|\mathbf{u})$ and $q_*(\mathbf{u})$. This estimation is to be done using logged data $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ and the target policy $\pi_*(\mathbf{u}|\mathbf{a})$.

6.2 OPE as Counterfactual Inference

We explain below how our CME estimator (18) can be applied to the OPE task. To this end, we consider the marginal distribution of a reward under the target environment:

$$\mathbb{P}_*(r) := \int_{\mathcal{U}\times\mathcal{A}} \mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a}) \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a}) = \int_{\mathcal{U}\times\mathcal{A}} \mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a}) \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a}). \tag{33}$$

Note that the expected reward (32) is the mean of this distribution. We first show that the distribution $\mathbb{P}_*(r)$ can be interpreted as a counterfactual distribution, by formulating the OPE task using the potential outcome framework[6] in Section 3.

5. More precisely, this assumption may be identified with the policy invariance assumptions PI-1 and PI-2 in Section 2.2 of Heckman and Vytlacil (2007), where s and ω there correspond to $\mathbf{a}$ and $\mathbf{u}$ in our setting, respectively, and tuple (a, b, τ) there essentially corresponds to a policy in our setting.
6. Our formulation of the OPE task using the potential outcome framework is different from the existing formulation, *e.g.*, Kallus and Zhou (2018), where each action $\mathbf{a} \in \mathcal{A}$ is defined as a treatment, and for

Random variables. Consider a hypothetical subject in population. This subject is associated with covariates $X := (U, A)$, where $U \in \mathcal{U}$ is context features and $A \in \mathcal{A}$ is an action taken. As such, we define the covariate space as $\mathcal{X} := \mathcal{U} \times \mathcal{A}$, the product of the context feature space $\mathcal{U}$ and action space $\mathcal{A}$. In a recommendation system, for instance, U is the user features and A is a recommended list of items. We define two treatments 0 and 1 as exposing the subject to *the environment during the data collection phase* and *that during the evaluation phase*, respectively; and the associated potential outcomes Y_0^* and Y_1^* as the *rewards* under the respective treatments 0 and 1. Let $T \in \{0, 1\}$ be a treatment indicator.

In a recommendation system, for example, an environment may refer to the situation where a user is about to choose an item, such as the calendar year when this takes place. For instance, treatment 0 may refer to the environment in the year 2000, and treatment 1 the environment in year 2020. Consider a user with the features U and the recommended items A, and suppose that A consists of items which were popular in 2000 but are outdated in 2020. Then this user may have chosen an item from A if it were in 2000, but may not choose any item in 2020, *i.e.*, we may have $Y_0^* \neq Y_1^*$.

For ease of understanding, consider a finite population of N subjects with

$$(\mathbf{y}_{i0}^*, \mathbf{y}_{i1}^*, \mathbf{x}_i, t_i)_{i=1}^N \tag{34}$$

being i.i.d. realizations of the random variables (Y_0^*, Y_1^*, X, T). That is, the i-th subject is associated with covariates $\mathbf{x}_i := (\mathbf{u}_i, \mathbf{a}_i)$ consisting of context features $\mathbf{u}_i$ and action $\mathbf{a}_i$. The treatment assignment $t_i \in \{0, 1\}$ indicates which environment the i-th subject is exposed to. Thus, the potential outcomes $\mathbf{y}_{i0}^*$ and $\mathbf{y}_{i1}^*$ are the rewards from the i-th subject (associated with $\mathbf{x}_i := (\mathbf{u}_i, \mathbf{a}_i)$) that would have been observed if she was exposed to the environment during the data collection phase (treatment 0) and that during the evaluation phase (treatment 1), respectively.

Distributions of the potential outcomes. The distributions of the potential outcomes Y_0^*, Y_1^* are defined via their conditional distributions given $X = (U, A)$, *i.e.*, $\mathbb{P}_{Y_0^*|X}(\mathbf{y}|\mathbf{x})$ and $\mathbb{P}_{Y_1^*|X}(\mathbf{y}|\mathbf{x})$ where $\mathbf{y} = r$ and $\mathbf{x} = (\mathbf{u}, \mathbf{a})$. These are the conditional distributions of rewards $r \in \mathcal{R}$ given context features $\mathbf{u} \in \mathcal{U}$ and action $\mathbf{a} \in \mathcal{A}$, under the environment during the data collection phase (treatment 0) and that during the evaluation phase (treatment 1), respectively:

$$\mathbb{P}_{Y_0^*|X}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_0(r|\mathbf{u}, \mathbf{a}), \quad \mathbb{P}_{Y_1^*|X}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_*(r|\mathbf{u}, \mathbf{a}) \quad (\mathbf{y} = r, \quad \mathbf{x} = (\mathbf{u}, \mathbf{a})).$$

Note that our assumption $\mathbb{P}_0(r|\mathbf{u}, \mathbf{a}) = \mathbb{P}_*(r|\mathbf{u}, \mathbf{a})$ implies that

$$\mathbb{P}_{Y_0^*|X}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_{Y_1^*|X}(\mathbf{y}|\mathbf{x}), \tag{35}$$

each action $\mathbf{a}$ there is a corresponding potential outcome $Y_{\mathbf{a}}^*$. In our formulation, on the other hand, an action $\mathbf{a}$ taken for the subject is defined as a part of covariates $\mathbf{x} = (\mathbf{u}, \mathbf{a})$, and binary treatments, 0 and 1, are considered, each of which is defined as exposing the subject to a certain environment. Our formulation enables us to interpret the OPE task as counterfactual inference of changing the covariate distribution, so that our CME estimator can be naturally applied. Thus, our motivation of introducing this formulation is rather pragmatic, and we do not argue whether it is more reasonable than the existing one. One benefit may exist, however: In our formulation, we explicitly model the assumption on the conditional distributions of rewards being the same for the data collection and evaluation phases via the potential outcome notation (35), while this assumption is implicitly made in the existing formulation. This explicit statement of the assumption helps a researcher to understand when the OPE may be justified.

i.e., the conditional distributions of the potential outcomes (rewards) Y_0^* and Y_1^* are the same, given the covariates $X = \mathbf{x} := (\mathbf{u}, \mathbf{a})$.

For instance, consider a recommendation system with a finite population (34) with the i-th user equipped with covariates $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{a}_i)$ consisting of features $\mathbf{u}_i$ and recommended items $\mathbf{a}_i$. The identity (35) then implies that, for the i-th user, the distributions of the potential outcomes (rewards) $\mathbf{y}_{i0}^*$ and $\mathbf{y}_{i1}^*$ are the same. This means that this user should have the same stochastic behavior in choosing (or not choosing) an item from the recommended ones during the data collection (treatment 0) and evaluation (treatment 1) phases. In other words, the environmental factors that affect the user behavior should be the same for the data collection and evaluation phases. This excludes, for instance, the above example where the data collection phase is in year 2000 and the evaluation phase is in 2020, in which case user preferences are different.

Distributions of covariates. As in Section 3.3.2, we interpret here a policy as specifying a distribution on the covariate space $\mathcal{X} = \mathcal{U} \times \mathcal{R}$. More specifically, consider the conditioning $T = 0$ or $T = 1$, which imply that the hypothetical subject is exposed to the environment of the data collection phase ($T = 0$) or to that of the evaluation phase ($T = 1$). Then the corresponding distributions of the covariates $X_0 = X|(T = 0)$ and $X_1 = X|T = 1$ are given by the logging policy $\pi_0(\mathbf{u}|\mathbf{a})$ and the target policy $\pi_*(\mathbf{u}|\mathbf{a})$, respectively:

$$\mathbb{P}_{X_0}(\mathbf{x}) = \pi_0(\mathbf{a}|\mathbf{u})q_0(\mathbf{u}), \quad \mathbb{P}_{X_1}(\mathbf{x}) = \pi_*(\mathbf{a}|\mathbf{u})q_*(\mathbf{u}) \quad (\mathbf{x} = (\mathbf{u}, \mathbf{a}) \in \mathcal{U} \times \mathcal{A}). \tag{36}$$

To describe this, consider the finite population (34), and assume that for the i-th user the treatment indicator is $t_i = 0$, which implies that her data are given in the data collection phase. In this case, her covariates $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{a}_i)$ are generated according to the joint distribution $\pi_0(\mathbf{a}|\mathbf{u})q_0(\mathbf{u})$ involving the logging policy $\pi_0(\mathbf{a}|\mathbf{u})$. On the other hand, if $t_i = 1$, her covariates $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{a}_i)$ are generated from the joint distribution $\pi_*(\mathbf{a}|\mathbf{u})q_*(\mathbf{u})$ given by the target policy $\pi_*(\mathbf{a}|\mathbf{u})$. Note that in the OPE setting, if $t_i = 1$ we have access to neither $\mathbf{y}_{i0}^*$ nor $\mathbf{y}_{1i}^*$; note also that in this case this "user" may be imaginary, with covariates $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{a}_i)$ generated artificially.

Distributions of observed outcomes. The observed outcome (reward) from the hypothetical subject is defined as $Y = \mathbb{1}(T = 0)Y_0^* + \mathbb{1}(T = 1)Y_1^*$. Let $Y_0 = Y|(T = 0) = Y_0^*|(T = 0)$ and $Y_1 = Y|(T = 1) = Y_1^*|(T = 1)$ be the observed outcome Y conditioned on $T = 0$ or $T = 1$, respectively. Then Y_0 conditioned on $X_0 = X|(T = 0)$ and Y_1 conditioned on $X_1 = X|(T = 1)$ can be written in terms of the potential outcomes Y_0^* and Y_1^* as

$$Y_0|X_0 = Y_0^*|X, (T = 0), \quad Y_1|X_1 = Y_1^*|X, (T = 1).$$

Note that the potential outcomes Y_0^* and Y_1^* are independent to the treatment indicator T given the covariates X under the conditional exogeneity in Assumption 1. Therefore, $Y_0^*|X, (T = 0) = Y_0^*|X$ and $Y_1^*|X, (T = 1) = Y_1^*|X$.

Thus, we can identify the conditional distributions $\mathbb{P}_{Y_0|X_0}$ and $\mathbb{P}_{Y_1|X_1}$ as $\mathbb{P}_0(r|\mathbf{u}, \mathbf{a})$ and $\mathbb{P}_*(r|\mathbf{u}, \mathbf{a})$, respectively:

$$\mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_{Y_0^*|X}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_0(r|\mathbf{u}, \mathbf{a}), \quad \mathbb{P}_{Y_1|X_1}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_{Y_1^*|X}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_*(r|\mathbf{u}, \mathbf{a}), \tag{37}$$

where $\mathbf{y} = r$ and $\mathbf{x} = (\mathbf{u}, \mathbf{a})$. Therefore, the assumption $\mathbb{P}_0(r|\mathbf{u}, \mathbf{a}) = \mathbb{P}_*(r|\mathbf{u}, \mathbf{a})$ (or (35)) implies that

$$\mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_{Y_1|X_1}(\mathbf{y}|\mathbf{x}).$$

Reward distribution as a counterfactual distribution. Finally, the reward distribution (33) under the target environment can be written as a counterfactual distribution using the identities (36) and (37) (assuming the conditional exogeneity) as

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

with $\mathbf{y} = r$ and $\mathbf{x} = (\mathbf{u}, \mathbf{a})$. Thus, we can use the CME estimator (18) to estimate the kernel mean of this reward distribution, which is described below.

6.3 Off-Policy Evaluation by the CME Estimator

Let ℓ be a kernel on the outcome (reward) space $\mathcal{Y} = \mathcal{R}$ with $\mathscr{F}$ its RKHS. Then the mean embedding of the reward distribution under the target environment is defined as

$$\mu_{\mathbb{P}_*} := \int \ell(\cdot, r)\,\mathrm{d}\mathbb{P}_*(r) \in \mathscr{F}. \tag{38}$$

Note that this becomes the expected reward (32) if we define ℓ as a linear kernel: $\ell(r, r') := rr'$, which we use in our experiments in Section 7.2. In principle, however, the use of other nonlinear kernels (in particular characteristic kernels) makes the mean embedding more informative, and this may be beneficial in assessing the effectiveness of the target policy.

Kernel on covariates. To use the CME estimator, we also need to define a kernel k on the covariate space $\mathcal{X} = \mathcal{U} \times \mathcal{A}$. To this end, we first define kernels $k_{\mathcal{U}}$ and $k_{\mathcal{A}}$ on the context feature space $\mathcal{U}$ and the action space $\mathcal{A}$, respectively. Then we can define k as the *product kernel* of $k_{\mathcal{U}}$ and $k_{\mathcal{A}}$: $k((\mathbf{u}, \mathbf{a}), (\mathbf{u}', \mathbf{a}')) := k_{\mathcal{U}}(\mathbf{u}, \mathbf{u}')k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}')$ for $(\mathbf{u}, \mathbf{a}), (\mathbf{u}', \mathbf{a}') \in \mathcal{U} \times \mathcal{A}$.

Joint sample. Recall that logged data $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ are i.i.d. with the joint distribution $\mathbb{P}_0(\mathbf{u}, \mathbf{a}, r) = \mathbb{P}_0(r|\mathbf{u}, \mathbf{a})\pi_0(\mathbf{a}|\mathbf{u})q_0(\mathbf{u})$, which is identified as $\mathbb{P}_{X_0Y_0}(\mathbf{x}, \mathbf{y}) = \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x})\mathbb{P}_{X_0}(\mathbf{x})$ for $\mathbf{x} = (\mathbf{u}, \mathbf{a})$ and $\mathbf{y} = r$ because of (36) and (37). Thus, by defining $\mathbf{x}_i := (\mathbf{u}_i, \mathbf{a}_i)$ and $\mathbf{y}_i := r_i$, we have an i.i.d. sample $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ from the joint distribution $\mathbb{P}_{X_0Y_0}(\mathbf{x}, \mathbf{y})$.

Covariate sample. We also need to express $\mathbb{P}_{X_1}$ in terms of a sample in the form $(\mathbf{x}'_j)_{j=1}^m$. As in (36), the covariate distribution $\mathbb{P}_{X_1}$ is the joint distribution given by the target policy $\pi_*(\mathbf{a}|\mathbf{u})$ and the marginal distribution $q_*(\mathbf{u})$ of context features. Thus, if we can sample from both $\pi_*(\mathbf{a}|\mathbf{u})$ and $q_*(\mathbf{u})$ (the former is typically possible because it is defined by the designer of the target policy, while the latter depends on the problem), then $(\mathbf{x}'_j)_{j=1}^m$ may be given by

$$\mathbf{x}'_j := (\mathbf{u}^*_j, \mathbf{a}^*_j), \quad \text{where} \quad \mathbf{u}^*_j \sim q_*(\mathbf{u}), \quad \mathbf{a}^*_j \sim \pi_*(\mathbf{a}|\mathbf{u}^*_j), \quad j = 1, \ldots, m. \tag{39}$$

In the particular case where $q_*(\mathbf{u}) = q_0(\mathbf{u})$, we can use the sample $(\mathbf{u}_i)_{i=1}^n$ in the logged data $\{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$, which are from $q_0(\mathbf{u})$, as a sample from $q_*(\mathbf{u})$: $\mathbf{u}^*_j := \mathbf{u}_j$ for $j = 1, \ldots, m := n$. Note that even if $q_*(\mathbf{u}) \neq q_0(\mathbf{u})$, we can use the CME estimator as long as we have a sample of context features $(\mathbf{u}^*_j)_{j=1}^m$ from the target environment, *i.e.*, the covariate shift setting of Uehara et al. (2020).

Algorithm. The resulting algorithm is described in Algorithm 2, which only requires matrix operations and thus is simple to implement. We note that the expected reward (32) under the target environment can be estimated as $\hat{\mu}_{\mathbb{P}_*(r)} = \sum_{i=1}^n \beta_i r_i$; this is obtained by setting ℓ as a linear kernel on $\mathbb{R}$.

Algorithm 2 Off-Policy Evaluation using the CME estimator (18)

Requirement: A kernel $k_{\mathcal{U}}$ on the context space $\mathcal{U}$, a kernel $k_{\mathcal{A}}$ on the action space $\mathcal{A}$, a kernel ℓ on the reward space $\mathcal{R}$, and a regularization constant $\varepsilon > 0$.
Input: Logged data $(\mathbf{u}_i, \mathbf{a}_i, r_i)_{i=1}^n$, a target policy $\pi_*(\mathbf{u}|\mathbf{a})$ and a sample of context features $(\mathbf{u}_j^*)_{j=1}^m$. (If $q_*(\mathbf{u}) = q_0(\mathbf{u})$, set $\mathbf{u}_j^* := \mathbf{u}_j, j = 1, \ldots, m := n$.)
for $j = 1$ **to** n **do**
  $\mathbf{a}_j^* \sim \pi_*(\mathbf{a} \,|\, \mathbf{u}_j^*)$
end for
Compute $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} := k_{\mathcal{U}}(\mathbf{u}_i, \mathbf{u}_j) k_{\mathcal{A}}(\mathbf{a}_i, \mathbf{a}_j)$, $i, j = 1, \ldots, n$.
Compute $\widetilde{\mathbf{K}} \in \mathbb{R}^{n \times m}$ with $\widetilde{\mathbf{K}}_{ij} := k_{\mathcal{U}}(\mathbf{u}_i, \mathbf{u}_j^*) k_{\mathcal{A}}(\mathbf{a}_i, \mathbf{a}_j^*)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.
Compute $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_n)^\top = (\mathbf{K} + n\epsilon \mathbf{I})^{-1} \widetilde{\mathbf{K}} \mathbf{1}_m \in \mathbb{R}^n$, where $\mathbf{1}_m := \frac{1}{m}(1, \ldots, 1)^\top \in \mathbb{R}^m$.
Output: An estimate $\hat{\mu}_{\mathbb{P}_*} = \sum_{i=1}^n \beta_i \ell(\cdot, r_i)$ of the mean embedding (38) or an estimate $\hat{R}_* := \sum_{i=1}^n \beta_i r_i$ of the expected reward (32).

Extensions. Note that the above method of approximating the covariate distribution $\mathbb{P}_{X_1}$ via the sampling procedure in (39) does not fully exploit the information of the target policy $\pi_*(\mathbf{a}|\mathbf{u})$, since for each $\mathbf{u}_j^*$ we only sample one action $\mathbf{a}_j^* \sim \pi_*(\mathbf{a}|\mathbf{u}_j^*)$. In Appendix A, we discuss extensions of Algorithm 2 to make use of more information from the target policy.

7. Experiments

This section provides empirical results that demonstrate the advantages of the proposed framework. The codes to reproduce the experiments are available at `https://github.com/sorawitj/counterfactual-mean-embedding`.

7.1 Simulations: Distributional Treatment Effects

We first conduct simulation experiments on distributional treatment effects in Section 7.1.1 and on distributional effects of covariate distributions in Section 7.1.2.

7.1.1 Distributional Treatment Effects (DTE)

We first deal with the identification of DTE (Sections 3.1.1), defined as the difference between the distributions $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$ of two potential outcomes $Y_0^*, Y_1^* \in \mathbb{R}$. As discussed in Section 5.2, this identification problem can be formulated as hypothesis testing of the null hypothesis $H_0 : \mathbb{P}_{Y_1^*} = \mathbb{P}_{Y_0^*}$ against the alternative $H_1 : \mathbb{P}_{Y_1^*} \neq \mathbb{P}_{Y_0^*}$. For this purpose, we assume that i.i.d. observations $\{(\mathbf{x}_i, t_i, \mathbf{y}_i)\}_{i=1}^N$ of random variables (X, T, Y) are available, where $X \in \mathbb{R}^5$ is covariates, $T \in \{0, 1\}$ is a treatment indicator, and $Y = \mathbb{1}(T = 0) Y_0^* + \mathbb{1}(T = 1) Y_1^* \in \mathbb{R}$ is the observed outcome.

The purpose here is to demonstrate the validity of our approach to identifying DTE, described in Sections 3.5.1 and Section 5.2. To this end, we compare it with a baseline approach that uses an estimate of ATE (9) as a test statistic. For simplicity, we call here our approach "DTE" and the baseline "ATE". We consider the following three scenarios:

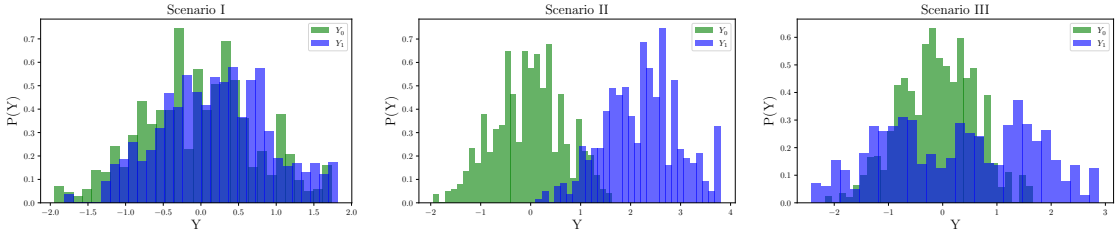


Figure 2: Histograms of observed outcomes $(\mathbf{y}_i)_{i=1}^N$ from the data $\{(\mathbf{x}_i, t_i, \mathbf{y}_i)\}_{i=1}^N$ generated under the three scenarios in Section 7.1.1, with $N = 500$. For each scenario, the green histogram consists of outcomes $\mathbf{y}_i$ with $t_i = 0$, which are i.i.d. with $Y_0 = Y|(T = 0)$), and the blue histogram consists of $\mathbf{y}_i$ with $t_i = 1$, which are i.i.d. with $Y_1 = Y|(T = 1)$. Note that $Y_0 = Y_0^*|(T = 0)$ and $Y_1 = Y_1^*|(T = 1)$, so the distributions of Y_0 and Y_1 (described here) are slightly different from those of the potential outcomes Y_0^* and Y_1^*.

Scenario I. There exists no treatment effect so that the distributions of the potential outcomes Y_0^*, Y_1^* are the same: $\mathbb{P}_{Y_0^*} = \mathbb{P}_{Y_1^*}$. Hence, we expect that both ATE and DTE do not detect any treatment effect.

Scenario II. There exists a treatment effect that only makes the means of $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$ different: *i.e.*, the mean-shift scenario. Hence, we expect that both ATE and DTE can detect the treatment effect.

Scenario III. There exists a treatment effect that does not change the means of $\mathbb{P}_{Y_0^*}$ and $\mathbb{P}_{Y_1^*}$, but changes their higher order moments. Hence, we expect that ATE fails to detect any treatment effect, whereas DTE with non-linear kernels can detect the difference.

To realize these scenarios, we define the random variables X, T, Y_0^* and Y_1^* as

$$X \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{x}} \mathbf{I}_5), \quad T \sim \text{Bernoulli}\left(\frac{1}{1 + \exp\left(-\boldsymbol{\alpha}^\top X - \alpha_0\right)}\right)$$
$$Y_0^* = \beta^\top X + \varepsilon_0, \quad Y_1^* = \beta^\top X + b + \varepsilon_1,$$

where $\varepsilon_0, \varepsilon_1 \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ are independent noises. Throughout the experiment, we set $\boldsymbol{\beta} = [0.1, 0.2, 0.3, 0.4, 0.5]^\top$, $\boldsymbol{\alpha} = [0.05, 0.04, 0.03, 0.02, 0.01]^\top$, $\alpha_0 = 0.05$, and $\sigma_\varepsilon^2 = \sigma_{\mathbf{x}}^2 = 0.1$. We set $b = 0$ for the Scenario I and $b = 2$ for the Scenario II. For Scenario III, we set $b = 2z - 1$, where $z \in \{0, 1\}$ is an independent Bernoulli random variable $z \sim \text{Bernoulli}(0.5)$ generated for every observation. By construction, the conditional exogeneity $Y_0^*, Y_1^* \perp\!\!\!\perp T|X$ in Assumption 1 is satisfied. For each scenario, we generate data $\{(\mathbf{x}_i, t_i, \mathbf{y}_i)\}_{i=1}^N$ as i.i.d. observations of X, T and $Y = \mathbb{1}(T = 0)Y_0^* + \mathbb{1}(T = 1)Y_1^*$, with $N \in \{50, 100\}$. Figure 2 describes the empirical distributions of observed outcomes for the three scenarios, where $Y_0 = Y|(T = 0)$ and $Y_1 = Y|(T = 1)$.

For DTE and ATE, we perform the following tests using data $\{(\mathbf{x}_i, t_i, \mathbf{y}_i, e_i)\}_{i=1}^n$ augmented with propensity scores $e_i := e(\mathbf{x}_i) := \mathbb{E}[T \,|\, X = \mathbf{x}_i]$. For DTE, we use the unbiased KTE estimate in (23) as a test statistic, with the Gaussian kernel $\ell(\mathbf{y}, \mathbf{y}') = \exp(-\|\mathbf{x} - \mathbf{y}\|_2^2 / 2\sigma^2)$ whose bandwidth parameter σ is chosen using the median heuristic

		$N = 50$		$N = 100$	
		ATE	**DTE**	**ATE**	**DTE**
Scenario I:	No Treatment Effect	0.013	0.012	0.013	0.012
Scenario II:	Mean Shift Effect	1.000	1.000	1.000	1.000
Scenario III:	High-order Treatment Effect	**0.012**	**0.224**	**0.012**	**0.639**

Table 1: The frequencies of rejecting the null hypothesis $H_0 : \mathbb{P}_{Y_1^*} = \mathbb{P}_{Y_0^*}$ when the null hypothesis is true (*i.e.*, the probability of the Type-I error in Scenario I) and when the alternative hypothesis $H_1 : \mathbb{P}_{Y_1^*} \neq \mathbb{P}_{Y_0^*}$ is true (*i.e.*, the power of the test in Scenario II & III), computed from 1000 repetitions. The significance level α is 0.01.

(Garreau et al., 2017). For ATE, we also use (23) as a test statistic, but with the linear kernel $\ell(\mathbf{y}, \mathbf{y}') = \mathbf{y}^\top \mathbf{y}'$, resulting in a test that distinguishes only the means of two distributions. We use the bootstrap procedure described in Section 5.2 to construct the distribution of the test statistic under the null $H_0 : \mathbb{P}_{Y_0^*} = \mathbb{P}_{Y_1^*}$, with $B = 10,000$ bootstrap samples. The significance level α is set to 0.01 in all experiments.

Table 1 reports the frequencies of rejecting the null hypothesis $H_0 : \mathbb{P}_{Y_0^*} = \mathbb{P}_{Y_1^*}$ over 1000 repetitions, for each of the three scenarios. When the null hypothesis H_0 is true (Scenario I), these are the frequencies of Type-I errors, which are well calibrated approximately at the designed level $\alpha = 0.01$ for both ATE and DTE. When the alternative hypothesis $H_1 : \mathbb{P}_{Y_0^*} \neq \mathbb{P}_{Y_1^*}$ is true (Scenarios II and III), these represent test powers (*i.e.*, one minus the probability of Type II error). In Scenario II, both ATE and DTE successfully reject the null hypothesis, capable of detecting the mean shift effect in the potential outcome distributions. In Scenario III, where the treatment effects do not appear in the mean but in the higher order moments, DTE has significantly higher power than ATE, demonstrating that DTE can identify higher order distributional effects.

7.1.2 Distributional Effects of Covariate Distributions

We next consider the identification of distributional effects of covariate distributions (see Sections 3.1.3, 3.3 and 3.5.3). As before, let $Y_0^*, Y_1^* \in \mathbb{R}$ be potential outcomes, $T \in \{0, 1\}$ be a treatment indicator, $Y = \mathbb{1}(T = 0)Y_0^* + \mathbb{1}(T = 1)Y_1^*$ be the observed outcome, and $X \in \mathbb{R}^5$ be covariates. Let $X_0 = X|(T = 0)$, $X_1 = X|(T = 1)$ and $Y_0 = Y|(T = 0) = Y_0^*|(T = 0)$. Here we are interested in the distributional effects defined as

$$\begin{aligned}\mathbb{P}_{Y_0^* \,|\, T}(\cdot \,|\, 0) - \mathbb{P}_{Y_0^* \,|\, T}(\cdot \,|\, 1) &= \mathbb{P}_{Y\langle 0|0\rangle} - \mathbb{P}_{Y\langle 0|1\rangle} \\ &= \int \mathbb{P}_{Y_0|X_0}(\cdot|\mathbf{x})\mathbb{P}_{X_0}(\mathbf{x}) - \int \mathbb{P}_{Y_0|X_0}(\cdot|\mathbf{x})\mathbb{P}_{X_1}(\mathbf{x})\end{aligned}$$

where the first identity holds under the conditional exogeneity. As discussed in Section 5.2, the identification of this distributional effect can be cast as testing the null hypothesis $H_0 : \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0) = \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$ against the alternative $H_1 : \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0) \neq \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$.

For this experiment, we define the joint distribution of T, X and Y_0^* by first specifying the distribution of T, and then specifying the conditional distributions of X, Y_0^* and Y_1^* given

T (note that Y_1^* is not relevant in this experiment). To this end, we define the distribution $\mathbb{P}_T$ of T as $\mathbb{P}_T(0) = \mathbb{P}_T(1) = 1/2$. Then, we define the conditional distributions of X and Y_0^* given T as

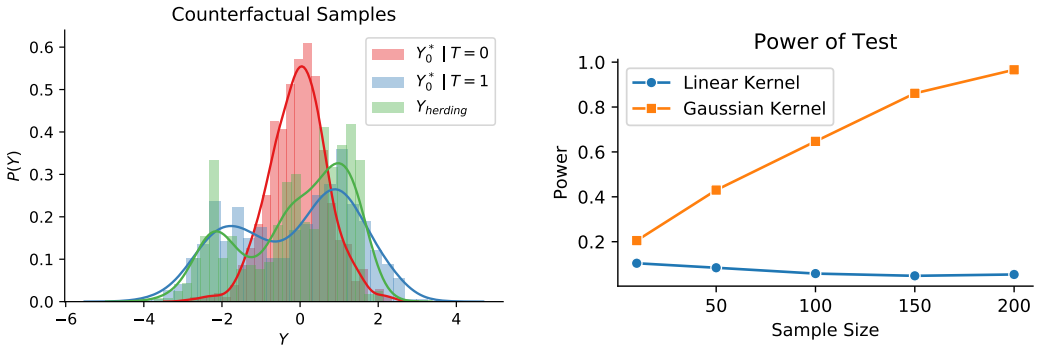
$$Y_0^* \,|\, (T=0) = \boldsymbol{\beta}^\top X_0 + \varepsilon_0, \quad X_0 = X|(T=0) \sim \mathcal{N}(\mathbf{0}, \sigma_x \mathbf{I}_5),$$
$$Y_0^* \,|\, (T=1) = \boldsymbol{\beta}^\top X_1 + \varepsilon_1, \quad X_1 = X|(T=1) \sim \sum_{j=1}^{3} \gamma_j \mathcal{N}(\boldsymbol{\nu}_j, \sigma_x \mathbf{I}_5),$$

where $\varepsilon_0, \varepsilon_1 \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ are independent, $\boldsymbol{\beta} = [0.1, 0.2, 0.3, 0.4, 0.5]^\top$, $\sigma_\varepsilon = \sigma_x = 0.1$, and $\gamma_1 = \gamma_2 = \gamma_3 = 1/3$. We set $\boldsymbol{\nu}_1 = [-5, 2.5, 0, 0, 2.5]$, $\boldsymbol{\nu}_2 = [2.5, 2.5, 0, 0, -5]$, and $\boldsymbol{\nu}_3 = [2.5, -5, 0, 0, 2.5]$, so that X_0 and X_1 have the same zero mean. By construction, $Y_0^* \,|\, T = 0$ and $Y_0^* \,|\, T = 1$ have the same mean, which is zero, while their higher-order moments differ. In other words, the distributional effects of the covariate distributions appear only in the higher-order moments. We generate data $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ as i.i.d. observations of (X_0, Y_0) (recall that $Y_0 = Y_0^*|(T=0)$) and $(\mathbf{x}'_j)_{j=1}^m$ as i.i.d. observations of X_1, where $n = m$ (which amounts to $\mathbb{P}_T(0) = \mathbb{P}_T(1) = 1/2$).

We estimate the embedding $\mu_{Y_0|T=1} = \int \ell(\cdot, \mathbf{y}) \, \mathrm{d}\mathbb{P}_{Y_0|T}(\mathbf{y}|1)$ of the counterfactual distribution $\mathbb{P}_{Y_0|T}(\cdot|1) = \mathbb{P}_{Y\langle 0|1\rangle}$ with the CME estimator (18) based on $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ and $(\mathbf{x}'_j)_{j=1}^m$. We set the kernel ℓ on the outcome space as the Gaussian kernel $\ell(\mathbf{y}, \mathbf{y}') = \exp(-\|\mathbf{y} - \mathbf{y}'\|_2^2 / 2\sigma_Y^2)$ whose bandwidth parameter σ_Y is chosen by the median heuristic using $(\mathbf{y}_i)_{i=1}^n$. We also set the kernel k on the covariate space as the Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\sigma_X^2)$, whose parameter σ_X as well as the regularization constant ε in the CME estimator are chosen by 5-fold cross validation from $\sigma_X \in \{0.01, 0.1, 1, 10\}$ and $\varepsilon \in \{0.01, 0.1, 1, 10\}$. This cross validation is done by regarding the joint sample $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ as training data for regression from $\mathbf{x}_i$ to $\mathbf{y}_i$, and by performing kernel ridge regression with kernel k and regularization parameter ε, motivated by the interpretation of conditional mean embedding as kernel ridge regression (Grünewälder et al., 2012).

We apply Algorithm 1 to the resulting CME estimate $\hat{\mu}_{Y\langle 0|1\rangle} = \sum_{i=1}^n \beta_i \ell(\cdot, \mathbf{y}_i)$ to generate counterfactual samples $(\mathbf{y}'_j)_{j=1}^n$. We now have $(\mathbf{y}_i)_{i=1}^n$ as an i.i.d. sample from $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0)$ and $(\mathbf{y}'_j)_{j=1}^n$ as an approximate sample of the counterfactual distribution $\mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$. As discussed in Section 5.2, we can test the null hypothesis $H_0 : \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0) = \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$ against the alternative $H_1 : \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 0) \neq \mathbb{P}_{Y_0^*|T}(\cdot \,|\, 1)$ by performing a two sample test using the samples $(\mathbf{y}_i)_{i=1}^n$ and $(\mathbf{y}'_j)_{j=1}^n$. For this purpose, we perform the kernel two-sample test with the unbiased MMD statistic (Gretton et al., 2012, Eq. 3), with permutation-based bootstrapping using $B = 10,000$ bootstrap samples and with significance level $\alpha = 0.01$. For comparison, we also perform the same kernel two-sample test, but with linear kernel $\ell(\mathbf{y}, \mathbf{y}') = \mathbf{y}^\top \mathbf{y}'$, resulting in a test that only uses the means of $(\mathbf{y}_i)_{i=1}^n$ and $(\mathbf{y}'_j)_{j=1}^n$.

Figure 3 describes the experimental results. Figure 3(a) illustrates the observed outcomes $(\mathbf{y}_i)_{i=1}^n$ of $Y_0^*|(T=0)$ (red), a counterfactual sample of $Y_0^*|(T=1)$ (blue) and the approximate counterfactual sample $(\mathbf{y}'_j)_{j=1}^n$ generated with Algorithm 1 applied to our CME estimate (green). Note that the sample of $Y_0^*|(T=1)$ is shown here for an illustration purpose; in practice we never have access to such a sample, but we can generate it here as we know the ground-truth model. For illustration, we also show the corresponding density curves obtained from the respective samples using kernel density estimation. The approxi-



(a) (b)

Figure 3: Results of the experiments in Section 7.1.2. (a) Histograms of observed outcomes $(\mathbf{y}_i)_{i=1}^n$ of $Y_0^*|(T=0)$ (red), a counterfactual sample of $Y_0^*|(T=1)$ (blue) and the approximate counterfactual sample $(\mathbf{y}'_j)_{j=1}^n$ generated with Algorithm 1 applied to our CME estimate (green), obtained from data with size $n = 500$. For illustration, we also show density curves obtained from the corresponding samples of the same colors, estimated with kernel density estimation. (b) Powers of the two sample tests based on the generated counterfactual sample $(\mathbf{y}'_j)_{j=1}^n$ and observed outcomes $(\mathbf{y}_i)_{i=1}^n$, using the unbiased MMD statistic with the Gaussian or the linear kernel with significance level $\alpha = 0.01$. The powers are obtained from 1,000 repetitions, for each of different sample sizes.

mate counterfactual sample $(\mathbf{y}'_j)_{j=1}^n$ resembles that from the ground-truth model, supporting the validity of our CME estimator (18) and the sampling method (Algorithm 1).

Figure 3(b) describes the test powers (i.e., the frequencies of rejecting the null hypothesis $H_0 : \mathbb{P}_{Y_0^*|T}(\cdot \mid 0) = \mathbb{P}_{Y_0^*|T}(\cdot \mid 1)$) over 1,000 repetitions of the above testing procedure for each case of using the Gaussian or the linear kernel for computing the test statistic, for different sample sizes. The test with the linear kernel has very low power. This implies that the mean of the generated counterfactual sample $(\mathbf{y}'_j)_{j=1}^n$ is close to the mean of the observed sample $(\mathbf{y}_i)_{i=1}^n$ from $Y_0^*|(T=0)$ since the kernel two-sample test with the linear kernel only uses the information of the sample means. On the other hand, the power of the test with the Gaussian kernel increases as the size n of observed data increases, suggesting that the higher-order moments of the generated counterfactual sample $(\mathbf{y}'_j)_{j=1}^n$ differ substantially from those of the observed sample $(\mathbf{y}_i)_{i=1}^n$. These observations suggest that the approximate counterfactual sample $(\mathbf{y}'_j)_{j=1}^n$ has properties consistent with the ground-truth counterfactual distribution $Y_0^*|T=1$. Thus our CME estimator (18) and Algorithm 1 are capable of producing an approximate counterfactual sample based on which a test for distributional effects can be constructed.

7.2 Off-Policy Evaluation

We conduct experiments on the off-policy evaluation (OPE) task for a recommendation system, described in Section 6. Let $\eta : \mathcal{U} \times \mathcal{A} \to \mathbb{R}$ be the regression function that takes a

pair $(\mathbf{u}, \mathbf{a})$ of user features $\mathbf{u} \in \mathcal{U}$ and recommendation $\mathbf{a} \in \mathcal{A}$ as an input and outputs the conditional expectation of the reward r:

$$\eta(\mathbf{u}, \mathbf{a}) := \mathbb{E}[r \,|\, \mathbf{u}, \mathbf{a}] := \int_{\mathbb{R}} r \, \mathrm{d}\mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a}) = \int_{\mathbb{R}} r \, \mathrm{d}\mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a}),$$

where $\mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a}) = \mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a})$ is the conditional distribution of the reward r given the pair $(\mathbf{u}, \mathbf{a})$, which is assumed to be invariant under the target and logging environments.

For a given target policy $\pi_*(\mathbf{a}|\mathbf{u})$, the OPE task is to estimate the expected reward under the target environment defined by

$$R_* := \int_{\mathcal{U}\times\mathcal{A}} \int_{\mathbb{R}} r \, \mathrm{d}\mathbb{P}_*(r \,|\, \mathbf{u}, \mathbf{a}) \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a}) = \int_{\mathcal{U}\times\mathcal{A}} \eta(\mathbf{u}, \mathbf{a}) \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a}),$$

where $\pi_*(\mathbf{u}, \mathbf{a}) = \pi_*(\mathbf{a} \,|\, \mathbf{u}) q_*(\mathbf{u}) = \pi_*(\mathbf{a} \,|\, \mathbf{u}) q_0(\mathbf{u})$ is the joint distribution of context features $\mathbf{u} \in \mathcal{U}$ and action $\mathbf{a} \in \mathcal{A}$. Here we consider the standard setting where the marginal distributions of the user features $\mathbf{u}$ are the same under the target and logging environments: $q_*(\mathbf{u}) = q_0(\mathbf{u})$. The above estimation is to be done based on the logged data $\mathcal{D}_{\text{init}} := \{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ obtained from the joint distribution $\mathbb{P}_0(\mathbf{u}, \mathbf{a}, r) = \mathbb{P}_0(r \,|\, \mathbf{u}, \mathbf{a}) \pi_0(\mathbf{a}, \mathbf{u})$ during the data collection phase, where $\pi_0(\mathbf{u}, \mathbf{a}) = \pi_0(\mathbf{u}|\mathbf{a}) q_0(\mathbf{u})$.

We compare our approach in Algorithm 2, which we call `CME` below, to the following benchmark estimators using both simulated and real-world data.

Direct method with a parametric regressor (DM). The direct method (Dudík et al., 2011) first learns the regression function η based on the logged data $\mathcal{D}_{\text{init}}$ with a regression model of one's choice. Let $\hat{\eta} : \mathcal{U} \times \mathcal{A} \to \mathbb{R}$ be the learned regressor. Then the expected reward R_* is estimated as

$$\widehat{R}_{\text{DM}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbf{a} \sim \pi_*(\mathbf{a} \,|\, \mathbf{u}_i)}[\hat{\eta}(\mathbf{u}_i, \mathbf{a})].$$

The direct method obtains the approximation $\hat{\eta}$ based on the logged data $\mathcal{D}_{\text{init}} = \{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$, in which input pairs $(\mathbf{u}_i, \mathbf{a}_i)$ are generated from the covariate distribution $\pi_0(\mathbf{u}, \mathbf{a}) = \pi_0(\mathbf{u}|\mathbf{a}) q_0(\mathbf{u})$ that is different from the target covariate distribution $\pi_*(\mathbf{u}, \mathbf{a}) = \pi_*(\mathbf{u}|\mathbf{a}) q_0(\mathbf{u})$. Recall that we interpret the paired variables $(\mathbf{u}, \mathbf{a})$ as "covariates" in our discussion. This situation is known as *covariate shift* in the literature. It is well known that under the covariate shift, a *parametric* regression model may produce a significant bias (Shimodaira, 2000). That is, the approximation quality of the learned model $\hat{\eta}$ obtained with the logged data $\mathcal{D}_{\text{init}}$ may be good with respect to the covariate distribution $\pi_0(\mathbf{u}, \mathbf{a})$ under the data collection environment, but can be poor with respect to the target covariate distribution $\pi_*(\mathbf{u}, \mathbf{a})$, e.g., $\|\eta - \hat{\eta}\|^2_{L_2(\pi_*)} := \int (\eta(\mathbf{u}, \mathbf{a}) - \hat{\eta}(\mathbf{u}, \mathbf{a}))^2 \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a})$ may be large. This in turn may induce a large bias in the estimation of the expected reward $R_* = \int \eta(\mathbf{u}, \mathbf{a}) \, \mathrm{d}\pi_*(\mathbf{u}, \mathbf{a})$. To demonstrate this, we use a 3-layer feedforward neural network, which is an (overparametrized) parametric model, as a regressor for the direct method.

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 the *propensity scores* under the target and initial policies (Horvitz and Thompson, 1952; Precup et al., 2000). The `wIPS` estimator is defined by

$$\widehat{R}_{\mathrm{wIPS}} = \left(\sum_{i=1}^{n} w_i r_i\right) \Big/ \left(\sum_{i=1}^{n} w_i\right),$$

where $w_i := \pi_*(\mathbf{a}_i, \mathbf{u}_i)/\pi_0(\mathbf{a}_i, \mathbf{u}_i) = \pi_*(\mathbf{a}_i|\mathbf{u}_i)/\pi_0(\mathbf{a}_i|\mathbf{u}_i)$ 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 (Cassel et al., 1976; Dudík et al., 2011). The estimator is given by

$$\widehat{R}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_{\mathbf{a}\sim\pi_*(\mathbf{a}\,|\,\mathbf{u}_i)}[\hat{\eta}(\mathbf{u}_i, \mathbf{a})] + w_i(r_i - \hat{\eta}(\mathbf{u}_i, \mathbf{a}_i))\right).$$

It has been proved to be unbiased if at least one of the estimators, $\hat{\eta}$ and π_*/π_0 is correctly specified; see, *e.g.*, Cassel et al. (1976); Dudík et al. (2011).

Slate estimator. The `Slate` estimator, proposed for recommendation systems, makes use of the structure within a recommendation (= action) by assuming a certain linearity assumption on the regression function with respect to the recommendation (Swaminathan et al., 2017). More precisely, Swaminathan et al. (2017) consider a recommendation system in which an action $\mathbf{a} \in \mathcal{A}$ is an ordered list (called *slate*) of $K \in \mathbb{N}$ items chosen from $M \in \mathbb{N}$ possible items. Let $\mathbf{1}_\mathbf{a} \in \mathbb{R}^{KM}$ be the indicator vector whose (k, m)-th element is 1 if $\mathbf{a}$ contains the item $m \in \{1, \ldots, M\}$ in the slot $k \in \{1, \ldots, K\}$, and 0 otherwise. Swaminathan et al. (2017, Assumption 1) then model the regression function $\eta(\mathbf{u}, \mathbf{a})$ as a linear function of this indicator vector: $\eta(\mathbf{u}, \mathbf{a}) = \mathbf{w}_\mathbf{u}^\top \mathbf{1}_\mathbf{a}$, where $\mathbf{w}_\mathbf{u}$ is an unknown feature vector of the context $\mathbf{u}$ (Note that $\mathbf{w}_u$ can be a nonlinear function of $\mathbf{u} \in \mathcal{U}$). Under this assumption, the authors derive the slate estimator as

$$\widehat{R}_{\mathrm{slate}} = \frac{1}{n}\sum_{i=1}^{n} r_i \cdot \mathbf{q}_{\mathbf{u}_i}^\top \Gamma_{\mathbf{u}_i}^\dagger \mathbf{1}_{\mathbf{a}_i},$$

where $\Gamma_{\mathbf{u}_i}^\dagger$ is the Moore-Penrose pseudoinverse of the matrix $\Gamma_{\mathbf{u}_i} := \mathbb{E}_{\mathbf{a}\sim\pi_*(\mathbf{a}\,|\,\mathbf{u}_i)}[\mathbf{1}_\mathbf{a}\mathbf{1}_\mathbf{a}^\top] \in \mathbb{R}^{KM\times KM}$, and $\mathbf{q}_{\mathbf{u}_i} := \mathbb{E}_{\mathbf{a}\sim\pi_*(\mathbf{a}\,|\,\mathbf{u}_i)}[\mathbf{1}_\mathbf{a}] \in \mathbb{R}^{KM}$. Thanks to the linearity assumption, the slate estimator may enjoy a lower variance than the `wIPS` estimator, while the assumption may also lead to a non-vanishing bias if it does not hold.

For the `CME`, we use a kernel defined as $k((\mathbf{u}, \mathbf{a}), (\mathbf{u}', \mathbf{a}')) := k_\mathcal{U}(\mathbf{u}, \mathbf{u}')k_\mathcal{A}(\mathbf{a}, \mathbf{a}')$ where $k_\mathcal{U}(\mathbf{u}, \mathbf{u}') := \exp\left(-\|\mathbf{u} - \mathbf{u}'\|_2^2/2\sigma_u^2\right)$ and $k_\mathcal{A}(\mathbf{a}, \mathbf{a}') := \exp\left(-\|\mathbf{a} - \mathbf{a}'\|_2^2/2\sigma_a^2\right)$. For this experiment, the linear kernel $\ell(r, r') := rr'$ is used as a reward kernel since we only compare the estimation of the expected reward. The regularization parameter ε is selected by the cross validation procedure in Appendix B, while we determined σ_u and σ_a by the median heuristic, *i.e.*, $\sigma_u^2 = \mathrm{median}\{\|\mathbf{u}_i - \mathbf{u}_j\|_2^2\}_{1\le i<j\le n}$ and $\sigma_a^2 = \mathrm{median}\{\|\mathbf{a}_i - \mathbf{a}_j\|_2^2\}_{1\le i<j\le n}$.

Before proceeding, we point out here a connection between our approach and the direct method. Assume that we use kernel ridge regression to obtain the approximation $\hat{\eta}$ of

the regression function η using the logged data: $\hat{\eta}(\mathbf{u}, \mathbf{a}) = \mathbf{r}^\top (\mathbf{K} + n\epsilon I)^{-1} \widetilde{\mathbf{k}}(\mathbf{u}, \mathbf{a})$, where $\mathbf{r} := (r_1, \ldots, r_n)^\top \in \mathbb{R}^n$ and $\widetilde{\mathbf{k}}(\mathbf{u}, \mathbf{a}) := (k((\mathbf{u}_j, \mathbf{a}_j), (\mathbf{u}, \mathbf{a})))_{j=1}^n \in \mathbb{R}^n$ Then, the estimate of the direct method can be related to the CME estimate as

$$\begin{aligned} \widehat{R}_{\mathrm{DM}} &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbf{a} \sim \pi_*(\mathbf{a} \,|\, \mathbf{u}_i)} [\hat{\eta}(\mathbf{u}_i, \mathbf{a})] \approx \frac{1}{n} \sum_{i=1}^n \hat{\eta}(\mathbf{u}_i, \mathbf{a}_i^*) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{r}^\top (\mathbf{K} + n\epsilon I)^{-1} \widetilde{\mathbf{k}}(\mathbf{u}_i, \mathbf{a}_i^*) = \mathbf{r}^\top (\mathbf{K} + n\epsilon I)^{-1} \frac{1}{n} \sum_{i=1}^n \widetilde{\mathbf{k}}(\mathbf{u}_i, \mathbf{a}_i^*), \end{aligned} \tag{40}$$

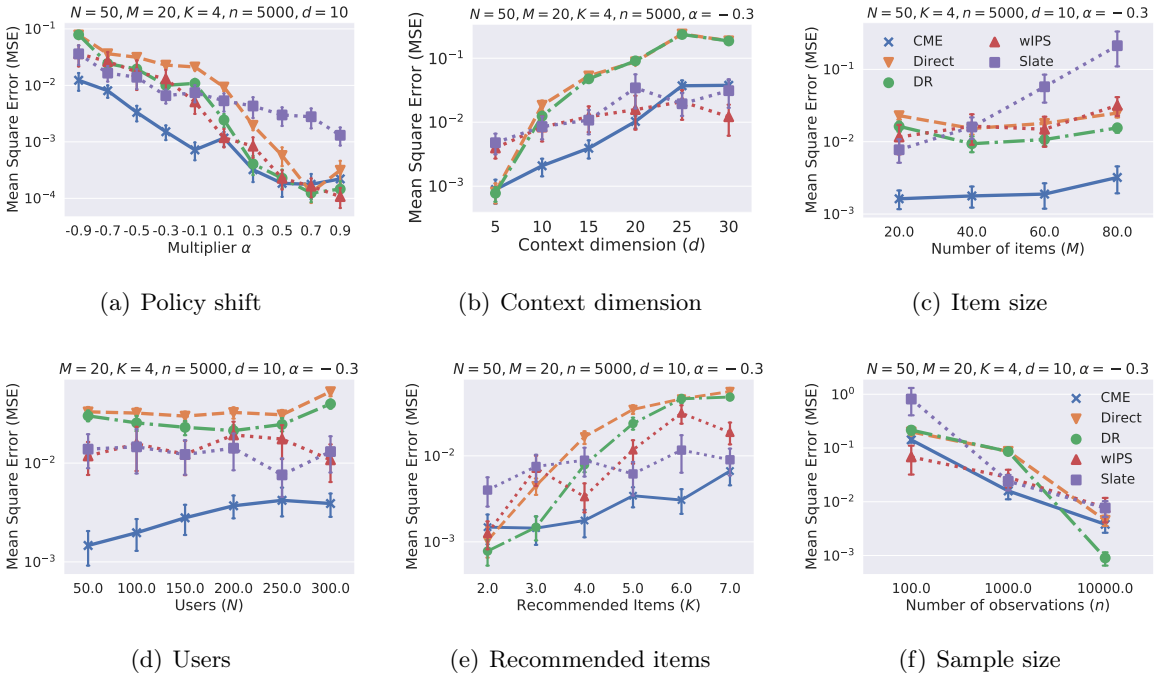
where the approximation in the first line is a Monte Carlo approximation based on a single draw $\mathbf{a}_i^*$ from the target policy $\pi_*(\mathbf{a} \,|\, \mathbf{u}_i)$ for each i. As we can see from (40), the estimate has the same form as the CME estimate given in Algorithm 2 when the output kernel ℓ is a linear kernel, i.e., when we are only interested in the expected reward.

In this sense, the CME estimate can be interpreted as the direct method with kernel ridge regressor. Note that the kernel ridge regression is a *nonparametric* method (as long as the RKHS of the kernel on covariates is infinite dimensional such as the RKHS of the Gaussian kernel), and thus less prone to the effects of covariate shift. In fact, our convergence results in Section 4 show that the CME is consistent and thus asymptotically unbiased. This explains why our method, even if it can be related to the direct method, works well in the off-policy evaluation task compared to the direct method using a parametric model (as we will show shortly).

7.2.1 Simulated Data

We consider the following setting for our simulation experiment. 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 $m \in \{1, \ldots, M\}$ is represented by a feature vector $\mathbf{v}_m \in \mathbb{R}^d$ generated randomly as $\mathbf{v}_m \sim \mathcal{N}(0, \mathbf{I}_d)$, where $d \in \mathbb{N}$. Hence, a recommendation is an ordered list $\mathbf{a} = (\mathbf{v}_{m_1}, \mathbf{v}_{m_2}, ..., \mathbf{v}_{m_K}) \in \mathbb{R}^{d \times K}$, where $m_1, m_2, \ldots, m_K \subset \{1, \ldots, M\}$. Likewise, each user $j \in \{1, \ldots, N\}$ has a feature vector $\mathbf{u}_j \in \mathbb{R}^d$ generated as $\mathbf{u}_j \sim \mathcal{N}(0, \mathbf{I}_d)$, where $N \in \mathbb{N}$ is the number of users. The reward from a user 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 $\theta_{ij} = \mathbb{P}(\mathrm{click} \,|\, \mathbf{a}_i, \mathbf{u}_j) = 1/(1 + \exp(-\overline{\mathbf{a}}_i^\top \mathbf{u}_j + \epsilon_{ij}))$ be the probability of a click, where $\overline{\mathbf{a}}_i$ is the mean vector of the item vectors listed in $\mathbf{a}_i$, and $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ is an independent noise. The reward from user j receiving recommendation $\mathbf{a}_i$ is defined as $r_{ij} \sim \mathrm{Bernoulli}(\theta_{ij})$.

In this experiment, we consider the following policy setup. For each user j, a policy $\pi(\mathbf{a}|\mathbf{u})$ generates a list $\mathbf{a} = (\mathbf{v}_{m_1}, \mathbf{v}_{m_2}, ..., \mathbf{v}_{m_K})$ of K recommended items by sampling without replacement with respect to a multinomial distribution over all items. The probability of item $l \in \{1, \ldots, M\}$ being selected for user j is $\exp(\mathbf{b}_j^\top \mathbf{v}_l) / \sum_{k=1}^M \exp(\mathbf{b}_j^\top \mathbf{v}_k)$, where $\{\mathbf{b}_j\}_{j=1}^N$ are parameter vectors of the policy $\pi(\mathbf{a}|\mathbf{u})$. Note that we obtain an optimal policy if $\mathbf{b}_j = \mathbf{u}_j$ for all $j \in \{1, \ldots, N\}$. To construct initial policy $\pi_0(\mathbf{a}|\mathbf{u})$ and target policy $\pi_*(\mathbf{a}|\mathbf{u})$, we first randomly generate user feature vectors $\mathbf{u}_1, \ldots, \mathbf{u}_N$. Then, 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 $p_{jk} \sim \mathrm{Bernoulli}(0.5)$. That is, the parameter vector $\mathbf{b}_j^*$ is equal to the user feature vector with about half of its entries randomly set to zero. For the initial policy π_0, we set $\mathbf{b}_j = \alpha \mathbf{b}_j^*$ where $\alpha \in [-1, 1]$. The



(a) Policy shift (b) Context dimension (c) Item size

(d) Users (e) Recommended items (f) Sample size

Figure 4: 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, (c) the number of available items M, (d) the number of users N, (e) the number of recommended items K, and (f) the number of observations n. Each error bar represents a 95% confidence interval.

parameter α controls how similar the policies are. If $\alpha = 1$, we obtain $\pi_0 = \pi_*$, whereas π_0 and π_* differ the most when $\alpha = -1$.

We generate two datasets $\mathcal{D}_{\text{init}} = \{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ and $\mathcal{D}_{\text{target}} = \{(\mathbf{u}_i^*, \mathbf{a}_i^*, r_i^*)\}_{i=1}^n$ using $\pi_0(\mathbf{a}|\mathbf{u})$ and $\pi_*(\mathbf{a}|\mathbf{u})$, respectively, where $\mathbf{u}_i = \mathbf{u}_i^*$ for $i = 1, \ldots, n$. Note that the target rewards $r_1^*, r_2^*, \ldots, r_n^*$ are only used for evaluation. Our task is to estimate the expected reward of the target policy from the remaining information. We perform 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 $\varepsilon \in \{10^{-8}, \ldots, 10^0\}$ for our `CME`. We repeat the experiments 30 times independently to obtain the mean square errors (MSE) and their 95% confidence intervals in the estimation of the expected reward for each estimator.

We investigate the behavior of different estimators as we vary different experimental conditions including the degree of difference between initial and target policies (α), the context dimensionality (d), the number of items (M), the number of users (N), the number of recommended items (K), and the number of observations (n). Figure 4 depicts the experimental results (note that vertical axis is in log scale). In brief, we find that *a*) 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, *b*) the `Slate` estimator does not perform well in this setting because its linearity assumption does not hold, *c*) all estimators deteriorate as the context dimension increases, but the effect appears to be more

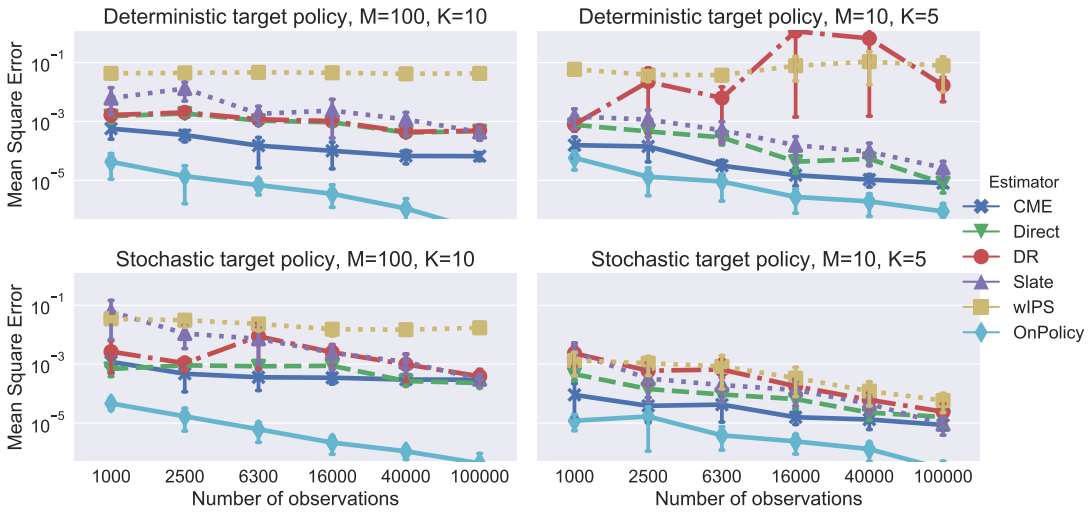


Figure 5: The performance of different estimators on the MSLR-WEB30K dataset.

pronounced for the `Direct`, `DR`, and `CME` estimators than for the `wIPS` and `Slate` estimators as they do not rely directly on the covariates, *d*) the opposite effect is observed if we increase the number of available items M, as illustrated in Figure 4(c), and *e*) the `CME` estimator achieves better performance than other estimators in most experiments.

7.2.2 Real Data

For our real data experiment, we use the data from the Microsoft Learning to Rank Challenge dataset (MSLR-WEB30K) (Qin and Liu, 2013) and treat them as an off-policy evaluation problem. We follow the same experiment setting as described in Swaminathan et al. (2017, Section 4.1). The data contains a set of queries and the corresponding URLs. Each pair of query q and URL u is represented by a feature vector $f_{q,u}$ and accompanied by a relevance judgment $\rho(q,u) \in \{0,...,4\}$. We consider the expected reciprocal rank (ERR) (Chapelle et al., 2009) as our reward function, which is defined as $\mathrm{ERR}(q,u) := \sum_{k=1}^{K} \frac{1}{k} \prod_{j=1}^{k-1}(1 - R(q,u_j))R(q,u_k)$, where $R(q,u) := \frac{2^{\rho(q,u)}-1}{2^{\mathrm{maxrel}}}$ with $\mathrm{maxrel} := 4$. In order to obtain distinct initial and target policies $\pi_0(\mathbf{a}|\mathbf{u})$ and $\pi_*(\mathbf{a}|\mathbf{u})$, the feature vector $f_{q,u}$ is split into URL features $f_{q,u}^{\mathrm{url}}$ and body features $f_{q,u}^{\mathrm{body}}$, which are used to train two regression models to predict the relevance score $\rho(q,u)$: a Lasso regression model is trained to predict $\rho(q,u)$ from $f_{q,u}^{\mathrm{url}}$ (denoted by $\mathrm{lasso}_{\mathrm{url}}$), and a regression tree model is trained to predict $\rho(q,u)$ from $f_{q,u}^{\mathrm{body}}$ (denoted by $\mathrm{tree}_{\mathrm{body}}$). These two regression models are then used in the initial and target policies as will be described below.

In order to generate logged data $\mathcal{D}_{\mathrm{init}}$, we first sample a query q uniformly from the dataset, and select top M candidate URLs based on the relevance scores predicted by the $\mathrm{tree}_{\mathrm{body}}$ model. We then generate K recommended URLs out of these top M candidates using an initial policy π_0 and compute its corresponding reward. The initial policy is a *stochastic* policy which recommends K URLs by sampling without replacement according to a multinomial distribution parameterized by $p_\alpha(u|q) \propto 2^{-\alpha[\log_2 \mathrm{rank}(u,q)]}$, where $\mathrm{rank}(u,q)$ is the ERR of the relevance score $\rho(q,u)$ predicted by the $\mathrm{tree}_{\mathrm{body}}$ model and $\alpha \geq 0$ is

an exploration rate parameter. For target data $\mathcal{D}_{\text{target}}$, we consider both *stochastic* and *deterministic* target policies. The *stochastic* target policy $\pi_*(\mathbf{a}|\mathbf{u})$ is similar to the initial policy described earlier except that it employs $\text{lasso}_{\text{url}}$ model for the predicted relevance scores, but this makes the two policies distinct; their top-10 rankings are only overlapping by 2.5 URLs, on average. On the other hand, the *deterministic* target policy directly selects top-K URLs ranked by the predicted relevance scores obtained from the $\text{lasso}_{\text{url}}$ model. In this experiment, we set the exploration rate parameter $\alpha = 1$ for the stochastic initial policy, and set $\alpha = 2$ for the stochastic target policy.

We compare our estimator (`CME`) with the benchmark estimators `Direct`, `wIPS`, `DR` and `Slate`. In addition, we include 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 make use of the Nyström approximation method (Williams and Seeger, 2001). We repeat the experiments 10 times independently to obtain the mean square errors (MSEs) and their 95% confidence intervals in the estimation of the expected reward for each estimator.

Figure 5 depicts the results. In short, our `CME` dominates other estimators in most of experimental conditions (note that the 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. The reason is that, in the deterministic setting, the propensity score adjustment requires that the treatments picked by logged and target policies match exactly. Otherwise, the propensity ratio $\pi_*(\mathbf{a}|\mathbf{u})/\pi_0(\mathbf{a}|\mathbf{u})$ will be zero. The exact match is likely not to happen in practice and leads to higher variance in estimation of rewards. 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 action space, while the `Slate` estimator is less sample-efficient, especially in the large action space.

8. Discussion

This paper presents a general-purpose kernel mean representation of counterfactual distributions called the counterfactual mean embedding (CME). It draws insights and tools from kernel methods in machine learning and the potential outcome framework in causal inference. We show that our estimator of counterfactual distributions exhibits appealing theoretical properties, and also serves as a practical tool for causal inference. Ultimately, we hope that our work will be useful not only for researchers in disciplines that rely on the potential outcome framework, such as social and biomedical sciences, but also for researchers in machine learning and statistics to develop novel methodology for counterfactual inference, since several important open questions still remain: *e.g.*, the use of high-order moments of counterfactual distributions, and how to handle a hidden confounder and an instrumental variable.

One promising application of our framework is in generating a sample from the counterfactual distribution. For instance, neuroscientists can visualize the fMRI images of subjects under alternative setups without explicitly conducting invasive experiments. In this case, the outcome variable corresponds to an fMRI image. Let $G_{\boldsymbol{\theta}}$ be a generative model over the outcome parametrized by a parameter vector $\boldsymbol{\theta}$. The choice of $G_{\boldsymbol{\theta}}$ can range from a mixture of Gaussians to deep generative models, *e.g.*, generative adversarial networks (GAN). An

estimate of the counterfactual distribution, denoted by $G^*_{\boldsymbol{\theta}}$, can be obtained via an optimization problem: $G^*_{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\| \hat{\mu}_{Y\langle 0|1\rangle} - \hat{\mu}_{G_{\boldsymbol{\theta}}} \right\|^2_{\mathscr{F}}$, where $\hat{\mu}_{Y\langle 0|1\rangle}$ is our CME estimate and $\hat{\mu}_{G_{\boldsymbol{\theta}}}$ denotes the mean embedding of $G_{\boldsymbol{\theta}}$ in the RKHS $\mathscr{F}$. Counterfactual sample generation is ubiquitous for qualitative analysis in many application domains. We leave it as an open problem to future research.

Acknowledgments

We express our sincere gratitude to the Action Editor and the anonymous reviewers, whose comments greatly helped improve the quality of the paper. We also 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). Sorawit Saengkyongam is supported by a research grant (18968) from VILLUM FONDEN.

Appendix A. Possible Extensions for Off-Policy Evaluation

Here we describe possible extensions of the proposed approach to the off-policy evaluation task described in Section 6.3 (Algorithm 2). As mentioned there, Algorithm 2 only generates one action $\mathbf{a}_j^* \sim \pi_*(\mathbf{a}|\mathbf{u}_j^*)$ for each $\mathbf{u}_j^*$, which does not fully exploit the information of the target policy $\pi_*(\mathbf{a}|\mathbf{u}_j^*)$. We show a possible approach to using more information from the target policy, thereby improving the quality of the algorithm.

First, notice that the weight vector $\boldsymbol{\beta} \in \mathbb{R}^n$ in Algorithm 2 depends on the sample $(\mathbf{x}'_j)_{j=1}^m = (\mathbf{u}_j^*, \mathbf{a}_j^*)_{j=1}^m$ only through the vector $\tilde{\mathbf{K}}\mathbf{1}_m \in \mathbb{R}^m$, where $\tilde{\mathbf{K}} = (k(\mathbf{x}_i, \mathbf{x}'_j)) \in \mathbb{R}^{n\times m}$ and $\mathbf{1}_m = \frac{1}{m}(1, \ldots, 1)^\top \in \mathbb{R}^m$. This vector can be written as

$$\tilde{\mathbf{K}}\mathbf{1}_m = (\frac{1}{m}\sum_{j=1}^m k(\mathbf{x}_i, \mathbf{x}'_j))_{i=1}^n = (\hat{\mu}_{X_1}(\mathbf{x}_i))_{i=1}^n \in \mathbb{R}^n, \tag{41}$$

where $(\mathbf{x}_i)_{i=1}^n = (\mathbf{u}_i, \mathbf{a}_i)_{i=1}^n$ are from the logged data, and $\hat{\mu}_{\mathbb{P}_{X_1}}$ is an empirical approximation of the mean embedding $\mu_{\mathbb{P}_{X_1}}$ of the covariate distribution $\mathbb{P}_{X_1}$, given by

$$\hat{\mu}_{X_1} = \frac{1}{m}\sum_{j=1}^m k(\cdot, \mathbf{x}'_j), \quad \mu_{X_1} = \int k(\cdot, \mathbf{x}) d\mathbb{P}_{X_1}(\mathbf{x}). \tag{42}$$

This implies that the role of the sample $(\mathbf{x}'_j)_{j=1}^m$ is essentially to approximate the kernel mean $\mu_{\mathbb{P}_{X_1}}$. In fact, the quality of the CME estimator (based on which Algorithm 2 is constructed) depends on the sample $(\mathbf{x}'_j)_{j=1}^m$ only through the the approximation error $\|\hat{\mu}_{X_1} - \mu_{X_1}\|_{\mathscr{F}}$ (see *e.g.*, the proof of Theorem 30).

Thus, Algorithm 2 may be improved by constructing a better approximation, say $\check{\mu}_{X_1}$, of the kernel mean $\hat{\mu}_{X_1}$, and replace (41) in the computation of the weight vector $\boldsymbol{\beta}$ by the evaluations of this new approximation:

$$\boldsymbol{\beta} := (\mathbf{K} + n\varepsilon\mathbf{I})^{-1}\boldsymbol{v} \in \mathbb{R}^n, \quad \boldsymbol{v} := (\check{\mu}_{X_1}(\mathbf{x}_i)) \in \mathbb{R}^n,$$

where $\mathbf{K} := (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$.

To construct $\check{\mu}_{X_1}$, recall that the kernel k is given as a product kernel $k(\mathbf{x}, \mathbf{x}') = k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}')k_{\mathcal{U}}(\mathbf{u}, \mathbf{u}')$ for $\mathbf{x} = (\mathbf{u}, \mathbf{a})$, $\mathbf{x} = (\mathbf{u}', \mathbf{a}')$, and rewrite the kernel mean μ_{X_1} as

$$\begin{aligned}
\mu_{X_1}(\mathbf{x}) = \mu_{\pi_*}(\mathbf{u}, \mathbf{a}) &= \int_{\mathcal{U}}\int_{\mathcal{A}} k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}')k_{\mathcal{U}}(\mathbf{u}, \mathbf{u}')\, \mathrm{d}\pi_*(\mathbf{a}'|\mathbf{u}')\, \mathrm{d}q_*(\mathbf{u}') \\
&= \int_{\mathcal{U}} \left(\int_{\mathcal{A}} k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}')\, \mathrm{d}\pi_*(\mathbf{a}'|\mathbf{u}')\right) k_{\mathcal{U}}(\mathbf{u}, \mathbf{u}')\, \mathrm{d}q_*(\mathbf{u}') \\
&\approx \frac{1}{m}\sum_{j=1}^m \left(\int_{\mathcal{A}} k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}')\, \mathrm{d}\pi_*(\mathbf{a}'|\mathbf{u}_j^*)\right) k_{\mathcal{U}}(\mathbf{u}, \mathbf{u}_j^*).
\end{aligned}$$

Thus, if we can approximate the integral in the last expression accurately, we can obtain a good approximation for the kernel mean.

One approach is to generate $M > 1$ actions $\mathbf{a}_{j1}, \ldots, \mathbf{a}_{jM} \sim \pi_*(\mathbf{a}|\mathbf{u}_j^*)$ from the target policy for each $\mathbf{u}_j^*$, and the approximate the integral as

$$\frac{1}{M}\sum_{\nu=1}^M k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}_{j\nu}^*) \approx \int_{\mathcal{A}} k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}')\, \mathrm{d}\pi_*(\mathbf{a}'|\mathbf{u}_j^*)$$

Thus, a new approximation of μ_{X_1} may be defined as

$$\check{\mu}_{X_1}(\mathbf{x}) := \check{\mu}_{\pi_*}(\mathbf{u}, \mathbf{a}) := \frac{1}{m} \sum_{j=1}^{m} \left(\frac{1}{M} \sum_{\nu=1}^{M} k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}^*_{j\nu}) \right) k_{\mathcal{U}}(\mathbf{u}, \mathbf{u}^*_j),$$

where $\mathbf{x} := (\mathbf{u}, \mathbf{a})$. Note that $M = 1$ recovers Algorithm 2.

Another approach is to *exactly* compute the integral $\int_{\mathcal{A}} k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}') \, \mathrm{d}\pi_*(\mathbf{a}'|\mathbf{u}^*_j)$ when it is possible, and define a new approximation of μ_{X_1} as

$$\check{\mu}_{X_1}(\mathbf{x}) := \check{\mu}_{\pi_*}(\mathbf{u}, \mathbf{a}) := \frac{1}{m} \sum_{j=1}^{m} \left(\int_{\mathcal{A}} k_{\mathcal{A}}(\mathbf{a}, \mathbf{a}') \, \mathrm{d}\pi_*(\mathbf{a}'|\mathbf{u}^*_j) \right) k_{\mathcal{U}}(\mathbf{u}, \mathbf{u}^*_j)$$

This essentially is the case of $M = \infty$. For instance, the integral can be computed analytically when the kernel $k_{\mathcal{A}}$ is Gaussian and the target policy $\pi_*(\mathbf{a}|\mathbf{u})$ is an additive Gaussian noise model of the form $\pi_*(\mathbf{a}|\mathbf{u}) = \mathcal{N}(\mathbf{a}|F(\mathbf{u}), \sigma^2)$ for some function $F : \mathcal{U} \to \mathcal{A}$ and $\sigma^2 > 0$. This way of using an analytic integral for approximating the kernel mean is studied in Nishiyama et al. (2020); we refer to this paper for details and other examples.

Appendix B. Cross Validation Procedure for Counterfactual Prediction

Here we describe an approach to cross validation for model selection in counterfactual prediction, focusing on the problem of off-policy evaluation (OPE). We use below the notation defined in Section 6. Unlike the standard situation in machine learning, performing cross validation directly on the logged data $\mathcal{D} := \{(\mathbf{u}_i, \mathbf{a}_i, r_i)\}_{i=1}^n$ may lead to a biased estimate on the performance measure for counterfactual prediction, resulting in sub-optimal model selection. This is due to the covariate shift – the change of the covariate distribution from $\pi_0(\mathbf{u}, \mathbf{a})$ in the data collection environment to $\pi_*(\mathbf{u}, \mathbf{a})$ in the target environment. To correct for the bias due to this distributional shift, we propose the following procedure for cross validation. A similar cross validation approach has been proposed by Sugiyama et al. (2007).

Let $\mathcal{M} := \{1, \ldots, M\}$ be a set of indicators for candidate models. (*e.g.*, each $m \in \mathcal{M}$ may represent a specific choice of the kernel k and regularization constant ε in our CME estimator.)

1. Split $\mathcal{D}$ into K folds: $\mathcal{D}_k = \{(\mathbf{u}_j, \mathbf{a}_j, r_j)\}_{j=q(k-1)+1}^{qk}$ for $k = 1, \ldots, K$ and $q = \lfloor n/K \rfloor$.
2. For each model $m \in \mathcal{M}$:
   (a) For each fold $k = 1, 2, \ldots, K$:
      i. Calculate $w_1, \ldots, w_q \geq 0$ 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 m-th parameter setting whose $\bar{e}_m$ is the 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.

Appendix C. Proofs for Section 3

C.1 Proof of Lemma 3

Proof We only show $\mathbb{P}_{Y\langle 0|0\rangle}(\mathbf{y}) = \mathbb{P}_{Y_0^*|T}(\mathbf{y}|0)$; the other identity $\mathbb{P}_{Y\langle 1|1\rangle}(\mathbf{y}) = \mathbb{P}_{Y_1^*|T}(\mathbf{y}|1)$ can be shown similarly. First notice that $Y_0 := Y \,|\, (T=0) = \sum_{t=0}^{1} \mathbb{1}(T=t) Y_t^* \,|\, (T=0) = Y_0^* \,|\, (T=0)$. From this and Definition 2, it follows that $\mathbb{P}_{Y\langle 0|0\rangle}(\mathbf{y}) = \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x}) \,\mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) = \int \mathbb{P}_{Y|T,X}(\mathbf{y}|0,\mathbf{x}) \,\mathrm{d}\mathbb{P}_{X|T}(\mathbf{x}|0) = \int \mathbb{P}_{Y_0^*|T,X}(\mathbf{y}|0,\mathbf{x}) \,\mathrm{d}\mathbb{P}_{X|T}(\mathbf{x}|0) = \mathbb{P}_{Y_0^*|T}(\mathbf{y}|0)$. ■

C.2 Proof of Lemma 4

Proof We only show here $\mathbb{P}_{Y\langle 0|1\rangle} = \mathbb{P}_{Y_0^*|T=1}$; the other identity $\mathbb{P}_{Y\langle 1|0\rangle} = \mathbb{P}_{Y_1^*|T=0}$ can be shown similarly. We first derive basic identities: (a) The conditional exogeneity implies that $\mathbb{P}_{Y_0^*|T,X}(\mathbf{y}|1,\mathbf{x}) = \mathbb{P}_{Y_0^*|X}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_{Y_0^*|T,X}(\mathbf{y}|0,\mathbf{x})$ for almost every $\mathbf{x}$ with respect to the distribution of X; (b) Recalling that $Y = \mathbb{1}(T=0)Y_0^* + \mathbb{1}(T=1)Y_1^*$, we have $Y \,|\, (T=0) = Y_0^* \,|\, (T=0)$; (c) By Definition 2, we have $\mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_{(Y|T=0)|(X|T=0)}(\mathbf{y}|\mathbf{x}) = \mathbb{P}_{Y|T,X}(\mathbf{y}|0,\mathbf{x})$. Using these, we have

$$\begin{aligned}\mathbb{P}_{Y_0^*|T}(\mathbf{y}|1) &= \int \mathbb{P}_{Y_0^*|T,X}(\mathbf{y}|1,\mathbf{x}) \,\mathrm{d}\mathbb{P}_{X|T}(\mathbf{x}|1) \overset{(a)}{=} \int \mathbb{P}_{Y_0^*|T,X}(\mathbf{y}|0,\mathbf{x}) \,\mathrm{d}\mathbb{P}_{X|T}(\mathbf{x}|1) \\ &\overset{(b)}{=} \int \mathbb{P}_{Y|T,X}(\mathbf{y}|0,\mathbf{x}) \,\mathrm{d}\mathbb{P}_{X|T}(\mathbf{x}|1) \overset{(c)}{=} \int \mathbb{P}_{Y_0|X_0}(\mathbf{y}|\mathbf{x}) \,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) = \mathbb{P}_{Y\langle 0|1\rangle}(\mathbf{y}),\end{aligned}$$

as required. ■

C.3 Proof of Theorem 5

Proof We prove here $\mathbb{E}[\hat{\mu}_{Y_1^*}] = \mu_{Y_1^*}$; the proof of $\mathbb{E}[\hat{\mu}_{Y_0^*}] = \mu_{Y_0^*}$ is similar and thus omitted. First note that, since $(\mathbf{x}_i, t_i, \mathbf{y}_i)_{i=1}^N$ are i.i.d. with (X,T,Y) and $m := \sum_{i=1}^N t_i$, we have

$$\mathbb{E}[\hat{\mu}_{Y_1^*}] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{N} \frac{t_i \ell(\cdot, \mathbf{y}_i)}{e(\mathbf{x}_i)}\right] = \mathbb{E}\left[\frac{T\ell(\cdot, Y)}{e(X)}\right].$$

By the definition of Y, the conditional exogeneity and the definition of the propensity $e(\mathbf{x})$, we then have

$$\mathbb{E}\left[\frac{T\ell(\cdot, Y)}{e(X)}\right] = \mathbb{E}\left[\frac{T\ell(\cdot, Y_1^*)}{e(X)}\right] = \mathbb{E}_X\left[\mathbb{E}\left[\frac{T\ell(\cdot, Y_1^*)}{e(X)}\middle| X\right]\right]$$

$$= \mathbb{E}_X\left[\frac{\mathbb{E}[T|X]\mathbb{E}[\ell(\cdot,Y_1^*)|X]}{e(X)}\right] = \mathbb{E}_X\left[\frac{e(X)\mathbb{E}[\ell(\cdot,Y_1^*)|X]}{e(X)}\right]$$
$$= \mathbb{E}_X\left[\mathbb{E}[\ell(\cdot,Y_1^*)|X]\right] = \mathbb{E}[\ell(\cdot,Y_1^*)] = \mu_{Y_1^*},$$

where $\mathbb{E}_X$ denotes the expectation with respect to X. ■

C.4 Proof of Theorem 6

.

Proof We derive the convergence rate of $\hat{\mu}_{Y_1^*}$; the rate of $\hat{\mu}_{Y_0^*}$ can be derived in a similar way, and thus is omitted. Let $\tilde{Y}_1^*$ be an independent copy of Y_1^*.

$$\mathbb{E}\left[\left\|\hat{\mu}_{Y_1^*} - \mu_{Y_1^*}\right\|_{\mathscr{F}}^2\right] = \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^N \frac{t_i\ell(\cdot,\mathbf{y}_i)}{e(\mathbf{x}_i)} - \mu_{Y_1^*}\right\|_{\mathscr{F}}^2\right]$$
$$= \underbrace{\mathbb{E}\left[\frac{1}{m^2}\sum_{i,j}\frac{t_it_j\ell(\mathbf{y}_i,\mathbf{y}_j)}{e(\mathbf{x}_i)e(\mathbf{x}_j)}\right]}_{(A)} - 2\underbrace{\mathbb{E}\left[\frac{1}{m}\sum_i \frac{t_i\mathbb{E}_{Y_1^*}[\ell(\mathbf{y}_i,Y_1^*)]}{e(\mathbf{x}_i)}\right]}_{(B)} + \mathbb{E}[\ell(Y_1^*,\tilde{Y}_1^*)]. \quad (43)$$

We first deal with the term (A). It can be expanded as

$$(A) = \frac{1}{m^2}\mathbb{E}\left[\sum_{i\neq j}\frac{t_it_j\ell(\mathbf{y}_i,\mathbf{y}_j)}{e(\mathbf{x}_i)e(\mathbf{x}_j)}\right] + \frac{1}{m^2}\mathbb{E}\left[\sum_i \frac{t_i^2\ell(\mathbf{y}_i,\mathbf{y}_i)}{e^2(\mathbf{x}_i)}\right].$$

Let $(\tilde{X},\tilde{T},\tilde{Y}_0^*,\tilde{Y}_1^*)$ be an independent copy of (X,T,Y_0^*,Y_1^*), and write $\tilde{Y} = \mathbb{1}(\tilde{T}=0)\tilde{Y}_0^* + \mathbb{1}(\tilde{T}=1)\tilde{Y}_1^*$. Since in the first term of (A), $(\mathbf{x}_i,t_i,\mathbf{y}_i)$ and $(\mathbf{x}_j,t_j,\mathbf{y}_j)$ are independent but distributed as (X,T,Y), the first term can be written as

$$\frac{1}{m^2}\mathbb{E}\left[\sum_{i\neq j}\frac{T\tilde{T}\ell(Y,\tilde{Y})}{e(X)e(\tilde{X})}\right] = \frac{m-1}{m}\mathbb{E}\left[\frac{T\tilde{T}\ell(Y,\tilde{Y})}{e(X)e(\tilde{X})}\right].$$

For the right hand side, we have

$$\mathbb{E}\left[\frac{T\tilde{T}\ell(Y,\tilde{Y})}{e(X)e(\tilde{X})}\right] \overset{(a)}{=} \mathbb{E}\left[\frac{T\tilde{T}\ell(Y_1^*,\tilde{Y}_1^*)}{e(X)e(\tilde{X})}\right] = \mathbb{E}_{\tilde{X},\tilde{T},\tilde{Y}_1^*}\left[\frac{\tilde{T}}{e(\tilde{X})}\mathbb{E}_{X,T,Y_1^*}\left[\frac{T\ell(Y_1^*,\tilde{Y}_1^*)}{e(X)} \mid \tilde{Y}_1^*\right]\right]$$
$$= \mathbb{E}_{\tilde{X},\tilde{T},\tilde{Y}_1^*}\left[\frac{\tilde{T}}{e(\tilde{X})}\mathbb{E}_X\left[\frac{\mathbb{E}_{T,Y_1^*}[T\ell(Y_1^*,\tilde{Y}_1^*) \mid X]}{e(X)} \mid \tilde{Y}_1^*\right]\right]$$
$$\overset{(b)}{=} \mathbb{E}_{\tilde{X},\tilde{T},\tilde{Y}_1^*}\left[\frac{\tilde{T}}{e(\tilde{X})}\mathbb{E}_X\left[\frac{\mathbb{E}_T[T \mid X]\mathbb{E}_{Y_1^*}[\ell(Y_1^*,\tilde{Y}_1^*) \mid X]}{e(X)} \mid \tilde{Y}_1^*\right]\right]$$
$$\overset{(c)}{=} \mathbb{E}_{\tilde{X},\tilde{T},\tilde{Y}_1^*}\left[\frac{\tilde{T}}{e(\tilde{X})}\mathbb{E}_X\left[\mathbb{E}_{Y_1^*}[\ell(Y_1^*,\tilde{Y}_1^*) \mid X] \mid \tilde{Y}_1^*\right]\right] = \mathbb{E}_{\tilde{X},\tilde{T},\tilde{Y}_1^*}\left[\frac{\tilde{T}}{e(\tilde{X})}\mathbb{E}_{Y_1^*}[\ell(Y_1^*,\tilde{Y}_1^*)] \mid \tilde{Y}_1^*\right]$$

$$= \mathbb{E}_{\tilde{X}}\left[\frac{1}{e(\tilde{X})}\mathbb{E}_{\tilde{T},\tilde{Y}_1^*}\left[\tilde{T}\mathbb{E}_{Y_1^*}[\ell(Y_1^*,\tilde{Y}_1^*)] \mid \tilde{Y}_1^*\right] \mid \tilde{X}\right] \overset{(d)}{=} \mathbb{E}_{\tilde{X}}\left[\frac{1}{e(\tilde{X})}\mathbb{E}_{\tilde{T}}[\tilde{T}|\tilde{X}]\mathbb{E}_{Y_1^*,\tilde{Y}_1^*}[\ell(Y_1^*,\tilde{Y}_1^*) \mid \tilde{X}]\right]$$
$$\overset{(e)}{=} \mathbb{E}_{\tilde{X}}\left[\mathbb{E}_{Y_1^*,\tilde{Y}_1^*}[\ell(Y_1^*,\tilde{Y}_1^*) \mid \tilde{X}]\right] = \mathbb{E}_{Y_1^*,\tilde{Y}_1^*}[\ell(Y_1^*,\tilde{Y}_1^*)].$$

where (a) follows from the definitions of Y and $\tilde{Y}$, (b) from the conditional exogeneity, (c) from $\mathbb{E}[T \mid X] = e(X)$, (d) from the conditional exogeneity, and (e) from $\mathbb{E}[\tilde{T} \mid \tilde{X}] = e(\tilde{X})$. On the other hand, the second term of (A) can be written as

$$\frac{1}{m^2}\mathbb{E}\left[\sum_i \frac{t_i^2\ell(\mathbf{y}_i\mathbf{y}_i)}{e^2(\mathbf{x}_i)}\right] = \frac{1}{m}\mathbb{E}\left[\frac{T^2\ell(Y,Y)}{e^2(X)}\right].$$

For the last expression, we have

$$\mathbb{E}\left[\frac{T^2\ell(Y,Y)}{e^2(X)}\right] \overset{(a)}{=} \mathbb{E}\left[\frac{T\ell(Y,Y)}{e^2(X)}\right] \overset{(b)}{=} \mathbb{E}\left[\frac{T\ell(Y_1^*,Y_1^*)}{e^2(X)}\right]$$
$$= \mathbb{E}_X\left[\frac{\mathbb{E}_{T,Y_1^*}[T\ell(Y_1^*,Y_1^*) \mid X]}{e^2(X)}\right] \overset{(c)}{=} \mathbb{E}_X\left[\frac{\mathbb{E}_T[T|X]\mathbb{E}_{Y_1^*}[\ell(Y_1^*,Y_1^*) \mid X]}{e^2(X)}\right]$$
$$\overset{(d)}{=} \mathbb{E}_X\left[\frac{\mathbb{E}_{Y_1^*}[\ell(Y_1^*,Y_1^*) \mid X]}{e(X)}\right] \leq \frac{\sup_{\mathbf{y}\in\mathcal{Y}}\ell(\mathbf{y},\mathbf{y})}{\inf_{\mathbf{x}\in\mathcal{X}}e(x)} =: C_{\ell,e} \overset{(e)}{<} \infty,$$

where (a) follows from T taking values in $\{0,1\}$, (b) from Y being Y_1^* if $T=1$, (c) from the conditional exogeneity, (d) from $\mathbb{E}[T|X] = e(X)$, and (e) from our assumptions that $\sup_{\mathbf{y}\in\mathcal{Y}}\ell(\mathbf{y},\mathbf{y}) < \infty$ and $\inf_{\mathbf{x}\in\mathcal{X}}e(x) > 0$. Thus, the term (A) is upper-bounded as

$$(A) \leq \frac{m-1}{m}\mathbb{E}_{Y_1^*,\tilde{Y}_1^*}[\ell(Y_1^*,\tilde{Y}_1^*)] + \frac{1}{m}C_{\ell,e}.$$

Next we deal with the term (B), which can be written as

$$\mathbb{E}\left[\frac{1}{m}\sum_i \frac{t_i\mathbb{E}_{Y_1^*}[\ell(\mathbf{y}_i,Y_1^*)]}{e(\mathbf{x}_i)}\right] = \mathbb{E}\left[\frac{\tilde{T}\mathbb{E}_{Y_1^*}[\ell(\tilde{Y},Y_1^*)]}{e(\tilde{X})}\right],$$

where, as before, $(\tilde{X},\tilde{T},\tilde{Y}_0^*,\tilde{Y}_1^*)$ is an independent copy of (X,T,Y_0^*,Y_1^*) and $\tilde{Y} := \mathbb{1}(\tilde{T}=0)\tilde{Y}_0^* + \mathbb{1}(\tilde{T}=1)\tilde{Y}_1^*$. The right expression can be expanded as

$$\mathbb{E}\left[\frac{\tilde{T}\mathbb{E}_{Y_1^*}[\ell(\tilde{Y},Y_1^*)]}{e(\tilde{X})}\right] \overset{(a)}{=} \mathbb{E}\left[\frac{\tilde{T}\mathbb{E}_{Y_1^*}[\ell(\tilde{Y}_1^*,Y_1^*)]}{e(\tilde{X})}\right] = \mathbb{E}_{\tilde{X}}\left[\frac{\mathbb{E}_{\tilde{T},\tilde{Y}_1^*}\left[\tilde{T}\mathbb{E}_{Y_1^*}[\ell(\tilde{Y}_1^*,Y_1^*)] \mid \tilde{X}\right]}{e(\tilde{X})}\right]$$
$$\overset{(b)}{=} \mathbb{E}_{\tilde{X}}\left[\frac{\mathbb{E}_{\tilde{T}}[\tilde{T} \mid \tilde{X}]\mathbb{E}_{Y_1^*,\tilde{Y}_1^*}[\ell(\tilde{Y}_1^*,Y_1^*)] \mid \tilde{X}]}{e(\tilde{X})}\right] \overset{(c)}{=} \mathbb{E}_{\tilde{X}}\left[\mathbb{E}_{Y_1^*,\tilde{Y}_1^*}[\ell(\tilde{Y}_1^*,Y_1^*)] \mid \tilde{X}]\right] = \mathbb{E}_{Y_1^*,\tilde{Y}_1^*}[\ell(\tilde{Y}_1^*,Y_1^*)],$$

where (a) follows from $\tilde{Y}$ being $\tilde{Y}_1^*$ if $\tilde{T}=1$, (b) from the conditional exogeneity and (c) from $\mathbb{E}_{\tilde{T}}[\tilde{T} \mid \tilde{X}] = e(\tilde{X})$.

Using the obtained results for (A) and (B) in (43), we now have

$$\mathbb{E}\left[\left\|\hat{\mu}_{Y_1^*} - \mu_{Y_1^*}\right\|_{\mathscr{F}}^2\right] \leq \frac{m-1}{m}\mathbb{E}[\ell(Y_1^*,\tilde{Y}_1^*)] + \frac{1}{m}C_{\ell,e} - 2\mathbb{E}[\ell(Y_1^*,\tilde{Y}_1^*)] + \mathbb{E}[\ell(Y_1^*,\tilde{Y}_1^*)]$$

$$= \frac{1}{m}\left(C_{\ell,e} - \mathbb{E}[\ell(Y_1^*, \tilde{Y}_1^*)]\right),$$

which completes the proof. ■

Appendix D. Preliminaries to the Proofs for Section 4

We collect here preliminary results required for proving the theoretical results in Section 4. Thus, the interested reader may first look at Appendix E, where the proofs for the main theoretical results are presented. We use here the notation and basic definitions provided in Section 4 of the main body. In Appendix D.1, we introduce certain integral operators and collect basic facts regarding them. Based on them, in Appendix D.2 we present various lemmas needed for the proofs of the convernce results in Appendix E.

D.1 Integral Operators

To be rigorous, we employ the following notation used in Steinwart and Scovel (2012): For a measurable function $f : \mathcal{X} \to \mathbb{R}$, $[f]_\sim$ denotes the class of measurable functions that are $\mathbb{P}_{X_0}$-equivalent to f:

$$[f]_\sim := \{g : \mathcal{X} \to \mathbb{R} \mid \mathbb{P}_{X_0}(\{x \in \mathcal{X} \mid f(x) \neq g(x)\}) = 0\}.$$

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

$$Tf := \int k(\cdot, \mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) \in L_2(\mathbb{P}_{X_0}), \qquad f \in L_2(\mathbb{P}_{X_0}), \tag{44}$$

$$Sf := \int k(\cdot, \mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) \in \mathscr{H}, \qquad f \in L_2(\mathbb{P}_{X_0}), \tag{45}$$

$$\mathcal{C}_{XX} f := \int k(\cdot, \mathbf{x}) f(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) \in \mathscr{H}, \qquad f \in \mathscr{H}. \tag{46}$$

Note that while these operators look similar, they are different in their domains and ranges. In particular, $\mathcal{C}_{XX}$ is the covariance operator. Under Assumption 2 **(i)**, *i.e.*, $\sup_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x}) < \infty$, Steinwart and Scovel (2012, Lemma 2.3) implies that the operator $S^* : \mathscr{H} \to L_2(\mathbb{P}_{X_0})$ 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 (45). Since S^* is continuous, by Steinwart and Scovel (2012, Lemma 2.3), the operators T and $\mathcal{C}_{XX}$ can be written as

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

The following lemma summarizes conditions required for eigen-decompositions of (44), (45) and (46). In the sequel, "ONS" and "ONB" mean "orthonormal series" and "orthonormal basis," respectively. The set $I \subset \mathbb{N}$ is a set of indices, which is finite if the RKHS $\mathscr{H}$ is finite dimensional, and infinite if $\mathscr{H}$ is infinite dimensional.

Lemma 16 (Spectral decomposition of integral operators) *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) is satisfied. Then there exist at most countable families $(e_i)_{i\in I} \subset \mathscr{H}$ and $(\mu_i)_{i\in I} \subset (0,\infty)$ such that $\mu_1 \geq \mu_2 \geq \cdots > 0$, $(\mu_i^{1/2} e_i)_{i\in I}$ is an ONS in $\mathscr{H}$, $([e_i]_\sim)_{i\in I}$ is an ONS in $L_2(\mathbb{P}_{X_0})$, and*

$$Tf = \sum_{i\in I} \mu_i \langle [e_i]_\sim, f\rangle_{L_2(\mathbb{P}_{X_0})} [e_i]_\sim, \qquad f \in L_2(\mathbb{P}_{X_0}), \tag{47}$$

$$Sf = \sum_{i\in I} \mu_i \langle [e_i]_\sim, f\rangle_{L_2(\mathbb{P}_{X_0})} e_i, \qquad f \in L_2(\mathbb{P}_{X_0}), \tag{48}$$

$$\mathcal{C}_{XX} g = \sum_{i\in I} \mu_i \left\langle \mu_i^{1/2} e_i, g \right\rangle_{\mathscr{H}} \mu_i^{1/2} e_i, \qquad g \in \mathscr{H}, \tag{49}$$

where the convergence is in $L_2(\mathbb{P}_{X_0})$ for (47), and in $\mathscr{H}$ for (48) and (49).

Proof Since k and $\mathbb{P}_{X_0}$ satisfy Assumption 2 **(i)**, it follows from Steinwart and Scovel (2012, Lemma 2.3) that $\mathscr{H}$ is compactly embedded into $L_2(\mathbb{P}_{X_0})$. As a result, Steinwart and Scovel (2012, Lemma 2.12) implies that there exist at most countable families $(e)_{i\in I} \subset \mathscr{H}$ and $(\mu_i)_{i\in I} \subset (0,\infty)$ such that $\mu_1 \geq \mu_2 \geq \cdots > 0$, $([e_i]_\sim)_{i\in I}$ is an ONS in $L_2(\mathbb{P}_{X_0})$, $(\mu_i^{1/2} e_i)_{i\in I}$ is an ONS in $\mathscr{H}$, and (47) holds with convergence in $L_2(\mathbb{P}_{X_0})$.

To show (48), since $([e_i]_\sim)_{i\in I}$ is an ONS in $L_2(\mathbb{P}_{X_0})$, any $f \in L_2(\mathbb{P}_{X_0})$ can be written as

$$f = \sum_{i\in I} \langle [e_i]_\sim, f\rangle_{L_2(\mathbb{P}_{X_0})} [e_i]_\sim + f^\perp,$$

with convergence in $L_2(\mathbb{P}_{X_0})$, where $f^\perp \in L_2(\mathbb{P}_{X_0})$ is such that $\left\langle [e_i]_\sim, f^\perp \right\rangle_{L_2(\mathbb{P}_{X_0})} = 0$ for all $i \in I$. By Steinwart and Scovel (2012, Lemma 2.12, Eq.15) we have $\mu_i e_i = S[e_i]_\sim$ for all $i \in I$. It then holds that

$$Sf = \sum_{i\in I} \mu_i \langle [e_i]_\sim, f\rangle_{L_2(\mathbb{P}_{X_0})} e_i + Sf^\perp,$$

where the convergence is in $\mathscr{H}$ since S is continuous. Note that we have $Tf^\perp = 0$, since we have (47) and $\left\langle [e_i]_\sim, f^\perp \right\rangle_{L_2(\mathbb{P}_{X_0})} = 0$ for all $i \in I$. That is, $f^\perp$ is in the null space of T. Since the null spaces of S and T are equal (Steinwart and Scovel, 2012, Lemma 2.12, Eq.16), it follows that $Sf^\perp = 0$, which implies (48).

Finally we show (49). First note that $\mathcal{C}_{XX} e_i = SS^* e_i = S[e_i]_\sim = \mu e_i$ for all $i \in I$. Using this and (48), for any $g \in \mathscr{H}$ we have

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

where the convergence is in $\mathscr{H}$, which implies (49). ■

Definition 17 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) is satisfied. Let $(e_i)_{i\in I} \subset \mathscr{H}$ and $(\mu_i)_{i\in I} \subset (0,\infty)$ be as in Lemma 16. Then for a constant $\beta > 0$, the β-th power of T, S and $\mathcal{C}_{XX}$ are respectively defined by*

$$T^\beta f := \sum_{i\in I} \mu_i^\beta \langle [e_i]_\sim, f\rangle_{L_2(\mathbb{P}_{X_0})} [e_i]_\sim, \qquad f \in L_2(\mathbb{P}_{X_0}),$$

$$S^\beta f := \sum_{i\in I} \mu_i^\beta \langle [e_i]_\sim, f\rangle_{L_2(\mathbb{P}_{X_0})} e_i, \qquad f \in L_2(\mathbb{P}_{X_0}),$$

$$\mathcal{C}_{XX}^\beta f := \sum_{i\in I} \mu_i^\beta \left\langle \mu_i^{1/2} e_i, f\right\rangle_{\mathscr{H}} \mu_i^{1/2} e_i, \qquad f \in \mathscr{H}.$$

Lemma 18 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) and (ii) hold. Let $(e_i)_{i\in I} \subset \mathscr{H}$ and $(\mu_i)_{i\in I} \subset (0,\infty)$ be as in Lemma 16. Then, $([e]_\sim)_{i\in I}$ is an ONB of $L_2(\mathbb{P}_{X_0})$.*

Proof Since k and $\mathbb{P}_{X_0}$ satisfy Assumption 2 (i), it follows from Steinwart and Scovel (2012, Lemma 2.3) that $\mathscr{H}$ is compactly embedded into $L_2(\mathbb{P}_{X_0})$. Then one can use Steinwart and Scovel (2012, Theorem 3.1), which states that the assertion is equivalent to the Assumption 2 (ii) that the embedding $S^* : \mathscr{H} \to L_2(\mathbb{P}_{X_0})$ has a dense image in $L_2(\mathbb{P}_{X_0})$. ■

We provide a condition for the integral operator T_{prod} defined in Section 4.2 to admit an eigen-decomposition, which is needed for its power T_{prod}^β in (29) to be well-defined.

Lemma 19 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) and (ii) are satisfied. Let $T_{\mathrm{prod}} : L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}) \to L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$ be the integral operator defined as in Section 4.2, and let $(e_i)_{i\in I} \subset \mathscr{H}$ and $(\mu_i)_{i\in I} \subset (0,\infty)$ be as in Lemma 16. Then we have*

$$T_{\mathrm{prod}}\eta = \sum_{i,j\in I} \mu_i\mu_j \langle \eta, [e_i]_\sim \otimes [e_j]_\sim\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} [e_i]_\sim \otimes [e_j]_\sim, \quad \eta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0}),$$

where the convergence is in $L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$.

Proof By Assumption 2 (ii) that S^* has a dense image in $L_2(\mathbb{P}_{X_0})$ and Lemma 18, $([e_i]_\sim)_{i\in I}$ is an ONB in $L_2(\mathbb{P}_{X_0})$. This implies that $([e_i]_\sim \otimes [e_j]_\sim)_{i,j\in I}$ is an ONB in $L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$; see *e.g.*, Folland (1999, Ex. 61, p.178). Therefore any $\eta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$ can be written as

$$\eta = \sum_{i,j\in I} \langle \eta, [e_i]_\sim \otimes [e_j]_\sim\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} [e_i]_\sim \otimes [e_j]_\sim$$

with convergence in $L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})$. Note that $[e_i]_\sim\otimes[e_j]_\sim$ for any $i, j \in I$ is an eigenfunction of T_{prod} with the corresponding eigenvalue being $\mu_i\mu_j$, since

$$\begin{aligned} T_{\mathrm{prod}}([e_i]_\sim \otimes [e_j]_\sim) &= \int k(\cdot,\mathbf{x})[e_i]_\sim(\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) \otimes \int k(\cdot,\tilde{\mathbf{x}})[e_j]_\sim(\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_0}(\tilde{\mathbf{x}}) \\ &= (T[e_i]_\sim) \otimes (T[e_j]_\sim) = \mu_i\mu_i[e_i]_\sim \otimes [e_j]_\sim. \end{aligned}$$

The assertion follows from this and the above eigendecomposition of η in Assumption 4. ■

As a direct corollary of Lemma 19, we have the following result, which provides an eigenbasis expression of the range assumption $\theta \in \mathrm{Range}(T_{\mathrm{prod}}^\beta)$.

Corollary 20 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2* **(i)** *and* **(ii)** *are satisfied, and let $(e_i)_{i\in I} \subset \mathscr{H}$ and $(\mu_i)_{i\in I} \subset (0,\infty)$ be as in Lemma 16. Suppose that Assumption 4 is satisfied,* i.e., *$\theta \in \mathrm{Range}(T^{\beta}_{\mathrm{prod}})$ holds for $\theta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$ and $0 \le \beta \le 1$, where $T^{\beta}_{\mathrm{prod}}$ is defined in (29). Then there exist $(a_{i,j})_{i,j\in I} \subset \mathbb{R}$ such that $\sum_{i,j\in I} a_{i,j}^2 < \infty$ and*

$$\theta = \sum_{i,j\in I} a_{i,j} (\mu_i \mu_j)^{\beta} [e_i]_{\sim} \otimes [e_j]_{\sim},$$

where the convergence is in $L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$.

Proof The assumption $\theta \in \mathrm{Range}(T^{\beta}_{\mathrm{prod}})$ implies that there exists some $\eta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$ such that $\theta = T^{\beta}\eta$. Therefore

$$\theta = \sum_{i,j\in I} (\mu_i \mu_j)^{\beta} \left\langle \eta, [e_i]_{\sim} \otimes [e_j]_{\sim} \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} [e_i]_{\sim} \otimes [e_j]_{\sim}$$

where the convergence is in $L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$. Defining $a_{i,j} := \left\langle \eta, [e_i]_{\sim} \otimes [e_j]_{\sim} \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})}$, we have $\sum_{i,j\in I} a_{i,j}^2 < \infty$ from $\eta \in L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})$, which completes the proof. ■

Lastly, let $\mathcal{C}_{YX} : \mathscr{H} \to \mathscr{F}$ be the covariance operator of the random variables X_0 and Y_0 defined as (see *e.g.*, Fukumizu et al. (2013))

$$\mathcal{C}_{YX} f = \int \ell(\cdot, \mathbf{y}) f(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0 Y_0}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{X_0, Y_0}[\ell(\cdot, Y_0) f(X_0)], \quad f \in \mathscr{H} \tag{50}$$

where $\mathbb{P}_{X_0 Y_0}$ is the joint distribution of X_0 and Y_0. Under Assumption 2 **(i)**, this covariance operator satisfies

$$\langle \mathcal{C}_{YX} f, g \rangle_{\mathscr{F}} = \langle \mathbb{E}_{X_0, Y_0}[\ell(\cdot, X_0) f(X_0)], g \rangle_{\mathscr{F}} = \mathbb{E}_{X_0, Y_0}[g(Y_0) f(X_0)], \quad f \in \mathscr{H}, \quad g \in \mathscr{F}.$$

The conjugate operator of $\mathcal{C}_{YX}$ is denoted by $\mathcal{C}_{XY} : \mathscr{F} \to \mathscr{H}$ and given by

$$\mathcal{C}_{XY} g = \int k(\cdot, \mathbf{x}) g(\mathbf{y}) \, \mathrm{d}\mathbb{P}_{X_0 Y_0}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{X_0, Y_0}[k(\cdot, X_0) g(Y_0)],$$

since for any $f \in \mathscr{H}$ and $g \in \mathscr{F}$ it holds that

$$\langle f, \mathcal{C}_{YX} g \rangle_{\mathscr{H}} = \langle f, \mathbb{E}_{X_0, Y_0}[k(\cdot, X_0) g(Y_0)] \rangle_{\mathscr{H}} = \mathbb{E}_{X_0, Y_0}[f(X_0) g(Y_0)] = \langle \mathcal{C}_{YX} f, g \rangle_{\mathscr{F}}.$$

D.2 Lemmas

We collect here lemmas used in the proofs for the convergence results in Appendix E.

Lemma 21 *Let $\mathcal{X}$, k, $\mathbb{P}_{X_0}$, $\mathbb{P}_{X_1}$ and $g := \mathrm{d}\mathbb{P}_{X_1} / \mathrm{d}\mathbb{P}_{X_0}$ be such that Assumption 2* **(i)** *and* **(iii)** *are satisfied. Then we have $\mu_{X_1} = Sg$.*

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

$$\mu_{X_1} = \int k(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) = \int k(\cdot, \mathbf{x}) g(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) = Sg \in \mathscr{H},$$

where the expression Sg is justified from Assumption 2 **(iii)** that $g \in L_2(\mathbb{P}_{X_0})$. ■

Lemma 22 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) is satisfied. Then for any $f \in L_2(\mathbb{P}_{X_0})$ and $\varepsilon > 0$, we have $S^*(\mathcal{C}_{XX} + \varepsilon I)^{-1} S f = (T + \varepsilon I)^{-1} T f$.*

Proof Let $(e_i)_{i \in I} \subset \mathscr{H}$ and $(\mu_i)_{i \in I} \subset (0, \infty)$ as in Lemma 16. Then we have

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

as required. ■

Lemma 23 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) is satisfied. Then for any $f \in L_2(\mathbb{P}_{X_0})$ and $\alpha \geq 0$, we have $S T^\alpha f = \mathcal{C}_{XX}^{1/2+\alpha} S^{1/2} f$.*

Proof Let $(e_i)_{i \in I} \subset \mathscr{H}$ and $(\mu_i)_{i \in I} \subset (0, \infty)$ as in Lemma 16. Then we have

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

as required. ■

Lemma 24 *Let $\mathcal{X}$, k, $\mathbb{P}_{X_0}$, $\mathbb{P}_{X_1}$ and $g := \mathrm{d}\mathbb{P}_{X_1} / \mathrm{d}\mathbb{P}_{X_0}$ be such that Assumption 2 (i) and (iii) are satisfied. Assume also that Assumption 3 holds, i.e., $g \in \mathrm{Range}(T^\alpha)$ for a constant $\alpha \geq 0$. Then for any $\varepsilon > 0$, we have*

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

where $c_\alpha := \|h\|_{L_2(\mathbb{P}_{X_0})}$ is a constant defined by $h \in L_2(\mathbb{P}_{X_0})$ such that $g = T^\alpha h$.

Proof As in the assertion, write $g = T^\alpha h$ for $h \in L_2(\mathbb{P}_{X_0})$, which exists from the assumption $g \in \mathrm{Range}(T^\alpha)$. By Lemmas 21 and 23, we can then write μ_{X_1} as

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

Since $(\mu_i^{1/2} e_i)_{i \in I}$ is an ONS of $\mathscr{H}$, $([e_i]_\sim)_{i \in I}$ is an ONS in $L_2(\mathbb{P}_{X_0})$ and

$$S^{1/2} h = \sum_{i \in I} \langle [e_i]_\sim, h \rangle_{L_2(\mathbb{P}_{X_0})} \mu_i^{1/2} e_i,$$

it holds that $\|S^{1/2} h\|_{\mathscr{H}}^2 = \sum_{i \in I} \langle [e_i]_\sim, h \rangle_{L_2(\mathbb{P}_{X_0})}^2 \leq \|h\|_{L_2(\mathbb{P}_{X_0})}^2$ and thus $S^{1/2} h \in \mathscr{H}$. Therefore we have

$$\begin{aligned} \left\| (\mathcal{C}_{XX} + \varepsilon I)^{-1} \mu_{X_1} \right\|_{\mathscr{H}} &= \left\| (\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}^{1/2+\alpha} S^{1/2} h \right\|_{\mathscr{H}} \\ &\leq \left\| (\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}^{1/2+\alpha} \right\| \left\| S^{1/2} h \right\|_{\mathscr{H}} \leq \left\| (\mathcal{C}_{XX} + \varepsilon I)^{-1} \mathcal{C}_{XX}^{1/2+\alpha} \right\| \|h\|_{L_2(\mathbb{P}_{X_0})}. \end{aligned}$$

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

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

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

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

This completes the proof. ■

Lemma 25 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) and (ii) are satisfied. Then, for any $g \in L_2(\mathbb{P}_{X_0})$, we have*

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

Proof Let $(e_i)_{i \in I} \subset \mathscr{H}$ and $(\mu_i)_{i \in I} \subset (0, \infty)$ be as in Lemma 16. By Lemma 18, $([e_i]_\sim)_{i \in I}$ is an ONB of $L_2(\mathbb{P}_{X_0})$, which implies that g can be expanded using $([e_i]_\sim)_{i \in I}$. From this and Lemma 16, we then have

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

Thus, by Parseval's identity,

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

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

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

This completes the proof. ■

Lemma 26 *Let $\mathcal{X}$, k and $\mathbb{P}_{X_0}$ be such that Assumption 2 (i) is satisfied. Let $g \in L_2(\mathbb{P}_{X_0})$ be such that $g \in \mathrm{Range}(T^\alpha)$ for a constant $0 \leq \alpha \leq 1$. Then, for all $\varepsilon > 0$, we have*

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

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

Proof Let $(e_i)_{i\in I} \subset \mathscr{H}$ and $(\mu_i)_{i\in I} \subset (0,\infty)$ be as in Lemma 16. From $g \in \mathrm{Range}(T^\alpha)$ there exists $h \in L_2(\mathbb{P}_{X_0})$ such that $g = T^\alpha h$. Therefore g can be written as

$$g = T^\alpha h = \sum_{i\in I}\mu_i^\alpha b_i [e_i]_\sim, \tag{51}$$

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

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

Therefore, by Parseval's identity, we have

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

The rhs of the above equation can be bounded from above 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

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

where the last inequality follows from $([e_i]_\sim)_{i\in I}$ being an ONS in $L_2(\mathbb{P}_{X_0})$. ■

Remark 27 *Different from Lemma 25, Lemma 26 does not require Assumption 2* **(ii)** *that S^* has a dense image in $L_2(\mathbb{P}_{X_0})$. In Lemma 25, this condition is required to guarantee that $([e_i]_\sim)_{i\in I}$ is an ONB in $L_2(\mathbb{P}_{X_0})$, so that g can be expanded by this ONB. On the other hand, in Lemma 26, g can be written as (51), thanks to the assumption $g \in \mathrm{Range}(T^\alpha)$; this is the reason why Lemma 26 does not need Assumption 2* **(ii)**.

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

Lemma 28 *Let $\mathcal{X}$, $\mathcal{Y}$, k, ℓ, $\mathbb{P}_{X_0}$, $\mathbb{P}_{X_1}$, $g := \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$ and $\theta : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ defined in (28) be such that Assumption 2* **(i)** *and* **(iii)** *are satisfied. Then for any $\varepsilon > 0$, we have*

$$\begin{aligned}
&\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|^2_{\mathscr{F}} \\
&= \langle g_\varepsilon \otimes g_\varepsilon, \theta\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} - 2\left\langle g, (T+\varepsilon I)^{-1}T\int \theta(\cdot,\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})\right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&\quad + \iint \theta(\mathbf{x},\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})
\end{aligned}$$

where $g_\varepsilon := (T+\varepsilon I)^{-1}Tg$. In the second term of the right hand side, the inner product is well defined, since we have $(T+\varepsilon_n I)^{-1}T\int\theta(\cdot,\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \in L_2(\mathbb{P}_{X_0})$.

Proof First note that, because ℓ is bounded (Assumption 2 **(i)**), the function θ in (28) satisfies $\theta \in L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})$. Therefore, the right hand side of the assertion is well defined. The left hand side of the assertion can be written as

$$\begin{aligned}
&\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|^2_{\mathscr{F}} \qquad (52)\\
&= \|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon I)^{-1}\mu_{X_1}\|^2_{\mathscr{F}} - 2\left\langle \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon I)^{-1}\mu_{X_1}, \mu_{Y\langle 0|1\rangle}\right\rangle_{\mathscr{F}} + \|\mu_{Y\langle 0|1\rangle}\|^2_{\mathscr{F}}.
\end{aligned}$$

As in the proof of Fukumizu et al. (2013, Thm. 8), the third term in (52) can be written as

$$\|\mu_{Y\langle 0|1\rangle}\|^2_{\mathscr{F}} = \iint \theta(\mathbf{x},\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}). \qquad (53)$$

We thus derive the expressions for the first two terms in (52) in the sequel.

The first term in (52). 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 property of $\mathcal{C}_{YX}$ that $\langle \mathcal{C}_{YX} f, h\rangle_{\mathscr{F}} = \mathbb{E}_{X_0,Y_0}[f(X_0)h(Y_0))]$ for any $h \in \mathscr{F}$ and the expression $\theta(\mathbf{x},\tilde{\mathbf{x}}) = \mathbb{E}_{Y_0,\tilde{Y}_0}[\ell(Y_0,\tilde{Y}_0)|X_0 = \mathbf{x}, \tilde{X}_0 = \tilde{\mathbf{x}}]$, we have

$$\begin{aligned}
\|\mathcal{C}_{YX} f\|^2_{\mathscr{F}} &= \langle \mathcal{C}_{YX} f, \mathcal{C}_{YX} f\rangle_{\mathscr{F}} = \mathbb{E}_{X_0,Y_0}[f(X_0)(\mathcal{C}_{YX} f)(Y_0))] \\
&= \mathbb{E}_{X_0,Y_0}[f(X_0)\mathbb{E}_{\tilde{X}_0,\tilde{Y}_0}[\ell(Y_0,\tilde{Y}_0)f(\tilde{X}_0)]] \\
&= \mathbb{E}_{X_0,\tilde{X}_0}[f(X_0)f(\tilde{X}_0)\mathbb{E}_{Y_0,\tilde{Y}_0}[\ell(Y_0,\tilde{Y}_0)|X_0,\tilde{X}_0]] \quad (\because \text{Fubini theorem}) \\
&= \mathbb{E}_{X_0,\tilde{X}_0}[f(X_0)f(\tilde{X}_0)\theta(X_0,\tilde{X}_0)], \qquad (54)
\end{aligned}$$

where the use of Fubini's theorem is enabled by ℓ and f being bounded, the latter implied by k being bounded. Now define $f := (\mathcal{C}_{XX}+\varepsilon I)^{-1}\mu_{X_1} \in \mathscr{H}$. With this choice of f, the quantity $\|\mathcal{C}_{YX} f\|^2_{\mathscr{F}}$ is equal to the first term in (52). From (54), it follows that

$$\|\mathcal{C}_{YX} f\|^2_{\mathscr{F}} = \mathbb{E}_{X_0,\tilde{X}_0}[f(X_0)f(\tilde{X}_0)\theta(X_0,\tilde{X}_0)]$$

$$
\begin{aligned}
&= \iint f(\mathbf{x}) f(\tilde{\mathbf{x}}) \theta(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\tilde{\mathbf{x}}) = \langle S^* f \otimes S^* f, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \\
&= \left\langle S^* (\mathcal{C}_{XX} + \varepsilon I)^{-1} \mu_{X_1} \otimes S^* (\mathcal{C}_{XX} + \varepsilon I)^{-1} \mu_{X_1}, \theta \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \\
&= \left\langle S^* (\mathcal{C}_{XX} + \varepsilon I)^{-1} S g \otimes S^* (\mathcal{C}_{XX} + \varepsilon I)^{-1} S g, \theta \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \quad (\because \text{Lemma 21}) \\
&= \left\langle (T + \varepsilon I)^{-1} T g \otimes (T + \varepsilon I)^{-1} T g, \theta \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \quad (\because \text{Lemma 22}) \\
&= \langle g_\varepsilon \otimes g_\varepsilon, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})}, \qquad (55)
\end{aligned}
$$

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

The second term in (52). First we have

$$
\begin{aligned}
\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0 = \mathbf{x}] &= \mathbb{E}_{Y_0}\left[\int \mathbb{E}_{\tilde{Y}_0}[\ell(Y_0, \tilde{Y}_0)|\tilde{X}_0 = \tilde{\mathbf{x}}] \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) | X_0 = \mathbf{x} \right] \\
&= \int \mathbb{E}_{Y_0, \tilde{Y}_0}\left[\ell(Y_0, \tilde{Y}_0) | X_0 = \mathbf{x}, \tilde{X}_0 = \tilde{\mathbf{x}} \right] \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \quad (\because \text{Fubini}) \\
&= \int \theta(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \qquad (56)
\end{aligned}
$$

where $(\tilde{X}_0, \tilde{Y}_0)$ is an independent copy of (X_0, Y_0). Note that for the first expression in (56), we have $\mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0 = \cdot] \in L_2(\mathbb{P}_{X_0})$ since ℓ is bounded. Using this and (56), we have

$$
\begin{aligned}
\mathcal{C}_{XY} \mu_{Y\langle 0|1\rangle} &= \mathbb{E}_{X_0, Y_0}[k(\cdot, X_0) \mu_{Y\langle 0|1\rangle}(Y_0)] = \mathbb{E}_{X_0}\left[k(\cdot, X_0) \mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0] \right] \\
&= S \mathbb{E}_{Y_0}[\mu_{Y\langle 0|1\rangle}(Y_0)|X_0 = \cdot] = S \int \theta(\cdot, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}). \qquad (57)
\end{aligned}
$$

Now for the second term in (52), we have

$$
\begin{aligned}
&\left\langle \mathcal{C}_{YX} (\mathcal{C}_{XX} + \varepsilon_n I)^{-1} \mu_{X_1}, \mu_{Y\langle 0|1\rangle} \right\rangle_{\mathscr{F}} = \left\langle \mu_{X_1}, (\mathcal{C}_{XX} + \varepsilon_n I)^{-1} \mathcal{C}_{XY} \mu_{Y\langle 0|1\rangle} \right\rangle_{\mathscr{H}} \\
&= \left\langle S g, (\mathcal{C}_{XX} + \varepsilon_n I)^{-1} S \int \theta(\cdot, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{\mathscr{H}} \quad (\because \text{ Lemma 21 and (57)}) \\
&= \left\langle g, S^* (\mathcal{C}_{XX} + \varepsilon_n I)^{-1} S \int \theta(\cdot, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&= \left\langle g, (T + \varepsilon_n I)^{-1} T \int \theta(\cdot, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(P_{X_0})} \quad (\because \text{Lemma 22}).
\end{aligned}
$$

This completes the proof. ■

Appendix E. Proofs for Section 4

We provide proofs for the convergence results presented in Section 4 of the main paper. The proofs rely on several lemmas collected and proved in Appendix D. The notation and definitions follow those in these sections. In the following, for any bounded linear operator

$A : V \to W$ between normed vector spaces V and W, we denote by $\|A\|$ its operator norm: $\|A\| := \sup_{\|v\|_V \leq 1} \|A(v)\|_W$, where $\|\cdot\|_V$ and $\|\cdot\|_W$ denote the norms of V and W, respectively.

We first show that the CME estimator $\hat{\mu}_{\langle 0|1\rangle}$ in (18) can be expressed in terms of certain empirical covariance operators. Given an i.i.d. sample $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ from $\mathbb{P}_{X_0 Y_0}$, the covariance operators $\mathcal{C}_{XX} : \mathscr{H} \to \mathscr{H}$ in (46) and $\mathcal{C}_{YX} : \mathscr{H} \to \mathscr{F}$ in (50) can be respectively approximated by $\widehat{\mathcal{C}}_{XX} : \mathscr{H} \to \mathscr{H}$ and $\widehat{\mathcal{C}}_{YX} : \mathscr{H} \to \mathscr{F}$, defined as

$$\widehat{\mathcal{C}}_{XX} f := \frac{1}{n} \sum_{i=1}^n k(\cdot, \mathbf{x}_i) f(\mathbf{x}_i), \quad \widehat{\mathcal{C}}_{YX} f = \frac{1}{n} \sum_{i=1}^n \ell(\cdot, \mathbf{y}_i) f(\mathbf{x}_i), \quad f \in \mathscr{H}.$$

Under Assumption 2 **(i)** that the kernels k and ℓ are bounded, these satisfy $\|\widehat{\mathcal{C}}_{XX} - \mathcal{C}_{XX}\| = O_p(n^{-1/2})$ and $\|\widehat{\mathcal{C}}_{YX} - \mathcal{C}_{YX}\| = O_p(n^{-1/2})$ as $n \to \infty$. Similarly, given an i.i.d. sample $(\mathbf{x}'_j)_{j=1}^m$ from $\mathbb{P}_{X_1}$, the kernel mean $\mu_{X_1} := \int k(\cdot, \mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x})$ of $\mathbb{P}_{X_1}$ can be estimated as $\hat{\mu}_{X_1} := \frac{1}{m} \sum_{i=1}^m k(\cdot, \mathbf{x}_j)$, with the error rate $\|\mu_{X_1} - \hat{\mu}_{X_1}\|_{\mathscr{H}} = O_p(m^{-1/2})$ as $m \to \infty$ under Assumption 2 **(i)**.

Proposition 29 *Let $\hat{\mu}_{\langle 0|1\rangle}$ be the CME estimator in (18). Then we have*

$$\hat{\mu}_{\langle 0|1\rangle} = \widehat{\mathcal{C}}_{YX} (\widehat{\mathcal{C}}_{XX} + \varepsilon I)^{-1} \hat{\mu}_{X_1}. \tag{58}$$

Proof Define $g := (\widehat{\mathcal{C}}_{XX} + \varepsilon I)^{-1} \hat{\mu}_{X_1}$. Since $\hat{\mu}_{X_1} = (\widehat{\mathcal{C}}_{XX} + \varepsilon I) g = \frac{1}{n} \sum_{j=1}^n k(\cdot, \mathbf{x}_j) g(\mathbf{x}_j) + \varepsilon g$, we have $\hat{\mu}_{X_1}(\mathbf{x}_\ell) = \frac{1}{n} \sum_{j=1}^n k(\mathbf{x}_\ell, \mathbf{x}_j) g(\mathbf{x}_j) + \varepsilon g(\mathbf{x}_\ell) = \frac{1}{n} (\mathbf{K}\boldsymbol{g})_\ell + \varepsilon \boldsymbol{g}_\ell$ for all $\ell = 1, \ldots, n$, where $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\boldsymbol{g} = (g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n))^\top \in \mathbb{R}^n$. Therefore $\boldsymbol{\mu} = \frac{1}{n} (\mathbf{K} + n\varepsilon \mathbf{I}) \boldsymbol{g}$, where $\boldsymbol{\mu} := (\hat{\mu}_{X_1}(\mathbf{x}_1), \ldots, \hat{\mu}_{X_1}(\mathbf{x}_n))^\top = \widetilde{\mathbf{K}} \mathbf{1}_m$, where $\mathbf{1}_m = (1/m, \ldots, 1/m)^\top$ and $\widetilde{\mathbf{K}} \in \mathbb{R}^{n \times m}$ with $\widetilde{\mathbf{K}}_{ij} = k(\mathbf{x}_i, \mathbf{x}'_j)$. Thus $\boldsymbol{g} = n(\mathbf{K} + n\varepsilon \mathbf{I})^{-1} \boldsymbol{\mu}$. Lastly, the right hand side of (58) can be expressed as $\frac{1}{n} \sum_{i=1}^n \ell(\cdot, \mathbf{y}_i) g(\mathbf{x}_i) = \sum_{i=1}^n \beta_i \ell(\cdot, \mathbf{y}_i)$, where $\beta = (\beta_1, \ldots, \beta_n)^\top = n^{-1} \boldsymbol{g} = (\mathbf{K} + n\varepsilon \mathbf{I})^{-1} \boldsymbol{\mu}$, which is the expression of the CME estimator $\hat{\mu}_{\langle 0|1\rangle}$ in (18). ■

E.1 Convergence Rates of the Stochastic Error

The proofs of Theorems 8 and 13 rely on the following result, which characterizes the "stochastic error" of the CME estimator. As stated in Assumption **(iv)**, we assume $m = n$ in the following.

Theorem 30 *Let $\mathcal{X}$ be a measurable space, k be a measurable kernel on $\mathcal{X}$ and $\mathbb{P}_{X_0}$ be a probability measure on $\mathcal{X}$ such that Assumption 2 **(i)**, **(iii)** and **(iv)** are satisfied. Assume that the Radon-Nikodym derivative $g := \mathrm{d}\mathbb{P}_{X_1} / \mathrm{d}\mathbb{P}_{X_0}$ satisfies $g \in \mathrm{Range}(T^\alpha)$ for a constant $\alpha \geq 0$. Then for any $\varepsilon_n > 0$ such that $\varepsilon_n \to 0$ as $n \to \infty$, we have*

$$\|\widehat{\mathcal{C}}_{YX} (\widehat{\mathcal{C}}_{XX} + \varepsilon_n I)^{-1} \hat{\mu}_{X_1} - \mathcal{C}_{YX} (\mathcal{C}_{XX} + \varepsilon_n I)^{-1} \mu_{X_1}\|_{\mathscr{F}} = O_p\left(n^{-1/2} \varepsilon_n^{\min(-1+\alpha, -1/2)}\right) \qquad (n \to \infty)$$

Proof As in the proof of Fukumizu et al. (2013, Theorem 11), the lhs can be bounded as

$$\|\widehat{\mathcal{C}}_{YX} (\widehat{\mathcal{C}}_{XX} + \varepsilon_n I)^{-1} \hat{\mu}_{X_1} - \mathcal{C}_{YX} (\mathcal{C}_{XX} + \varepsilon_n I)^{-1} \mu_{X_1}\|_{\mathscr{F}}$$

$$\leq \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}(\hat{\mu}_{X_1}-\mu_{X_1})\|_{\mathscr{F}}+\|(\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{F}}$$
$$+\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}(\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{F}} \tag{59}$$

By Baker (1973, Theorem 1), $\widehat{\mathcal{C}}_{YX}$ 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 $\widehat{\mathcal{W}}_{YX}:\mathscr{H}\to\mathscr{F}$ with $\|\widehat{\mathcal{W}}_{YX}\|\leq 1$, where $\widehat{\mathcal{C}}_{YY}^{1/2}:\mathscr{F}\to\mathscr{F}$ and $\widehat{\mathcal{C}}_{XX}^{1/2}:\mathscr{H}\to\mathscr{H}$ are such that $\widehat{\mathcal{C}}_{YY}=\widehat{\mathcal{C}}_{YY}^{1/2}\widehat{\mathcal{C}}_{YY}^{1/2}$ and $\widehat{\mathcal{C}}_{XX}=\widehat{\mathcal{C}}_{XX}^{1/2}\widehat{\mathcal{C}}_{XX}^{1/2}$. Therefore,

$$\begin{aligned}\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\| &= \|\widehat{\mathcal{C}}_{YY}^{1/2}\widehat{\mathcal{W}}_{YX}\widehat{\mathcal{C}}_{XX}^{1/2}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\| \\ &\leq \|\widehat{\mathcal{C}}_{YY}^{1/2}\|\ \|(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1/2}\|\leq \|\widehat{\mathcal{C}}_{YY}^{1/2}\|\varepsilon_n^{-1/2}.\end{aligned} \tag{60}$$

Thus, the rate of the first term in (59) is

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

Next, the rate of the second term in (59) is given by

$$\begin{aligned}\|(\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{F}} &\leq \|\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX}\|\|(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{H}} \\ &\leq \|\widehat{\mathcal{C}}_{YX}-\mathcal{C}_{YX}\|c_\alpha\varepsilon_n^{\min(-1/2+\alpha,0)} \quad (\because \text{Lemma 24}) \\ &= O_p\left(n^{-1/2}\varepsilon_n^{\min(-1/2+\alpha,0)}\right),\end{aligned}$$

where c_α is a constant depending only on α and g. Finally, for the third term in (59), the rate is given as

$$\begin{aligned}&\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}(\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX})(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{F}} \\ &\leq \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\|\|\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX}\|\|(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{H}} \\ &\leq \|\widehat{\mathcal{C}}_{YY}^{1/2}\|\varepsilon_n^{-1/2}\|\mathcal{C}_{XX}-\widehat{\mathcal{C}}_{XX}\|c_\alpha\varepsilon_n^{\min(-1/2+\alpha,0)} \quad (\because (60) \text{ and Lemma 24}) \\ &= O_p\left(n^{-1/2}\varepsilon_n^{\min(-1+\alpha,-1/2)}\right).\end{aligned}$$

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 (59). This completes the proof. ■

E.2 Proof of Theorem 8

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

$$\begin{aligned}&\left\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1}-\mu_{Y\langle 0|1\rangle}\right\|_{\mathscr{F}} \\ &\leq \left\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1}-\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\right\|_{\mathscr{F}} \qquad (61) \\ &\quad +\left\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}-\mu_{Y\langle 0|1\rangle}\right\|_{\mathscr{F}}, \qquad (62)\end{aligned}$$

where (61) can be interpreted as the stochastic error and (62) as the approximation error. Note that the assumption $g\in L_2(\mathbb{P}_{X_0})$ enables the use of Theorem 30 with $\alpha=0$, which

implies that the estimation error (61) converges to 0 at rate $O_p(n^{-1/2}\varepsilon_n^{-1})$ as $n \to \infty$, provided that $\varepsilon_n \to 0$ and $n^{1/2}\varepsilon_n \to \infty$ as $n \to \infty$.

Here we aim to prove that the approximation error (62) goes to zero as $\varepsilon_n \to 0$. Note that to this end, we cannot apply the proof of Theorem 8 in Fukumizu et al. (2013), since it relies on stronger assumptions than ours. We do this by using Lemma 28, which shows that the approximation error can be written as

$$\begin{aligned}\left\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}-\mu_{Y\langle 0|1\rangle}\right\|_{\mathscr{F}}^2 &= \langle g_{\varepsilon_n}\otimes g_{\varepsilon_n},\theta\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} && (63)\\ &\quad -2\left\langle g,(T+\varepsilon_n I)^{-1}T\int\theta(\cdot,\tilde{x})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})\right\rangle_{L_2(\mathbb{P}_{X_0})} && (64)\\ &\quad +\iint\theta(\mathbf{x},\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}),\end{aligned}$$

where $g_{\varepsilon_n} := (T+\varepsilon_n I)^{-1}Tg$ with $g = \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$ being the Radon-Nikodym derivative. Below we show the convergence limits of (63) and (64) as $\varepsilon_n \to 0$, which conclude the proof.

Convergence of (63). We will show that

$$\langle g_{\varepsilon_n}\otimes g_{\varepsilon_n},\theta\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} \to \iint\theta(\mathbf{x},\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \quad (\varepsilon_n\to 0). \tag{65}$$

Note that we have

$$\langle g\otimes g,\theta\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} = \iint g(\mathbf{x})g(\tilde{\mathbf{x}})\theta(\mathbf{x},\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_0}(\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_0}(\tilde{\mathbf{x}}) = \iint\theta(\mathbf{x},\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\mathbf{x})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}).$$

Therefore it suffices to show that

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

Note that by the Cauchy-Schwartz inequality, we have

$$\begin{aligned}&\left|\langle g_{\varepsilon_n}\otimes g_{\varepsilon_n},\theta\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}-\langle g\otimes g,\theta\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}\right|\\ =\ &\left|\langle g_{\varepsilon_n}\otimes g_{\varepsilon_n}-g\otimes g,\theta\rangle_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}\right| \le \|g_{\varepsilon_n}\otimes g_{\varepsilon_n}-g\otimes g\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}\,\|\theta\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}.\end{aligned}$$

Thus we focus on showing that

$$\|g_{\varepsilon_n}\otimes g_{\varepsilon_n}-g\otimes g\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}\to 0 \quad (\varepsilon_n\to 0). \tag{66}$$

By the triangle inequality we have

$$\begin{aligned}&\|g_{\varepsilon_n}\otimes g_{\varepsilon_n}-g\otimes g\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}\\ &\le\ \|g_{\varepsilon_n}\otimes g_{\varepsilon_n}-g\otimes g_{\varepsilon_n}\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}+\|g\otimes g_{\varepsilon_n}-g\otimes g\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}.\end{aligned} \tag{67}$$

The first term of (67) can be written as

$$\begin{aligned}\|g_{\varepsilon_n}\otimes g_{\varepsilon_n}-g\otimes g_{\varepsilon_n}\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} &= \|(g_{\varepsilon_n}-g)\otimes g_{\varepsilon_n}\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}\\ &= \|g_{\varepsilon_n}-g\|_{L_2(\mathbb{P}_{X_0})}\|g_{\varepsilon_n}\|_{L_2(\mathbb{P}_{X_0})}\to 0 \quad (\varepsilon_n\to 0) \quad (\because \text{Lemma 25}),\end{aligned}$$

Similarly, the second term of (67) can be written as

$$\begin{aligned}\|g\otimes g_{\varepsilon_n}-g\otimes g\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})} &= \|g\otimes(g_{\varepsilon_n}-g)\|_{L_2(\mathbb{P}_{X_0}\otimes\mathbb{P}_{X_0})}\\ &= \|g\|_{L_2(\mathbb{P}_{X_0})}\|g_{\varepsilon_n}-g\|_{L_2(\mathbb{P}_{X_0})}\to 0 \quad (\varepsilon_n\to 0) \quad (\because \text{Lemma 25}).\end{aligned}$$

We have shown (66), which concludes (65).

Convergence of (64). We show that as $\varepsilon_n \to 0$,

$$\left\langle g, (T+\varepsilon_n I)^{-1} T \int \theta(\cdot, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(\mathbb{P}_{X_0})} \to \iint \theta(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}). \tag{68}$$

From Lemma 25, as $\varepsilon_n \to 0$, the lhs converges to

$$\begin{aligned} \left\langle g, \int \theta(\cdot, \tilde{x}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(\mathbb{P}_{X_0})} &= \iint \theta(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) g(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) \\ &= \iint \theta(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}). \end{aligned}$$

Thus we have shown (68). The proof completes by substituting (65) and (68) in (63) and (64) respectively. ■

E.3 Proof of Theorem 13

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

$$\begin{aligned} &\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathscr{F}} \\ &\quad\leq \|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{F}} \qquad (69) \\ &\qquad + \|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathscr{F}}, \qquad (70) \end{aligned}$$

where (69) is the estimation error, and (70) is the approximation error. By Theorem 30, the estimation error decays at the rate

$$\|\widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX}+\varepsilon_n I)^{-1}\hat{\mu}_{X_1} - \mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}\|_{\mathscr{F}} = O_p\left(n^{-1/2}\varepsilon_n^{\min(-1+\alpha,-1/2)}\right) \tag{71}$$

as $n \to \infty$. Hence we focus below on deriving a convergence rate for the approximation error. We then determine the optimal schedule for the decay of the regularization constant ε_n as $n \to \infty$ in order to derive a convergence rate for the overall error.

Rate for the approximation error (70). We will show that the approximation error decays at the rate

$$\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathscr{F}} = O\left(\varepsilon_n^{(\alpha+\beta)/2}\right) \quad (\varepsilon_n \to 0). \tag{72}$$

First note that, by the definition of $g = \mathrm{d}\mathbb{P}_{X_1}/\mathrm{d}\mathbb{P}_{X_0}$, we have

$$\begin{aligned} \iint \theta(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) &= \iint \theta(\mathbf{x}, \tilde{\mathbf{x}}) g(\mathbf{x}) g(\tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_0}(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_0}(\tilde{\mathbf{x}}) \\ &= \langle g \otimes g, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})}. \end{aligned} \tag{73}$$

Therefore, using Lemma 28 and the notation $g_{\varepsilon_n} := (T+\varepsilon_n I)^{-1} T g$, we can bound the square of the approximation error (70) as

$$\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1} - \mu_{Y\langle 0|1\rangle}\|_{\mathscr{F}}^2$$

$$
\begin{aligned}
&= \langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} - 2 \left\langle g, (T + \varepsilon_n I)^{-1} T \int \theta(\cdot, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&\quad + \iint \theta(\mathbf{x}, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \\
&\leq \left| \langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} - \langle g \otimes g, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \right| \qquad (74) \\
&\quad + 2 \left| \langle g \otimes g, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} - \left\langle g, (T + \varepsilon_n I)^{-1} T \int \theta(\cdot, \tilde{\mathbf{x}}) \, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(\mathbb{P}_{X_0})} \right|,
\end{aligned}
$$

Bound on the first term in (74). By Corollary 20 (which follows from Assumption 4), $\theta = \sum_{i,j \in I} a_{i,j} (\mu_i^\beta [e_i]_\sim) \otimes (\mu_j^\beta [e_j]_\sim)$ with $\sum_{i,j \in I} a_{ij}^2 < \infty$. Using this, we have

$$
\begin{aligned}
&\langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} - \langle g \otimes g, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \\
&\quad = \left\langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \sum_{i,j \in I} a_{i,j} (\mu_i^\beta [e_i]_\sim) \otimes (\mu_j^\beta [e_j]_\sim) \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \\
&\qquad - \left\langle g \otimes g, \sum_{i,j \in I} a_{i,j} (\mu_i^\beta [e_i]_\sim) \otimes (\mu_j^\beta [e_j]_\sim) \right\rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \\
&\quad = \sum_{i,j \in I} a_{i,j} \left\langle g_{\varepsilon_n}, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \left\langle g_{\varepsilon_n}, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&\qquad - \sum_{i,j \in I} a_{i,j} \left\langle g, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \left\langle g, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&\quad = \sum_{i,j \in I} a_{i,j} \left\langle g_{\varepsilon_n} - g, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \left\langle g_{\varepsilon_n}, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&\qquad + \sum_{i,j \in I} a_{i,j} \left\langle g, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \left\langle g_{\varepsilon_n} - g, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left| \langle g_{\varepsilon_n} \otimes g_{\varepsilon_n}, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} - \langle g \otimes g, \theta \rangle_{L_2(\mathbb{P}_{X_0} \otimes \mathbb{P}_{X_0})} \right| \\
&\quad \leq \left| \sum_{i,j \in I} a_{i,j} \left\langle g_{\varepsilon_n} - g, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \left\langle g_{\varepsilon_n}, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \right| \\
&\qquad + \left| \sum_{i,j \in I} a_{i,j} \left\langle g, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \left\langle g_{\varepsilon_n} - g, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \right| \\
&\quad \leq \sqrt{\sum_{i,j \in I} a_{i,j}^2} \sqrt{\sum_{i \in I} \left\langle g_{\varepsilon_n} - g, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})}^2 \sum_{j \in I} \left\langle g_{\varepsilon_n}, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})}^2} \\
&\qquad + \sqrt{\sum_{i,j \in I} a_{i,j}^2} \sqrt{\sum_{i \in I} \left\langle g, \mu_i^\beta [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})}^2 \sum_{j \in I} \left\langle g_{\varepsilon_n} - g, \mu_j^\beta [e_j]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})}^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{\sum_{i,j\in I} a_{i,j}^2}\left\|T^\beta(g_{\varepsilon_n}-g)\right\|_{L_2(\mathbb{P}_{X_0})}\left\|T^\beta g_{\varepsilon_n}\right\|_{L_2(\mathbb{P}_{X_0})} \\
&\quad + \sqrt{\sum_{i,j\in I} a_{i,j}^2}\left\|T^\beta g\right\|_{L_2(\mathbb{P}_{X_0})}\left\|T^\beta(g_{\varepsilon_n}-g)\right\|_{L_2(\mathbb{P}_{X_0})} \\
&= \sqrt{\sum_{i,j\in I} a_{i,j}^2}\left\|T^\beta(g_{\varepsilon_n}-g)\right\|_{L_2(\mathbb{P}_{X_0})}\left(\left\|T^\beta g_{\varepsilon_n}\right\|_{L_2(\mathbb{P}_{X_0})}+\left\|T^\beta g\right\|_{L_2(\mathbb{P}_{X_0})}\right) \qquad (75)
\end{aligned}
$$

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 26 (which can be used because $\alpha+\beta \le 1$), we have

$$
\left\|T^\beta g_{\varepsilon_n} - T^\beta g\right\|_{L_2(\mathbb{P}_{X_0})} = \left\|(T+\varepsilon_n I)^{-1}TT^\beta g - T^\beta g\right\|_{L_2(\mathbb{P}_{X_0})} \le c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta}, \qquad (76)
$$

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

$$
\begin{aligned}
\left\|T^\beta g_{\varepsilon_n}\right\|_{L_2(\mathbb{P}_{X_0})} &\le \left\|T^\beta g_{\varepsilon_n} - T^\beta g\right\|_{L_2(\mathbb{P}_{X_0})} + \left\|T^\beta g\right\|_{L_2(\mathbb{P}_{X_0})} \\
&\le c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta} + \left\|T^\beta g\right\|_{L_2(\mathbb{P}_{X_0})} \quad (\because (76))
\end{aligned}
$$

Therefore (75), and thus the first term in (74), is bounded by

$$
\sqrt{\sum_{i,j\in I} a_{i,j}^2}\, c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta}\left(c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta} + 2\|T^\beta g\|_{L_2(\mathbb{P}_{X_0})}\right). \qquad (77)
$$

Bound on the second term in (74). From the equivalence (73), (the half of) the second term in (74) can be written as

$$
\left|\left\langle g, \int \theta(\cdot,\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})\right\rangle_{L_2(\mathbb{P}_{X_0})} - \left\langle g, (T+\varepsilon_n I)^{-1}T\int \theta(\cdot,\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})\right\rangle_{L_2(\mathbb{P}_{X_0})}\right|. \qquad (78)
$$

Note that, using $\theta = \sum_{i,j\in I} a_{i,j}(\mu_i^\beta[e_i]_\sim)\otimes(\mu_j^\beta[e_j]_\sim)$, we can write

$$
\begin{aligned}
\int \theta(\cdot,\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) &= \int \sum_{i,j\in I} a_{i,j}(\mu_i^\beta[e_i]_\sim)\otimes(\mu_j^\beta[e_j]_\sim(\tilde{\mathbf{x}}))\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \\
&= \sum_{i,j\in I} a_{i,j}(\mu_i^\beta[e_i]_\sim)\int(\mu_j^\beta[e_j]_\sim(\tilde{\mathbf{x}}))\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\left\langle g, \int \theta(\cdot,\tilde{\mathbf{x}})\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})\right\rangle_{L_2(\mathbb{P}_{X_0})} &= \left\langle g, \sum_{i,j\in I} a_{i,j}(\mu_i^\beta[e_i]_\sim)\int(\mu_j^\beta[e_j]_\sim(\tilde{\mathbf{x}}))\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})\right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&= \sum_{i,j\in I} a_{i,j}\mu_i^\beta\left\langle g,[e_i]_\sim\right\rangle_{L_2(\mathbb{P}_{X_0})}\int(\mu_j^\beta[e_j]_\sim(\tilde{\mathbf{x}}))\,\mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}).
\end{aligned}
$$

Similarly,

$$\begin{aligned}
&\left\langle g, (T+\varepsilon_n I)^{-1} T \int \theta(\cdot, \tilde{\mathbf{x}})\, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&= \left\langle (T+\varepsilon_n I)^{-1} T g, \sum_{i,j\in I} a_{i,j} (\mu_i^\beta [e_i]_\sim) \int (\mu_j^\beta [e_j]_\sim(\tilde{\mathbf{x}}))\, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right\rangle_{L_2(\mathbb{P}_{X_0})} \\
&= \sum_{i,j\in I} a_{i,j} \mu_i^\beta \left\langle (T+\varepsilon_n I)^{-1} T g, [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \int (\mu_j^\beta [e_j]_\sim(\tilde{\mathbf{x}}))\, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}})
\end{aligned}$$

Because of the properties that $\mu_1 \geq \mu_2 \geq \cdots > 0$, that $([e_j]_\sim)_{j\in I}$ is an ONS in $L_2(\mathbb{P}_{X_0})$ and that $g \in L_2(\mathbb{P}_{X_0})$, we have

$$\begin{aligned}
&\sum_{j\in I} \left(\int (\mu_j^\beta [e_j]_\sim(\tilde{\mathbf{x}}))\, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right)^2 \leq \mu_1^{2\beta} \sum_{j\in I} \left(\int [e_j]_\sim(\tilde{\mathbf{x}})\, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right)^2 \\
= \;& \mu_1^{2\beta} \sum_{j\in I} \left(\int [e_j]_\sim(\tilde{\mathbf{x}}) g(\tilde{\mathbf{x}})\, \mathrm{d}\mathbb{P}_{X_0}(\tilde{\mathbf{x}}) \right)^2 = \mu_1^{2\beta} \sum_{j\in I} \langle [e_j]_\sim, g \rangle_{L_2(\mathbb{P}_{X_0})}^2 \leq \mu_1^{2\beta} \|g\|_{L_2(\mathbb{P}_{X_0})}^2 < \infty.
\end{aligned}$$

Therefore, using the Cauchy-Schwartz, the above identities and inequality, we have

$$\begin{aligned}
(78) &= \left| \sum_{i,j\in I} a_{i,j} \mu_i^\beta \left\langle g - (T+\varepsilon_n I)^{-1} T g, [e_i]_\sim \right\rangle_{L_2(\mathbb{P}_{X_0})} \int (\mu_j^\beta [e_j]_\sim(\tilde{\mathbf{x}}))\, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right| \\
&\leq \sqrt{\sum_{i,j\in I} a_{i,j}^2} \sqrt{\sum_{i\in I} \mu_i^{2\beta} \langle g - (T+\varepsilon_n I)^{-1} T g, [e_i]_\sim \rangle_{L_2(\mathbb{P}_{X_0})}^2 \sum_{j\in I} \left(\int (\mu_j^\beta [e_j]_\sim(\tilde{\mathbf{x}}))\, \mathrm{d}\mathbb{P}_{X_1}(\tilde{\mathbf{x}}) \right)^2} \\
&\leq \sqrt{\sum_{i,j\in I} a_{i,j}^2} \mu_1^\beta \|g\|_{L_2(\mathbb{P}_{X_0})} \sqrt{\sum_{i\in I} \mu_i^{2\beta} \langle g - (T+\varepsilon_n I)^{-1} T g, [e_i]_\sim \rangle_{L_2(\mathbb{P}_{X_0})}^2} \\
&\leq \sqrt{\sum_{i,j\in I} a_{i,j}^2} \mu_1^\beta \|g\|_{L_2(\mathbb{P}_{X_0})} \left\| T^\beta \left(g - (T+\varepsilon_n I)^{-1} T g \right) \right\|_{L_2(\mathbb{P}_{X_0})},
\end{aligned}$$

where the last inequality follows from $([e_j]_\sim)_{j\in I}$ being an ONS in $L_2(\mathbb{P}_{X_0})$ and the definition of T^β. Note that we have $T^\beta g \in \mathrm{Range}(T^{\alpha+\beta})$ from the assumption $g \in \mathrm{Range}(T^\alpha)$. Therefore by Lemma 26,

$$\left\| T^\beta \left(g - (T+\varepsilon_n I)^{-1} T g \right) \right\|_{L_2(\mathbb{P}_{X_0})} = \|T^\beta g - (T+\varepsilon_n)^{-1} T T^\beta g\|_{L_2(\mathbb{P}_{X_0})} \leq c_{\alpha+\beta}\, \varepsilon_n^{\alpha+\beta}$$

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

$$(78) \leq \sqrt{\sum_{i,j\in I} a_{i,j}^2} \mu_1^\beta \|g\|_{L_2(\mathbb{P}_{X_0})} c_{\alpha+\beta}\, \varepsilon_n^{\alpha+\beta}. \tag{79}$$

Resulting approximation error rate. Using (77) and (79) in (74), the rate (72) is finally obtained as

$$\begin{aligned}
&\|\mathcal{C}_{YX}(\mathcal{C}_{XX}+\varepsilon_n I)^{-1}\mu_{X_1}-\mu_{Y\langle 0|1\rangle}\|_{\mathscr{F}}^2 \\
&\leq \sqrt{\sum_{i,j\in I} a_{i,j}^2}c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta}\left(c_{\alpha+\beta}\varepsilon_n^{\alpha+\beta}+2\|T^{\beta}g\|_{L_2(\mathbb{P}_{X_0})}+\mu_1^{\beta}\|g\|_{L_2(\mathbb{P}_{X_0})}\right) \\
&= O(\varepsilon_n^{\alpha+\beta}) \quad (\varepsilon_n\to 0).
\end{aligned}$$

Balancing the estimation and approximation error rates. For an arbitrary constant $c>0$ independent of n, let $\varepsilon_n = cn^{-b}$ for some constant $b>0$. We determine b by balancing the two rates (71) and (72). 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 $0\leq\alpha\leq 1$. The proof completes by substituting the resulting $\varepsilon_n = n^{-b}$ in (71) and (72). ■

References

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3427–3435. Curran Associates Inc., 2017.

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

Onur Atan, James Jordon, and Mihaela van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *AAAI Conference on Artificial Intelligence*, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning (ICML2012)*, pages 1359–1366, 2012.

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

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

Karsten Borgwardt, Arthur Gretton, Malte Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Found. Comput. Math. J.*, 7(4):331–368, 2007.

Claes M. Cassel, Carl E. Särndal, and Jan H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 621–630. ACM, 2009.

Y. Chen, M. Welling, and A. Smola. Super samples from kernel-herding. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press, 2010.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

J. Dick, F. Y. Kuo, and I. H. Sloan. High dimensional numerical integration - the Quasi-Monte Carlo way. *Acta Numerica*, 22(133-288), 2013.

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

N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. Wiley, 2000.

G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 132–141. AUAI Press, 2014.

Miroslav Dudík, John Langford, and Lihong Li. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1993.

R.A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications, 2nd Edition*. Wiley, 1999.

K Fukumizu, A Gretton, X Sun, and B Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496. Curran Associates, Inc., 2008.

Kenji Fukumizu, Francis Bach, and Michael Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Thomas Gärtner. A survey of kernels for structured data. *SIGKDD Explor. Newsl.*, 5(1): 49–58, 2003.

Marc Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2002.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Steffen Grünewälder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1823–1830, New York, NY, USA, 2012. Omnipress.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt 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.

James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics*, 6:4779–4874, 2007.

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

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

Paul Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

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

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 2004.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, New York, NY, USA, 2015.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems 29*, pages 181–189. Curran Associates, Inc., 2016.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 3020–3029, 2016.

Takafumi Kajihara, Motonobu Kanagawa, Keisuke Yamazaki, and Kenji Fukumizu. Kernel recursive ABC: Point estimation with intractable likelihood. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2400–2409. PMLR, 2018.

Nathan Kallus. A framework for optimal matching for causal inference. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 372–381, 2017.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1243–1251, 2018.

M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu. Filtering with state-observation examples via kernel monte carlo filter. *Neural Computation*, 28(2):382–444, 2016.

S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Proc. AISTATS 2015*, 2015.

John Langford, Alexander Strehl, and Jennifer Wortman. Exploration scavenging. In *Proceedings of the 25th International Conference on Machine Learning*, pages 528–535, 2008.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya 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.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pages 41–50. PMLR, 2020a.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2020b. Forthcoming.

Jerzy Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. Master's thesis, 7 1923. Excerpts reprinted in English, Statistical Science, Vol. 5, pp. 463–472. (D. M. Dabrowska, and T. P. Speed, Translators.).

Yu Nishiyama, Motonobu Kanagawa, Arthur Gretton, and Kenji Fukumizu. Model-based kernel sum rule: kernel bayesian inference with probabilistic models. *Machine Learning*, pages 1–34, 2020.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, page 759–766, 2000.

Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

Paul Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 2002.

Paul Rosenbaum and Donald Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

C. Rothe. Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, 155:56–70, 2010.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

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

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1670–1679, 2016.

Bernhard Schölkopf and Alexander Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2002.

Uri Shalit, Fredrik Johansson, and David Sontag. Bounding and minimizing counterfactual error. arXiv:1606.03976 Preprint, 2016.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3076–3085. PMLR, 2017.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems 32*, pages 4593–4605. Curran Associates, Inc., 2019.

S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard 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.

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

Le Song, Jonathan Huang, Alex Smola, and Kenji 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.

Le Song, Kenji Fukumizu, and Arthur 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.

Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

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

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.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8: 985–1005, 2007.

Dougal Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

Adith Swaminathan and Thorsten 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.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems 30*, pages 3632–3642. Curran Associates, Inc., 2017.

Ilya Tolstikhin, Bharath Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18:86:1–86:47, 2017.

Alexandre Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

Masatoshi Uehara, Masahiro Kato, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

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

Raymond K W Wong and Kwun Chuen Gary Chan. Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213, 2017.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.