Journal of Machine Learning Research 27 (2026) 1-48 Submitted 7/23; Revised 6/25; Published 1/26

Flexible Functional Treatment Effect Estimation

Jiayi Wang JIAYI.WANG2@UTDALLAS.EDU
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, TX 75080, USA

Raymond K. W. Wong RAYWONG@TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843, USA

Xiaoke Zhang XKZHANG@GWU.EDU
Department of Statistics
The George Washington University
Washington, DC 20052, USA

Kwun Chuen Gary Chan KCGCHAN@UW.EDU
Department of Biostatistics
University of Washington
Seattle, WA 98195, USA

Editor: Eric Laber

Abstract

We study treatment effect estimation with functional treatments where the average potential outcome functional is a function of functions, in contrast to continuous treatment effect estimation where the target is a function of real numbers. By considering a flexible scalar-on-function marginal structural model, a weight-modified kernel ridge regression (WMKRR) is adopted for estimation. The weights are constructed by directly minimizing the uniform balancing error resulting from a decomposition of the WMKRR estimator, instead of being estimated under a particular treatment selection model. Despite the complex structure of the uniform balancing error derived under WMKRR, finite-dimensional convex algorithms can be applied to efficiently solve for the proposed weights thanks to a representer theorem. The optimal convergence rate is shown to be attainable by the proposed WMKRR estimator without any smoothness assumption on the true weight function. Corresponding empirical performance is demonstrated by a simulation study and a real data application.

Keywords: Covariate balancing; functional data analysis; functional regression; reproducing kernel Hilbert space

1. Introduction

It is well known that for observational studies where a treatment is not randomly assigned, the estimation of average potential outcomes or contrasts (such as average treatment effects) is challenging due to possible confoundedness. This work focuses on estimating the treatment effect with a *functional* treatment, in contrast to the vast majority of existing

work that focuses on binary and continuous treatments. For motivational purposes, we present a few examples as follows. To investigate the causal effect of temperature patterns in a year on crop yields in the following harvest season, one could use daily maximum and daily minimum temperature trajectories as functional treatments (Wong et al., 2019). To assess human visceral adipose tissue, Zhang et al. (2021) used body shape as a functional treatment and studied its causal effect on the tissue. In addition, biomedical researchers may be interested in the causal effect of the activity profile on certain health indicators, such as body mass index and waist circumference, which are potential indicators of obesity level (Neovius et al., 2005). The activity pattern of an individual can be recorded by a tracker during a certain time period, e.g., as in the Physical Activity Monitor data from the National Health and Nutrition Examination Survey (NHANES) in 2005-2006. One could transform a trajectory of activity intensity into a distribution of the intensity values (re-spresented by kernel mean embedding (Muandet et al., 2017)) and take it as a functional treatment. See Example 2 and Section 5.2 for more details.

There exists extensive literature on the estimation of average treatment effect (ATE) for binary treatments, which is well summarized by several review papers (e.g., Imbens, 2004; Stuart, 2010; Ding and Li, 2018; Yao et al., 2021). Extensions to multi-level categorical treatments (e.g., Yang et al., 2016; Lopez and Gutman, 2017; Li, 2019) and continuous treatments (e.g., Robins et al., 2000; Hirano and Imbens, 2004; Imai and van Dyk, 2004; Imai and Ratkovic, 2014; Galvao and Wang, 2015; Zhu et al., 2015; Kennedy et al., 2017; Fong et al., 2018; Li et al., 2020; Bahadori et al., 2022) are also abundant. Although there is practical interest in the causal effect of functional variables, existing methods that can be applied directly to functional treatments are scarce. To the best of our knowledge, only three related works (Zhao et al., 2018; Zhang et al., 2021; Tan et al., 2022) are devoted to estimating the causal effects of functional treatments using observational data, but each has limitations that we seek to address. We will contrast our proposed method with these works in detail below. We note that functional variables are sometimes treated as confounders (e.g., McKeague and Qian, 2014; Laber and Staicu, 2018; Ciarleglio et al., 2018; Miao et al., 2022) and outcomes (Zhao et al., 2018; Lin et al., 2023). Such settings are fundamentally different from ours, where the functional variables are treatments, accompanied with vector-valued covariates and a real-valued outcome.

With an unconfoundedness assumption, two modeling strategies are commonly adopted to estimate causal effects. One is the outcome regression approach, which first estimates an outcome regression model by treating both the treatments and confounders as predictors, and then averages the regression prediction on a fixed treatment value over the observed covariate distribution. The functional linear model, the most common scalar-on-function model, is often employed for this purpose (Zhao et al., 2018). However, due to the limited flexibility in characterizing the outcome given the functional treatment and multivariate covariates using the functional linear model, inconsistent estimation can arise from model misspecifications. Although some complex scalar-on-function regression models are available, model misspecification still pose a major risk for causal effect estimation. The other approach is based on estimating weights that directly address the selection bias of functional treatments. Compared to the regression approach, weighting methods try to mimic randomized experiments that theoretically balance on all pre-treatment-assignment variables, and do not involve the direct use of outcome data in constructing weights. As a result,

they help to preserve the objectivity of the analysis and avoid data snooping (Rubin, 2007; Rosenbaum et al., 2010). We adopt the weighting approach in this paper.

One of the biggest challenges in weighting for causal effect estimation of functional treatments is that, unlike discrete or continuous variables, the density of functional treatment is not well established due to its intrinsically infinite dimension (Delaigle and Hall, 2010). This phenomenon also prevents a direct adaptation of existing estimators for continuous treatment effects which often requires estimating density functions. To overcome this issue, we define a weight function through reverse conditioning that is well-defined for functional treatments and properly adjusts for the selection bias. We also propose a novel estimation approach that directly computes weights via the idea of covariate balancing. Covariate balancing has recently become a popular approach in causal inference for observational studies due to its advantage of providing a stable estimation of weights. For example, covariate balancing methods have been developed for binary treatments (e.g., Hainmueller, 2012; Imai and Ratkovic, 2014; Qin and Zhang, 2007; Zubizarreta, 2015; Wong and Chan, 2018; Wang and Zubizarreta, 2020) and continuous treatments (e.g., Fong et al., 2018; Kallus and Santacatterina, 2019; Tübbicke, 2022), and for estimating conditional treatment effects (e.g., Wang et al., 2022).

Zhang et al. (2021) and Tan et al. (2022) also adopt the idea of covariate balancing to estimate the causal effect of functional treatments by weighting, but with several weaknesses. The approach by Zhang et al. (2021) relies on a *finite* approximation of the functional treatment by truncating the tail part of its functional principal components. The weights are estimated by balancing the correlation between the covariates and the top functional principal components, which is directly generalized from Fong et al. (2018) to handle continuous treatments. This approach has several drawbacks. First, there is likely information loss in selecting only several top functional principal components. Second, only balancing the *correlation* may not be enough to guarantee the consistency of the final causal effect estimator unless the true outcome regression has a certain simple parametric form in the selected top functional principal components. Furthermore, the theoretical properties of their approach are not studied. Instead of using an approximation of the functional treatment, Tan et al. (2022) construct functional stabilized weights by balancing a set of growing number of basis functions. However, they focus on a functional linear marginal structural model, which imposes additional structure to the causal effect which may well be misspecified. Compared with Zhang et al. (2021) and Tan et al. (2022), our proposed covariate balancing method is distinct in the following aspects. (1) Our proposed method does not rely on any finite approximation of the functional treatment. (2) The balancing weights are constructed to directly balance the difference between the final causal effect estimator and the true target function. (3) We do not require a linear functional marginal structural model. (4) Our estimator attains the optimal rate of convergence under mild conditions. We will further elaborate on these points as follows.

Inspired by the development in nonparametric functional regression under the framework of reproducing kernel Hilbert space (RKHS) (e.g., Kadri et al., 2010; Zhang et al., 2012; Oliva et al., 2015; Szabó et al., 2016; Kadri et al., 2016), we adopt the RKHS modeling for the functional treatment effects, which allows for a great flexibility (compared to a functional linear marginal structural model) in characterizing the effect of different functional treatments. In particular, we assume that the marginal structural model lies in

an RKHS of functions with a functional input. With the help of the closed-form solution of weight-modified kernel ridge regression (WMKRR), we then propose balancing weights that are capable of controlling the balancing error between a smoothed weighted average and the population mean of functionals in a large hypothesis class (see (11) for the explicit form). We will then show that the solution to the optimization objective lies in a finite-dimensional space by a representer theorem and that the resulting optimization is convex. The theoretical analysis of the balancing error is not a trivial generalization from that for the binary treatment effects, since the balancing structure is significantly more complicated due to the interplay between weighting and smoothing in the formulation, and the balancing error is a function with a functional input instead of a scalar. Furthermore, while ignored by Zhang et al. (2021) and Tan et al. (2022), a functional treatment is often not fully observable in practice and thus requires some pre-processing steps for recovery, which creates another layer of complication in the theoretical analysis. We provide a careful and detailed theoretical analysis to deal with all these complications. Asymptotic properties of the proposed estimator are derived under the complex dependency structure of the weights and kernel ridge regression. Under appropriate technical conditions, we are able to show that the proposed causal effect estimator can achieve the optimal nonparametric convergence rate, without additional modeling assumptions on the true weight function.

The rest of the paper is organized as follows. Section 2.1 provides the basic setup of the weight-modified kernel ridge regression. Section 2.2 introduces the construction of covariate balancing weights for functional treatments. Computational details and an algorithm to construct the proposed balancing weights are presented in Section 3. Section 4 develops the asymptotic properties of the proposed weighted estimator. The numerical performance of the proposed method is demonstrated by a simulation study in Section 5.1 and an application to a physical activity tracking data set in Section 5.2. The code is publicly available via the Github link: `https://github.com/jiayiwang1017/FFTEE`.

2. Functional Treatment Effect Estimation

2.1 Background and motivation

Let $A \in \mathcal{A}$ be a functional treatment defined on $\mathcal{T} \subset \mathbb{R}^d$, where d is the dimension of the input of the function A, and $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ be a p-dimensional confounder. Denote by $Y(a) \in \mathbb{R}$ the potential outcome had treatment a was given, for $a \in \mathcal{A}$. Suppose that $\{(A_i, \boldsymbol{X}_i, Y_i(\cdot)) : i = 1, \ldots, n\}$ are independent and identically distributed copies of $(A, \boldsymbol{X}, Y(\cdot))$. In practice, we do not observe all the potential outcomes per subject. In fact, only one particular case is observed. The observed outcome for the i-th subject is $Y_i := Y_i(A_i)$. As such, the available data is $\{(A_i, \boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$. The goal is to estimate the functional treatment effect (FTE) $\tau : \mathcal{A} \to \mathbb{R}$ defined by

$$\tau(a) := \mathbb{E}\{Y(a)\}, \qquad a \in \mathcal{A}. \tag{1}$$

Note that the domain of τ is a possibly infinite-dimensional function space $\mathcal{A}$, and therefore (non-parametric) estimations of τ are significantly harder than typical continuous treatment estimations. Throughout the paper, we impose the following two assumptions.

Assumption 1 (Weak unconfoundedness) *Let $D(a)$ be the indicator of receiving treatment a: $D(a) = 1$ if $A = a$; $D(a) = 0$ otherwise. We have*

$$Y(a) \perp\!\!\!\perp D(a) \mid \boldsymbol{X}, \text{ for any } a \in \mathcal{A}.$$

This assumption is the weak unconfoundedness assumption introduced in Imbens (2000), which only requires the pairwise independence of the treatment with each of the potential outcomes. It is less restrictive than the strong ignorability assumption (Rosenbaum and Rubin, 1983): $\{Y(a), a \in \mathcal{A}\} \perp\!\!\!\perp A \mid \boldsymbol{X}$, since weak unconfoundedness only requires the independence of the potential outcome $Y(a)$ and the treatment to be *local* at the treatment level of interest, *i.e.* $D(a)$ (Imbens, 2000).

Take ρ_U as the marginal distribution of a random object U, and $\rho_{U|V}$ as the conditional distribution of U given V. Define the weight function $w^* : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ as

$$w^*(a, \boldsymbol{x}) := \frac{\rho_{\boldsymbol{X}}(\boldsymbol{x})}{\rho_{\boldsymbol{X}|A}(\boldsymbol{x} \mid a)}, \tag{2}$$

which can be used to adjust the dependence between the treatment A and the confounder $\boldsymbol{X}$. In (2), we use densities of the covariates $\boldsymbol{X}$ instead of the densities of the treatment A, as is commonly done in ATE methods (e.g., Wong and Chan, 2018; Kennedy et al., 2017). As such, we are able to circumvent the challenge of establishing functional densities (Delaigle and Hall, 2010). However, this makes the direct estimation of $w^*(a, \boldsymbol{x})$ difficult as multivariate joint density of X appears in both the numerator and the denominator. Based on the definition of w^*, one can observe that

$$\mathbb{E}\{w^*(A, \boldsymbol{X})Y \mid A = a\} = \mathbb{E}\left[w^*(a, \boldsymbol{X})\mathbb{E}(Y \mid A = a, \boldsymbol{X}) \mid A = a\right] = \mathbb{E}\{Y(a)\}. \tag{3}$$

See detailed derivation in Appendix B.

Remark 1 *Indeed, based on the definition of w^*, one has $\mathbb{E}\{w^*(A, \boldsymbol{X})u(A, \boldsymbol{X}) \mid A = a\} = \mathbb{E}_{\boldsymbol{X} \sim \rho_{\boldsymbol{X}}}\{u(A, \boldsymbol{X}) \mid A = a\}$ for any function $u \in \mathbb{R}^{\mathcal{A} \times \mathcal{X}}$ such that the expectations exist and are finite.*

Intuitively, this means that the weight function w^* helps to adjust the conditional expectation of the observed outcome Y so that the weight-modified outcome $Z := w^*(A, \boldsymbol{X})Y$ is an *unbiased* observation of the treatment effect. This indicates that one can regress Z on A to recover τ. In addition to Assumption 1, we also require the following overlap condition:

Assumption 2 (Overlap) *There exists a positive constant C_1 such that $w^*(a, \boldsymbol{x}) \leq C_1$ for all $a \in \mathcal{A}, \boldsymbol{x} \in \mathcal{X}$.*

This assumption plays a similar role as the standard positivity assumption of the conditional treatment density under the settings of continuous treatments (e.g., Assumption 2 in Kennedy et al., 2017). To significantly relax this assumption, one typically requires a strong outcome regression model (as a function of x and a) that allows extrapolation to no-chance or low-chance regions in $\mathcal{A} \times \mathcal{X}$. We avoid making strong outcome regression modeling assumptions like functional linear models, at the expense of a strong overlap assumption.

If the true weights $w_i^* := w^*(A_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, are known and the functional treatments A_i, $i = 1, \ldots, n$, are fully observed, we can construct the adjusted outcomes $Z_i = w_i^* Y_i$

such that $\mathbb{E}(Z_i \mid A_i) = \tau(A_i)$ for every i, and then perform a regression over the data $\{(A_i, Z_i), i = 1, \ldots, n\}$ to estimate τ. Here we consider a nonparametric regression model that allows for a flexible modeling for τ. In particular, we assume that τ lies in an RKHS $\mathcal{H}_A$ with a reproducing kernel $K_A(\cdot, \cdot)$, where $K_A : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, with the corresponding inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_A}$ and norm $\| \cdot \|_{\mathcal{H}_A}$ respectively.

Remark 2 *Note that the reproducing kernel K_A is a bivariate function with function inputs. Based on an additional assumption that $\mathcal{A}$ is a Hilbert space, we provide some concrete examples of K_A that are easy to implement in practice. Recall that its inner product and norm are denoted by $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ and $\| \cdot \|_{\mathcal{A}}$ respectively. The linear kernel $K_A(a_1, a_2) = \langle a_1, a_2 \rangle_{\mathcal{A}} + c$ with a constant c is probably the simplest example. Its corresponding RKHS $\mathcal{H}_A$ contains linear functions of the form $\beta_0 + \langle \beta_1, \cdot \rangle_{\mathcal{A}}$, where $\beta_0 \in \mathbb{R}$ and $\beta_1 \in \mathcal{A}$. As for nonlinear kernels, examples include the Gaussian kernel with $K_A(a_1, a_2) = \exp\{-2\|a_1 - a_2\|^2_{\mathcal{A}}/\theta\}$ and exponential kernel with $K_A(a_1, a_2) = \exp\{-\|a_1 - a_2\|_{\mathcal{A}}/\theta\}$ respectively for some pre-specified $\theta > 0$.*

We consider the weight-modified kernel ridge regression (WMKRR) estimator for τ:

$$\tilde{\tau} := \underset{\tau \in \mathcal{H}_A}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(w_i^* Y_i - \tau(A_i) \right)^2 + \lambda \|\tau\|^2_{\mathcal{H}_A}, \tag{4}$$

where $\lambda > 0$ is a tuning parameter of the regularization. In (4), the norm $\| \cdot \|_{\mathcal{H}_A}$ in the penalty term measures the "roughness" of the underlying mapping, and therefore encourages a "smoother" solution to (4) as λ increases. We also allow the functional treatments A_i to be not fully observed, but densely observed (with noise). In such cases, A_i can be replaced with the estimated functions $\hat{A}_i$ in (4). Further discussion on this will be provided in Sections 2.2 and 4.

Unfortunately, the true weights w_i^*, $i = 1, \ldots, n$, are typically unknown in observational studies. A natural solution is to first directly estimate w_i^* based on its definition, and then construct the adjusted outcomes based on these estimates. This has been extensively studied in ATE estimation when the treatment A is a binary random variable (e.g., Feng et al., 2012; Hirano et al., 2003). However, this approach has several drawbacks. First, from the definition of w^* in (2), the estimation of w^* involves estimating the densities of $\boldsymbol{X}$ and $\boldsymbol{X} \mid A$. Even if one uses a *finite* approximation of A (see Remark 3 below), their estimations are still challenging. This is because to make Assumption 1 plausible, $\boldsymbol{X}$ should include all the confounders that affect both the treatment and outcome, so it is usually multivariate. This indicates that estimating multivariate density functions is required, which is known to be difficult: Parametric estimations of multivariate density functions have a risk of possible model misspecifications, while nonparametric estimations such as the kernel density estimation suffer from slow rates of convergence in multivariate settings. Second, the true weights are expected to achieve the balance for covariates in expectation. However, it is unclear if such balance is enough for finite samples, especially when the sample size is small and the covariates are sparse (Zubizarreta et al., 2011). Third, the inverse of the densities can result in instability, especially when the estimated densities are close to zero.

To overcome the aforementioned problems, we consider finding a stable set of weights that mimic the role of w_i^*, $i = 1, \ldots, n$, through the idea of covariate balancing. There

exists extensive literature on covariate balancing techniques for ATE estimation when the treatment is binary (e.g. Hainmueller, 2012; Imai and Ratkovic, 2014; Qin and Zhang, 2007; Zubizarreta, 2015; Wong and Chan, 2018; Wang and Zubizarreta, 2020) or continuous (e.g. Fong et al., 2018; Kallus and Santacatterina, 2019; Tübbicke, 2022), while we consider covariate balancing for the challenging setup where the treatment is functional.

Remark 3 *Zhang et al. (2021) introduce functional propensity scores that are based on the functional principal components (FPCs) of A and define balancing weights based on the lower-order FPCs. Then the input of the weight function becomes finite-dimensional, which resembles the setting of multivariate continuous treatments. However, this definition relies on the unsupervised dimension reduction of the process A and has the risk of missing important information. For example, if the higher-order FPC scores are more correlated with the potential outcome than the lower-order ones, their proposed weight that only depends on the latter may not properly account for all confounding. In contrast, our method to be shown below does not rely on such unsupervised truncation of A so it can avoid the information loss mentioned above.*

Remark 4 *Tan et al. (2022) introduce a functional stabilized weight (FSW) estimator. They consider a functional linear marginal structural model $\tau(a) = \alpha + \int_{\mathcal{T}} \beta(t)a(t)\mathrm{d}t$ for some scalar α and function β, which is restrictive and subject to the risk of model misspecifications. Instead of directly estimating the weight function w^*, they estimate its projection $w^*(a, \cdot) : \mathcal{X} \to \mathbb{R}$ for every fixed $a \in \mathcal{A}$ by attempting to maintain the covariate balance $\mathbb{E}\{w^*(A, \boldsymbol{X})b(\boldsymbol{X}) \mid A = a\} = \mathbb{E}\{b(\boldsymbol{X})\}$ for any integrable function b. A Nadaraya-Watson estimator is proposed to approximate the left-hand side of the equation. To obtain the sequence of estimated weights, they have to perform the optimization n times, separately for each observation. The convergence of their estimated weights depends on the smoothness of the projection $w^*(a, \cdot)$. In contrast, as shown below, our method does not require the restrictive linearity assumption above for the marginal structural model. Moreover, our proposed weights are calculated jointly via a single optimization, and we do not require any smoothness assumption for the function w^*. Furthermore, the final weighted causal effect estimator can achieve the optimal rate of convergence with a nonparametric modeling of τ,* i.e., *without assuming a linear functional marginal structural model on τ.*

2.2 Construction of weights

To motivate our construction, first suppose we have obtained a set of adjusted weights $\boldsymbol{w} = [w_1, \ldots, w_n]^\intercal$. From (4), we form an estimator of the treatment effect:

$$\hat{\tau}_{\boldsymbol{w}} := \operatorname*{argmin}_{\tau \in \mathcal{H}_A} \frac{1}{n} \sum_{i=1}^{n} \left(w_i Y_i - \tau(A_i)\right)^2 + \lambda \|\tau\|^2_{\mathcal{H}_A}. \tag{5}$$

Recall that $\mathcal{H}_A$ is the RKHS with the reproducing kernel K_A. We give the following definition which will be useful in expressing and analyzing the solution of (5).

Definition 1 *For $a \in \mathcal{A}$, $\mathcal{K}_a : \mathbb{R} \to \mathcal{H}_A$ is a Hilbert-Schmidt operator such that*

$$f(a) = \mathcal{K}_a^* f = \langle K_A(a, \cdot), f \rangle_{\mathcal{H}_A},$$

where $\mathcal{K}_a^*$ is the adjoint of $\mathcal{K}_a$. Define the operator $\mathcal{S}_a := \mathcal{K}_a\mathcal{K}_a^*$. Note that we have

$$\mathcal{S}_a : \mathcal{H}_A \to \mathcal{H}_A, \qquad (\mathcal{S}_a f)(\cdot) = f(a)K_A(a,\cdot) \text{ for any } f \in \mathcal{H}_A.$$

With Definition 1, the estimator (5) can be rewritten as:

$$\hat{\tau}_{\boldsymbol{w}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}w_iY_i\right), \tag{6}$$

where $\mathcal{I} : \mathcal{H}_A \to \mathcal{H}_A$ is an identity operator such that $\mathcal{I}f = f$ for any $f \in \mathcal{H}_A$. See, e.g., Caponnetto and De Vito (2007) and Smale and Zhou (2007) for more details.

Define $m(a,\boldsymbol{x}) = \mathbb{E}\{Y(a) \mid A = a, \boldsymbol{X} = \boldsymbol{x}\} = \mathbb{E}\{Y(a) \mid \boldsymbol{X} = \boldsymbol{x}\}$. We can then express

$$Y_i(a) = m(a, \boldsymbol{X}_i) + \epsilon_i(a), \qquad i = 1, \ldots, n, \tag{7}$$

where $\epsilon_i(a) = Y_i(a) - m(a,\boldsymbol{X}_i)$ satisfies $\mathbb{E}[\epsilon_i(a) \mid A_i = a, \boldsymbol{X}_i] = \mathbb{E}[\epsilon_i(a) \mid \boldsymbol{X}_i] = 0$. This allows the error to be heteroskedastic with respect to the functional treatment and other covariates, and leads to $\tau(a) = \mathbb{E}_{X\sim\rho_X}\{m(a,X)\}$. We assume that $\mathbb{E}[\epsilon_i^2(a) \mid \boldsymbol{X}_i] \le \sigma_0^2 < \infty$ for some constant $\sigma_0 > 0$ (not depending on $\boldsymbol{X}_i$, a and i). As $(Y_i(a), \boldsymbol{X}_i)$, $i = 1, \ldots, n$, are i.i.d., so are $\epsilon_i(a)$, $i = 1, \ldots, n$. According to (7), the observed data follow

$$Y_i = Y_i(A_i) = m(A_i, \boldsymbol{X}_i) + \epsilon_i(A_i) = m(A_i, \boldsymbol{X}_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{8}$$

where we write $\epsilon_i = \epsilon(A_i)$ for short. Clearly, $\mathbb{E}(\epsilon_i \mid A_i = a, \boldsymbol{X}_i) = \mathbb{E}(\epsilon_i(a) \mid A_i = a, \boldsymbol{X}_i) = \mathbb{E}(\epsilon_i(a) \mid \boldsymbol{X}_i) = 0$ and, similarly, $\mathbb{E}(\epsilon_i^2 \mid A_i = a, \boldsymbol{X}_i) \le \sigma_0^2$. As such, $\mathbb{E}(\epsilon_i \mid A_i, \boldsymbol{X}_i) = 0$, $\mathbb{E}(\epsilon_i^2 \mid A_i, \boldsymbol{X}_i) \le \sigma_0^2$. Following (8), we can decompose the difference between $\hat{\tau}_{\boldsymbol{w}}$ and τ as:

$$\hat{\tau}_{\boldsymbol{w}} - \tau = \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}w_iY_i\right) - \tau = I_1 + I_2,$$

$$\text{where} \quad I_1 = \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}w_im(A_i,\boldsymbol{X}_i)\right) - \tau, \tag{9}$$

$$\text{and} \quad I_2 = \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}w_i\epsilon_i\right). \tag{10}$$

Apparently the estimation error of $\hat{\tau}_{\boldsymbol{w}}$ can be bounded by properly controlling the magnitudes of I_1 in (9) and I_2 in (10). Roughly speaking, term (10) exhibits concentration (at zero) due to the independence among ϵ_i's conditional on the treatments and covariates. This will be rigorously shown in our theoretical analysis. The primary challenge lies in controlling (9) since m is unknown in practice. To address this, motivated by Wong and Chan (2018), Kallus and Santacatterina (2019) and Wang et al. (2022), we assume that m belongs to a certain class of functions and control (9) for every element in this class.

Explicitly, we assume that m lies in a tensor-product RKHS $\mathcal{H} := \mathcal{H}_A \otimes \mathcal{H}_X$ of functions defined on $\mathcal{A}\times\mathcal{X}$. Here $\mathcal{H}_X$ is an RKHS of functions defined on $\mathcal{X}$, with a reproducing kernel $K_X(\cdot,\cdot)$, where $K_X : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, and its corresponding inner product and norm of $\mathcal{H}_X$ are denoted by $\langle\cdot,\cdot\rangle_{\mathcal{H}_X}$ and $\|\cdot\|_{\mathcal{H}_X}$ respectively. Note that the assumption $m \in \mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_X$ is compatible with the aforementioned assumption $\tau \in \mathcal{H}_A$ due to the following proposition.

Proposition 1 *Under Assumption 5 in Section 4, if* $\sup_{u\in\mathcal{H}_X}|\mathbb{E}u(\boldsymbol{X})|\neq 0$, *we have* $\mathcal{H}_A=\{\mathbb{E}_{\boldsymbol{X}\sim\rho_X}g(\cdot,\boldsymbol{X}):g\in\mathcal{H}\}$.

To bound the magnitude of (9), we aim to find weights $\tilde{\boldsymbol{w}}=[\tilde{w}_1,\ldots,\tilde{w}_n]^\intercal$ such that

$$\Upsilon:=\sup_{u\in\mathcal{H}:\|u\|_{\mathcal{H}}\leq 1}\left\|\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{A_i}+\lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n\mathcal{K}_{A_i}\tilde{w}_iu(A_i,\boldsymbol{X}_i)\right)-\mathbb{E}_{\boldsymbol{X}\sim\rho_{\boldsymbol{X}}}u(\cdot,\boldsymbol{X})\right\| \quad (11)$$

is minimized with respect to some norm $\|\cdot\|$. Note that the objective (i.e., the norm) of the supremum (11) is proportional to $\|u\|_{\mathcal{H}}$. Therefore, we limit the space to $\mathcal{H}(1)=\{u\in\mathcal{H}:\|u\|_{\mathcal{H}}\leq 1\}$. Since $\tau(a)=\mathbb{E}_{\boldsymbol{X}\sim\rho_{\boldsymbol{X}}}m(a,\boldsymbol{X})$ and $m\in\mathcal{H}$, we have

$$\left\|\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{A_i}+\lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n\mathcal{K}_{A_i}\tilde{w}_im(A_i,\boldsymbol{X}_i)\right)-\tau\right\|\leq\Upsilon\|m\|_{\mathcal{H}}.$$

Since $\|m\|_{\mathcal{H}}<\infty$, bounding (11) provides a good control over the discrepancy (9) even if we do not know the true outcome function m.

While $\tilde{\boldsymbol{w}}$ is well motivated, the criterion in (11) is not directly applicable for the following reasons First, $\mathbb{E}_{\boldsymbol{X}\sim\rho_{\boldsymbol{X}}}u(\cdot,\boldsymbol{X})$ is usually unavailable since the distribution of $\boldsymbol{X}$ is unknown. Thus we propose to replace it with its empirical counterpart $\sum_{i=1}^n u(\cdot,\boldsymbol{X}_i)/n$. Second, the norm $\|\cdot\|$ in (11) needs to be chosen. A natural choice is $\mathcal{L}_2(\mathcal{A})$-norm $\|\cdot\|_{\mathcal{L}_2}$ defined by $\|f\|_{\mathcal{L}_2}=\sqrt{\mathbb{E}\{f^2(A)\}}$ for a function $f:\mathcal{A}\to\mathbb{R}$. In practice, we will use the empirical norm $\|\cdot\|_n$, which is defined by $\|f\|_n:=\sqrt{\sum_{k=1}^n f^2(A_k)/n}$. Finally, the functional treatments are often not fully observed in practice so we will need to recover A_i by $\hat{A}_i$. Two examples of $\hat{A}_i$ will be given in Examples 1 and 2 below.

By the above discussion, we use the following criterion for controlling (9):

$$Q(\boldsymbol{w},\lambda,u):=\left\|\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{\hat{A}_i}+\lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n\mathcal{K}_{\hat{A}_i}w_iu(\hat{A}_i,\boldsymbol{X}_i)\right)-\frac{1}{n}\sum_{j=1}^n u(\cdot,\boldsymbol{X}_j)\right\|_n^2. \quad (12)$$

In addition, to simultaneously control the second moment of (10), we introduce the following regularization term of the weights:

$$R(\boldsymbol{w},\lambda)=\frac{1}{n^2}\sum_{i=1}^n w_i^2\left\|\left(\frac{1}{n}\sum_{j=1}^n\mathcal{S}_{\hat{A}_j}+\lambda\mathcal{I}\right)^{-1}\mathcal{K}_{\hat{A}_i}\right\|_n^2. \quad (13)$$

Combining (12) and (13), we define the proposed balancing weights as

$$\hat{\boldsymbol{w}}=[\hat{w}_1,\ldots,\hat{w}_n]^\intercal:=\operatorname*{argmin}_{0\leq w_i\leq L,i=1,\ldots,n}\left[\sup_{u\in\mathcal{H}(1)}\{Q(\boldsymbol{w},\lambda,u)\}+\eta R(\boldsymbol{w},\lambda)\right], \quad (14)$$

where $\mathcal{H}(1)=\{u\in\mathcal{H}:\|u\|_{\mathcal{H}}\leq 1\}$, $\eta\geq 0$ is a tuning parameter and L is an upper bound for the estimated weights and is allowed to be infinity (which means that the domain of the optimization is $0\leq w_i<\infty, i=1,\ldots,N$). In sequel, we write $\hat{\tau}$ in short for $\hat{\tau}_{\hat{\boldsymbol{w}}}$ when $\hat{\boldsymbol{w}}$ is computed from (14).

Before we conclude this section, we provide some examples of $\hat{A}_i$ to recover A_i in practice, $i=1,\ldots,n$.

Example 1 ***Densely observed trajectories***. *For every $A_i \in \mathcal{A}$, $i = 1, \ldots, n$, its noisy observations $\{\gamma_{i,j} : j = 1, \ldots, N\}$ are measured at $\{t_{i,j} : j = 1, \ldots, N\}$, a dense grid of $\mathcal{T}$. More specifically, $\gamma_{i,j} = A_i(t_{i,j}) + \varepsilon_{i,j}$, where $\mathbb{E}(\varepsilon_{i,j}) = 0$. Under certain assumptions on the grid points, e.g., $t_{i,j}$, $j = 1, \ldots, N$, are i.i.d. copies of a random variable T with density ρ_T, common nonparametric regression procedures can be applied to each individual i to obtain $\hat{A}_i$, such as penalized spline regression (e.g., Eilers and Marx, 1996; Claeskens et al., 2009), smoothing spline regression (e.g., Rice and Rosenblatt, 1983; Gu and Gu, 2013) and other pre-smoothing procedures (e.g., Zhang and Chen, 2007; Miao et al., 2023).*

Example 2 ***Kernel mean embedding of distributions***. *Alternatively, we often observe exposure measurements $s_{il}, l = 1, \ldots, N$, from a distribution P_i (e.g., the activity profile as mentioned in Section 1). In this case, one would like to define functional treatment A_i using a functional representation of P_i. A popular choice is the kernel mean embedding of the distribution (e.g., Muandet et al., 2017), which is a functional representation of a distribution like probability density function (PDF) and cumulative distribution function (CDF). But, unlike PDF and CDF, kernel mean embeddings resides in a Hilbert space without any additional manifold structures such as positivity, sum-to-one constraint or monotonicity, and can be readily used as a functional treatment. More specifically, for the i-th individual, the kernel mean embedding A_i, $i = 1, \ldots, n$, is defined by*

$$A_i = \int_{\mathcal{T}} K_e(\cdot, s) \mathrm{d}P_i(s),$$

where P_i is the corresponding distribution function, and $K_e : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a reproducing kernel. Given observations $s_{il} \sim P_i$, $l = 1, \ldots, N$, we can take $\hat{A}_i$ as the empirical embedding

$$\hat{A}_i = \frac{1}{N} \sum_{l=1}^{N} K_e(\cdot, s_{i,l}).$$

3. Computational Details

In this section, we discuss the computation for (14).

3.1 Representer theorem, closed-form expression and convexity

3.1.1 Inner optimization

First, let us focus on the inner optimization of (14), i.e., $\sup_{u \in \mathcal{H}(1)} Q(\boldsymbol{w}, \lambda, u)$. Note that it is an infinite-dimensional optimization problem when $\mathcal{H}$ is infinite dimensional. The practical optimization relies on the following representer theorem, which shows that the solution indeed lies in a finite-dimensional space given data.

Theorem 1 (Representer theorem) *The solution to $\sup_{u \in \mathcal{H}(1)} Q(\boldsymbol{w}, \lambda, u)$ lies in the finite-dimensional space*

$$\mathcal{H}_n := \left\{ \sum_{i=1}^{n} \alpha_i K_A(\cdot, \hat{A}_i) K_X(\cdot, \boldsymbol{X}_i) + \sum_{i=1}^{n} \beta_i K_A(\cdot, \hat{A}_i) \left(\frac{1}{n} \sum_{j=1}^{n} K_X(\cdot, \boldsymbol{X}_j) \right) : \alpha_i, \beta_i \in \mathbb{R} \right\}.$$

With Theorem 1, we are able to take a further step and obtain a closed-form representation of $\sup_{u\in\mathcal{H}(1)} Q(\boldsymbol{w},\lambda,u)$. Define

$$\begin{aligned} \boldsymbol{G}_A &:= [K_A(\hat{A}_i,\hat{A}_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n}; \qquad \boldsymbol{G}_X := [K_X(\boldsymbol{X}_i,\boldsymbol{X}_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n}; \\ \bar{\boldsymbol{G}}_X &:= \left[\frac{1}{n}\sum_{j=1}^n K_X(\boldsymbol{X}_i,\boldsymbol{X}_j)\right]_{i=1}^n \in \mathbb{R}^n; \qquad \bar{g}_X := \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n K_X(\boldsymbol{X}_i,\boldsymbol{X}_j). \end{aligned} \tag{15}$$

Then, for any $u\in\mathcal{H}_n$, $Q(\boldsymbol{w},\lambda,u)$ can be written as

$$\begin{aligned} Q(\boldsymbol{w},\lambda,u) =& \frac{1}{n}\Big\|\boldsymbol{G}_A\left(\boldsymbol{G}_A+n\lambda\boldsymbol{I}\right)^{-1}\left[\boldsymbol{w}\circ\left\{\left(\boldsymbol{G}_A\circ\boldsymbol{G}_X\right)\boldsymbol{\alpha}+\left(\boldsymbol{G}_A\odot\bar{\boldsymbol{G}}_X^\intercal\right)^\intercal\boldsymbol{\beta}\right\}\right] \\ &-\left[\left(\boldsymbol{G}_A\odot\bar{\boldsymbol{G}}_X^\intercal\right)\boldsymbol{\alpha}+\bar{g}_X\boldsymbol{G}_A\boldsymbol{\beta}\right]\Big\|_2^2, \quad \text{for some } \boldsymbol{\alpha}\in\mathbb{R}^n \text{ and } \boldsymbol{\beta}\in\mathbb{R}^n, \end{aligned}$$

where $\circ$ is the element-wise product between matrices (vectors), $\odot$ is the column-wise Khatri-Rao product, and $\|\cdot\|_2$ is the Euclidean norm of a vector.

We next simplify the constraint $u\in\mathcal{H}(1)$ in maximizing $Q(\boldsymbol{w},\lambda,u)$. By Theorem 1, it suffices to only focus on the squared RKHS norm of a function $u\in\mathcal{H}_n$, which can be expressed as

$$\left\|\sum_{i=1}^n \alpha_i K_A(\cdot,A_i)K_X(\cdot,\boldsymbol{X}_i)+\sum_{i=1}^n\beta_i K_A(\cdot,A_i)\left(\frac{1}{n}\sum_{j=1}^n K_X(\cdot,\boldsymbol{X}_j)\right)\right\|_{\mathcal{H}}^2 = \boldsymbol{\gamma}^\intercal\boldsymbol{G}_F\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma}=[\boldsymbol{\alpha}^\intercal,\boldsymbol{\beta}^\intercal]^\intercal\in\mathbb{R}^{2n}$ and

$$\boldsymbol{G}_F = \begin{bmatrix} \boldsymbol{G}_A\circ\boldsymbol{G}_X & \left(\boldsymbol{G}_A\odot\bar{\boldsymbol{G}}_X^\intercal\right)^\intercal \\ \boldsymbol{G}_A\odot\bar{\boldsymbol{G}}_X^\intercal & \bar{g}_X\boldsymbol{G}_A \end{bmatrix}\in\mathbb{R}^{2n\times 2n}.$$

Note that we can decompose $\boldsymbol{G}_F$ as

$$\boldsymbol{G}_F = \boldsymbol{M}\boldsymbol{M}^\intercal = \begin{bmatrix}\boldsymbol{M}_1\\ \boldsymbol{M}_2\end{bmatrix}[\boldsymbol{M}_1^\intercal,\boldsymbol{M}_2^\intercal],$$

where $\boldsymbol{M}\in\mathbb{R}^{2n\times q}$, $\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathbb{R}^{n\times q}$, $q\le 2n$. Thus finally, we have

$$\begin{aligned} \sup_{u\in\mathcal{H}(1)} Q(\boldsymbol{w},\lambda,u) =& \frac{1}{n}\sup_{\boldsymbol{\gamma}^\intercal\boldsymbol{G}_F\boldsymbol{\gamma}=1}\left\|\boldsymbol{G}_A\left(\boldsymbol{G}_A+n\lambda\right)^{-1}\boldsymbol{I}\left\{\mathrm{diag}(\boldsymbol{w})\boldsymbol{M}_1\boldsymbol{M}^\intercal\boldsymbol{\gamma}\right\}-\left(\boldsymbol{M}_2\boldsymbol{M}^\intercal\right)\boldsymbol{\gamma}\right\|_2^2 \\ =& \frac{1}{n}\sup_{\|\boldsymbol{M}\boldsymbol{\gamma}\|_2=1}(\boldsymbol{M}^\intercal\boldsymbol{\gamma})^\intercal\left\{\boldsymbol{G}_A\left(\boldsymbol{G}_A+n\lambda\boldsymbol{I}\right)^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{M}_1-\boldsymbol{M}_2\right\}^\intercal \\ & \qquad\left\{\boldsymbol{G}_A\left(\boldsymbol{G}_A+n\lambda\right)^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{M}_1-\boldsymbol{M}_2\right\}(\boldsymbol{M}^\intercal\boldsymbol{\gamma}) \\ =& \frac{1}{n}\left[\sigma_{\max}\left\{\boldsymbol{G}_A\left(\boldsymbol{G}_A+n\lambda\boldsymbol{I}\right)^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{M}_1-\boldsymbol{M}_2\right\}\right]^2, \end{aligned} \tag{16}$$

where $\sigma_{\max}(\cdot)$ returns the largest singular value of the input matrix. As such, (16) is the closed-form representation of the objective function in the inner optimization of (14).

3.1.2 Convexity with respect to weights

We next show that the objective function in (14) is convex with respect to $\boldsymbol{w}$. First, the regularization term $R(\boldsymbol{w}, \lambda)$ in the outer minimization is a quadratic function of $\boldsymbol{w}$ and hence convex in $\boldsymbol{w}$. Moreover, the inner maximization has been rewritten as in (16), and its convexity in $\boldsymbol{w}$ is implied by the following lemma.

Lemma 1 *For fixed $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times q}$ and $\boldsymbol{D} \in \mathbb{R}^{n\times q}$, $\varrho(\boldsymbol{w}) = \left[\sigma_{\max}\left\{\boldsymbol{A}\mathrm{diag}(\boldsymbol{w})\boldsymbol{B} - \boldsymbol{D}\right\}\right]^2$ is a convex function.*

Finally, we collect the previous results and express (14) in a practical optimization form. Notice that

$$R(\boldsymbol{w}, \lambda) = \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{A_i} + \lambda \mathcal{I}\right)^{-1} \mathcal{K}_{A_i} \right\|_n^2 = \sum_{j=1}^{n} w_j^2 \left\{ \frac{1}{n}\sum_{i=1}^{n} \left\{ \left[\boldsymbol{G}_A \left(\boldsymbol{G}_A + n\lambda \boldsymbol{I}\right)^{-1} \right]_{i,j} \right\}^2 \right\}, \tag{17}$$

where $\boldsymbol{A}_{i,j}$ indicates the (i,j)-th element of matrix $\boldsymbol{A}$. Therefore we can show that (14) is equivalent to

$$\begin{aligned} \hat{\boldsymbol{w}} = \underset{0<w_i<L, i=1,\ldots,n}{\mathrm{argmin}} & \frac{1}{n}\left[\sigma_{\max}\left\{\boldsymbol{G}_A\left(\boldsymbol{G}_A + n\lambda\boldsymbol{I}\right)^{-1}\mathrm{diag}(\boldsymbol{w})\boldsymbol{M}_1 - \boldsymbol{M}_2\right\}\right]^2 \\ & + \eta \sum_{j=1}^{n} w_j^2 \left\{ \frac{1}{n}\sum_{i=1}^{n} \left\{ \left[\boldsymbol{G}_A \left(\boldsymbol{G}_A + n\lambda \boldsymbol{I}\right)^{-1} \right]_{i,j} \right\}^2 \right\}. \end{aligned} \tag{18}$$

Due to the convexity of the objective function, common algorithms such L-BFGS-B can be applied to solve (18) given the smoothing parameter λ and tuning parameter η.

3.2 Tuning parameter selection

Here we discuss how to select λ and η. The smoothing parameter λ needs to be provided in order to calculate the balancing error (12), and hence the weights. Recall that the weights are used to form a modified outcome for the WMKRR. Naturally, one would tune λ based on common methods for kernel ridge regression such as cross-validation, but this becomes very difficult in our case because of the complicated dependency between the weights and λ. To address this issue, we propose a simple solution which performs reasonably well in practice. The idea is to use a simple estimator of the adjusted response to guide the selection of λ. More specifically, we first obtain the adjusted responses with the `FCBPS` weights described in Zhang et al. (2021), as their weights do not depend on λ and can be computed quickly. Then we apply the leave-one-out cross-validation (LOOCV) to select λ based on the mean square error computed in the validation set. Finally we use the selected λ to compute the proposed weights without updating λ further.

As for the hyper-parameter η, it is related to the magnitude of weights so as to achieve a balance between (9) and (10). Here we propose to use a fitted outcome regression to help select the best η. To be specific, we fit a KRR to get an estimate for m, and denote it as $\hat{m}$. Then we take $\hat{\tau}_{\texttt{REG}} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(\cdot, \boldsymbol{X}_i)$ as the estimator for τ from the regression approach.

Denote by $\hat{w}_i^{(\eta)}$, $i = 1, \ldots, n$, the weights defined in (14) for each given η. We select the best η such that

$$V(\eta) := \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} + \lambda \mathcal{I} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{\hat{A}_i} \hat{w}_i^{(\eta)} \hat{m}(\hat{A}_i, \boldsymbol{X}_i) \right) - \hat{\tau}_{\texttt{REG}} \right\|_n^2 + \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} + \lambda \mathcal{I} \right)^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{\hat{A}_i} \hat{w}_i^{(\eta)} \left\{ Y_i - \hat{m}(\hat{A}_i, \boldsymbol{X}_i) \right\} \right] \right\|_n^2 \quad (19)$$

is the smallest. In Algorithm 1 in Section A in the supplementary material, we summarize the computation steps to obtain the proposed $\hat{\eta}$ with tuning parameter selection.

4. Theory

In this section we provide the rate of convergence for $\hat{\tau}$. We first introduce a few notations and assumptions. Given two Hilbert spaces $\mathcal{V}$ and $\mathcal{W}$, the operator norm of an operator $B \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ is defined as $\|B\|_{\mathcal{L}(\mathcal{V},\mathcal{W})} := \sup_{\|v\|_{\mathcal{V}} \leq 1} \|Bv\|_{\mathcal{W}}$.

Assumption 3 *$\{(A_i, \boldsymbol{X}_i, \epsilon_i(a), a \in \mathcal{A}), i = 1, \ldots, n\}$ are independent. For any $a \in \mathcal{A}$, $\mathbb{E}(\epsilon_i(a)|A_i, \boldsymbol{X}_i) = 0$ for $i = 1, \ldots, n$, and $\sup_{1 \leq i \leq n} \mathbb{E}(\epsilon_i(a)^2|A_i, \boldsymbol{X}_i) \leq \sigma_0^2$ for some constant $\sigma_0 > 0$.*

Assumption 4 *$\tau \in \mathcal{H}_A$ and $m \in \mathcal{H}$.*

Assumption 5 *K_X and K_A are positive definite kernels. K_X is continuous and the real function $(a_1, a_2) \mapsto \langle \mathcal{K}_{a_1} c_1, \mathcal{K}_{a_2} c_2 \rangle_{\mathcal{H}_A}$ is measurable for any $c_1, c_2 \in \mathbb{R}$. There exists a constant $C_2 > 0$ such that $\sup_{a \in \mathcal{A}} |K_A(a, a)| \leq C_2$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} |K_X(\boldsymbol{x}, \boldsymbol{x})|\} \leq C_2$.*

Assumption 6 *Either one of the following two conditions is satisfied.*

(a) Functional treatments A_i, $i = 1, \ldots, n$, are fully observed without error. In this case, we take $\hat{A}_i = A_i$. This corresponds to $\kappa = 0$ in (20) below.

(b) There exists a pseudometric $d : \mathcal{A} \times \mathcal{A} \to [0, \infty)$ such that:

(i) The mapping $\mathcal{K}_{(\cdot)} : \mathcal{A} \to \mathcal{L}(\mathbb{R}, \mathcal{H}_A)$ is Hölder continuous, i.e., there exist constants $H > 0$ and $0 < h \leq 1$ such that

$$\|\mathcal{K}_{a_1} - \mathcal{K}_{a_2}\|_{\mathcal{L}(\mathbb{R},\mathcal{H}_A)} \leq H[d(a_1, a_2)]^h, \quad a_1, a_2 \in \mathcal{A}.$$

(ii) All $\hat{A}_1, \ldots, \hat{A}_n$ are independent. Each $\hat{A}_i$ can estimate A_i at a uniform rate $\kappa = \kappa(n)$ over $i = 1, \ldots, n$, i.e.,

$$\max_{1 \leq i \leq n} d(\hat{A}_i, A_i) = \mathcal{O}_{\mathrm{p}}(\kappa)\,. \quad (20)$$

Assumption 3 is a standard assumption for the data, which allows for heteroscedasticity of the errors. Assumption 4 states that the function classes for the target function τ and the outcome model m are well-specified. Assumption 5 is a boundedness requirement for the reproducing kernels. It is satisfied for the majority of common kernels including Gaussian and exponential kernels discussed in Remark 2. As for Assumption 6, we only need one of the two specified conditions. When all functional treatments are fully observed without error (Assumption 6(a)), there is no need to recover A_i. So we can simply take $\hat{A}_i = A_i$. Otherwise, we need to recover A_i by $\hat{A}_i$. Assumption 6(b) specifies the related conditions for $\hat{A}_i$ in this case: the Hölder continuity of the operator $\mathcal{K}_a$ and rate of convergence for $\hat{A}_i$. For example, when we take d in Assumption 6(b) as the norm $\|\cdot\|$ used in constructing the kernel in Remark 2, the Gaussian kernel and exponential kernel mentioned in Remark 2 satisfy the Hölder continuity condition with $h = 1$ and $h = 1/2$ respectively. See Table 1 in Szabó et al. (2016) for more examples. As for the rate of convergence κ, it can be specified given different applications. As in Example 1 in Section 2.2, if every A_i, $i = 1, \ldots, n$, satisfies $\int_{\mathcal{T}} A_i^2(t)\rho_T(t)\mathrm{d}t < \infty$ (ρ_T was defined as in Example 1), one can fix the norm $\|f\| = [\int_{\mathcal{T}} f^2(t)\rho_T(t)\mathrm{d}t]^{1/2}$ for each $f \in \mathcal{A}$ and obtain a nonparametric convergence rate for κ with typical nonparametric regression approaches. For example, when A_i is a twice-differentiable univariate function, under appropriate assumptions, the smoothing spline regression can lead to $\kappa = N^{-2/5}\log n$ (Raskutti et al., 2012), where $\log n$ is due to the union bound over n functions; the local polynomial kernel smoothing leads to $\kappa = N^{-2/5}$ (Zhang and Chen, 2007). In Example 2 in Section 2.2, when $\hat{A}_i$, $i = 1, \ldots, n$, are empirical kernel embeddings, one can take $\|\cdot\| = \|\cdot\|_{\mathcal{H}_e}$, the RKHS norm of the kernel embeddings associated with the reproducing kernel K_e, and let $\kappa = N^{-1/2}\log n$. See Section A.1.10 in Szabó et al. (2015) for more detailed results.

We introduce additional terms before presenting the last technical assumption. Define $\mathcal{S}_a = \mathcal{K}_a\mathcal{K}_a^*$ and let $\mathcal{S} = \mathbb{E}_{A\sim\rho_A}\mathcal{S}_A$. Here, ρ_A indicates the marginal distribution of A. Define the trace norm $\mathrm{Tr}(\cdot)$ of a semi-positive-definite operator $B : \mathcal{B} \to \mathcal{B}$ as $\mathrm{Tr}(B) = \sum_l \langle Be_l^{\mathcal{B}}, e_l^{\mathcal{B}}\rangle_{\mathcal{B}}$ with $\{e_l^{\mathcal{B}}\}$ an orthonormal basis of $\mathcal{B}$. Under Assumption 5, $\mathrm{Tr}(\mathcal{S}) \le \sup_{a\in\mathcal{A}} \mathrm{Tr}(\mathcal{S}_a) \le C_2$. Then the spectral theorem yields

$$\mathcal{S} = \sum_{l=1}^{L} t_l \langle \cdot, e_l\rangle_{\mathcal{H}_A} e_l, \tag{21}$$

where $\{e_l\}_{l=1}^L \subset \mathcal{H}_{\mathcal{A}}$ such that $\langle e_l, e_{l'}\rangle_{\mathcal{H}_A} = 1$ if $l = l'$ and 0 otherwise, and $t_1 \ge t_2 \ge \cdots \ge t_L > 0$, with $\sum_{l=1}^L t_l = \mathrm{Tr}(\mathcal{S}) \le C_2$. Here L can be ∞. We also define

$$\mathcal{N}(\lambda) := \mathrm{Tr}\{(\mathcal{S} + \lambda\mathcal{I})^{-1}\mathcal{S}\} = \sum_{l=1}^{\infty} \frac{t_l}{t_l + \lambda}.$$

As to be shown in the theorems below, the rate of convergence for $\hat{\tau}$ depends on the decay of the eigenvalues of $\mathcal{S}$. We also note that the theorems below hold for general choices of kernels.

Theorem 2 *Under Assumptions 1–6, if* $\mathcal{N}(\lambda)(\lambda n)^{-1} = \mathcal{O}(1)$, $\lambda \leq \|\mathcal{S}\|_{\mathcal{L}(\mathcal{H}_A)}$, $\kappa^{2h} = \mathcal{O}(\lambda\mathcal{N}(\lambda)n^{-1})$, $\kappa^h = \mathcal{O}(\lambda)$ *and* $\sqrt{\sum_{l=1}^{\infty}\min\{t_l,\lambda\}}(\sqrt{n}\lambda)^{-1} = \mathcal{O}(1)$, *we have*

$$\sup_{u\in\mathcal{H}(1)} Q(\hat{\boldsymbol{w}},\lambda,u) = \mathcal{O}_{\mathrm{p}}\left[(1+\eta)\frac{\mathcal{N}(\lambda)}{n}+\lambda\right], \tag{22}$$

$$\textit{and}\quad R(\hat{\boldsymbol{w}},\lambda) = \mathcal{O}_{\mathrm{p}}\left[\frac{\mathcal{N}(\lambda)}{n}+\eta^{-1}\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda\right)\right], \tag{23}$$

where $\hat{\boldsymbol{w}} := (\hat{w}_1,\ldots,\hat{w}_n)^{\intercal}$.

Theorem 2 provides the orders of the balancing error (12) and the regularization (13) with respect to $\hat{\boldsymbol{w}}$ given the condition that $\hat{A}_i$ converges to A_i sufficiently fast. In the discussion after Theorem 3, we provide some specific example for the requirement of N to satisfy this condition. Based on Theorem 2, we can develop the rate of convergence for the proposed weighted estimator $\hat{\tau}$. For instance, if the decay of eigenvalues of $\mathcal{S}$ follows a polynomial rate as shown in Assumption 7 below, we are able to achieve a nonparametric convergence rate for τ.

Assumption 7 *There exists a constant* $b > 1$ *such that* $t_l \asymp l^{-b}$ *for any* $l \geq 1$.

This decay rate is also considered in Caponnetto and De Vito (2007) and Szabó et al. (2016).

Theorem 3 *Suppose that the conditions stated in Theorem 2 hold. Then*

$$\|\hat{\tau}-\tau\|_n = \mathcal{O}_{\mathrm{p}}\left(\sigma_0\eta^{-1/2}\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda\right)^{1/2}+\left(\frac{\mathcal{N}(\lambda)}{n}\right)^{1/2}\right)\|m\|_{\mathcal{H}}.$$

If we further assume Assumption 7 hold, $\lambda \asymp n^{-b/(1+b)}$, $\|\mathcal{S}\|_{\mathcal{L}(\mathcal{H}_A)} \geq \lambda$, $\kappa = \mathcal{O}(n^{-b/[h(1+b)]})$ *and* $\eta \asymp 1$, *then there exists some constant* $C_b > 0$ *such that*

$$\mathcal{N}(\lambda) \leq C_b\lambda^{-1/b}.$$

Also, we have

$$\|\hat{\tau}-\tau\|_n = \mathcal{O}_{\mathrm{p}}\left(n^{-\frac{b}{2(1+b)}}\right)\|m\|_{\mathcal{H}}.$$

Based on the conditions of κ listed in Theorem 2 and 3, when λ is chosen optimally, i.e., $\lambda \asymp n^{-b/(1+b)}$, one need $\kappa = \mathcal{O}(n^{-b/[(1+b)h]})$. Take the Gaussian kernel for K_A as an example. When A_i, $i = 1,\ldots,n$ are twice differential univariate functions and they are estimated using smoothing splines (see Example 1), this condition requires that $N \gg n^{5b/(2+2b)}(\log n)^{5/2}$. If empirical kernel mean embeddings are adopted for $\hat{A}_i$, $i = 1,\ldots,n$ (see Example 2), this condition requires that $N \gg n^{2b/(1+b)}(\log n)^2$. Theorem 3 provides the rate of convergence for $\hat{\tau}$ under Assumption 7. According to Caponnetto and De Vito (2007), this rate is minimax in estimating the target function $\tau \in \mathcal{H}_A$ using the i.i.d data $\{A_i, w_i^* Y_i\}_{i=1}^n$.

Remark 5 *The proposed causal effect estimator, based on* $\hat{\boldsymbol{w}}$*, enjoys the same minimax rate of convergence as the ordinary kernel ridge regression estimator using the modified outcome* $\{w_i^* Y_i\}_{i=1}^n$ *with the true but unknown weights. However, the theoretical analysis with weights obtained by* (14) *is significantly more complicated than the typical analysis for kernel ridge regression. One reason is that the responses in kernel ridge regression are typically assumed independent, while the adjusted responses* $\hat{w}_i Y_i$ *are all dependent since* $\hat{\boldsymbol{w}}$ *are obtained by* (14)*. Moreover, as we do not impose any modeling assumption of* $\boldsymbol{w}^*$*, the convergence between* $\hat{\boldsymbol{w}}$ *and* $\boldsymbol{w}^*$ *cannot be established or used to show the convergence of WMKRR. Instead we perform a careful analysis to control the uniform error* $\sup_{u \in \mathcal{H}(1)} Q(\hat{\boldsymbol{w}}, \lambda, u)$*, which leads to the convergence for* $\hat{\tau}$*.*

5. Numerical Studies

5.1 Simulation

We first compare the finite-sample performance of different estimators via a simulation study. We have 200 simulated datasets where $n = 200$ independent subjects are generated in each simulated data. The observations $(A_i, \boldsymbol{X}_i, Y_i)$ for the i-th subject, $i = 1, \ldots, n$, in each simulated data are i.i.d. copies of $(A, \boldsymbol{X}, Y)$ as below. We assume the functional variables A_i's are fully observed. The confounders $\boldsymbol{X} = [X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}]^\intercal \in \mathbb{R}^4$ follow the multivariate normal distribution with the zero mean vector and the identity covariance matrix. The functional treatment A is generated by $A(t) = \sum_{k=1}^4 A^{(k)} \sqrt{2} \sin(2\pi k t)$, $t \in [0, 1]$, where $A^{(1)} \mid \boldsymbol{X} \sim N(4X^{(1)}, 1)$, $A^{(2)} \mid \boldsymbol{X} \sim N(2\sqrt{3}X^{(2)}, 1)$, $A^{(3)} \mid \boldsymbol{X} \sim N(2\sqrt{2}X^{(3)}, 1)$ and $A^{(4)} \mid \boldsymbol{X} \sim N(2X^{(4)}, 1)$.

The outcome Y is generated by $Y \mid (A, \boldsymbol{X}) \sim N(m(A, \boldsymbol{X}), 1)$, where we consider three choices for m as follows. Let $\Psi(\boldsymbol{x}) = x^{(2)}(x^{(1)})^2 + (x^{(4)})^2 \sin(2x^{(3)})$ where $\boldsymbol{x} = [x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}]^\intercal$ and $\mu(t) = 2\sqrt{2}\sin(2\pi t) + \sqrt{2}\cos(2\pi t) + \sqrt{2}\sin(4\pi t)/2 + \sqrt{2}\cos(4\pi t)/2$.

- Setting 1: We let $m(a, \boldsymbol{x}) = 15\Psi(\boldsymbol{x}) + \int_{t=0}^1 a(t)\mu(t)\mathrm{d}t$. In this case, the treatment effect $\tau(a)$ is linear in a in the sense

$$\tau(a) = \int_{t=0}^1 a(t)\mu(t)\mathrm{d}t,$$

  and m is additive in $\Psi(\boldsymbol{x})$ and a.

- Setting 2: We let $m(a, \boldsymbol{x}) = 10\Psi(\boldsymbol{x}) + 0.5(a^{(1)})^2 + 4\sin(a^{(1)})$. Here m is additive in $\Psi(\boldsymbol{x})$ and a. Then the treatment effect is

$$\tau(a) = 0.5(a^{(1)})^2 + 4\sin(a^{(1)}),$$

  which is nonlinear in a.

- Setting 3: We let $m(a, \boldsymbol{x}) = [1 + 2/3\Psi(\boldsymbol{x})][0.5(a^{(1)})^2 + 4\sin(a^{(1)})]$. In this case, the treatment effect $\tau(a)$ has the same form as in Setting 3, but $\Psi(\boldsymbol{x})$ interacts with a in m.

In this simulation study, we compare the following estimators for τ.

1. `CFB`: our proposed (KRR) estimator where weights are obtained from (14).

2. `FCBPS`: the weighted least squares estimator proposed in Zhang et al. (2021) where the weights are obtained by using the parametric SFPS estimation described in Zhang et al. (2021).

3. `NPFCBPS`: the weighted least squares estimator proposed in Zhang et al. (2021) where the weights are obtained by using the non-parametric SFPS estimation described in Zhang et al. (2021).

4. `REG`: the regression estimator $\hat{\tau}_{\texttt{REG}}$ discussed in Section 3.2.

5. `FLM`: the regression estimator $\sum_{i=1}^{n} \hat{m}(\cdot, X_i)/n$ where m is estimated by the classical functional linear model, i.e., $\hat{m}(a, x) = \int_t a(t)\hat{\beta}(t)dt + x^{\intercal}\hat{\gamma}$ for some $\hat{\beta}$ and $\hat{\gamma}$.

6. `FGAM`: the regression estimator $\sum_{i=1}^{n} \hat{m}(\cdot, X_i)/n$ where m is estimated by the functional generalized additive model McLean et al. (2014), i.e., $\hat{m}(a, x) = \hat{f}_1(a) + \sum_{j=1}^{4} \hat{g}_j(x^{(j)})$ for nonlinear functions $\hat{f}_1$ and $\hat{g}_j$, $j = 1, \ldots, 4$.

7. `NW`: the estimator without adjusting the response, i.e., $w_i^* Y_i$ in (4) is replaced by the original response Y_i for $i = 1, \ldots, n$.

When performing the KRR for `CFB`, `REG` and `NW`, we take K_A and K_X both as Gaussian kernels. More specifically, $K_A(a_1, a_2) = (\sqrt{2\pi}\sigma_A)^{-1} \exp\{-\int_{t=0}^{1}(a_1(t) - a_2(t))^2 \mathrm{d}t/\sigma_A^2\}$ and $K_X(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\sqrt{2\pi}\sigma_X)^{-1} \exp\{-(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\intercal}(\boldsymbol{x}_1 - \boldsymbol{x}_2)/\sigma_X^2\}$, where σ_A and σ_X are selected by the median heuristic (Fukumizu et al., 2009; Garreau et al., 2017). For `FCBPS` and `NPFCBPS`, the number of FPC L is chosen such that the top L FPC scores explain 95% percentage of the variance. Both `FCBPS` and `NPFCBPS` assume a linear model for τ. Therefore only Setting 1 is correctly specified for them. For `CFB`, we select its tuning parameter by the procedure described in Section 3 (or Algorithm 1 in the supplementary material) For `REG` and `NW`, we use LOOCV to select the smoothing parameter λ in KRR. Besides the general regression estimator `REG`, we also include two additional regression estimators (`FLM` and `FGAM`) that use common scalar-on-function regression models to estimate the outcome function. `FLM` is the most restirctive model. `FGAM` is less restirctive than `FLM` and we use the function `plr` in R package `refund` to obtain the estimated m.

Two evaluation metrics are provided to assess the performance of these estimators. Take τ' as a generic estimator of τ.

1. Empirical MSE: the mean squared errors (MSE) on sample points: $\sum_{i=1}^{n}[\tau(A_i) - \tau'(A_i)]^2/n$.

2. Out-of-Sample MSE: the mean squared errors (MSE) measured on a set of new evaluation points: $\sum_{i=1}^{n'}[\tau(A_i') - \tau'(A_i')]^2/n'$, where $n' = 100$ and $A_i'(t) = \sum_{k=1}^{4} A_i'^{(k)}\sqrt{2}\sin(2\pi k t)$ with A_i' sampled from the marginal distribution of A_i, i.e., $A_i'^{(1)} \sim N(0, 17)$, $A_i'^{(2)} \sim N(0, 13)$, $A_i'^{(3)} \sim N(0, 9)$ and $A_i'^{(4)} \sim N(0, 5)$.

Tables 1 and 2 show the empirical MSEs and Out-of-Sample MSEs for the above estimators based on 200 simulated datasets respectively. For both evaluation metrics, `NW` has

bad performance among all settings, as it does not adjust for selection bias. For `FLM` and `FGAM`, even though they adjust for the confounders, they perform badly among all settings due to the misspecifications of outcome regression models. `FCBPS` and `NPFCBPS` perform worse than `CFB` and `REG` in Settings 2 and 3 as the assumption of linear model is violated in these two settings. Even though Setting 1 satisfies the linear assumption, the weights calculated from `FCBPS` and `NPFCBPS` are only able to balance the case where the outcome model m is linear in both a and x. Therefore, they do not perform as well as `CFB`. Overall, `CFB` achieves the smallest average of MSEs among all five estimators except for Setting 2 where it is outperformed only by `REG`. Moreover, the MSE associated with `CFB` has the smallest standard errors in all three settings, which demonstrates its attractive stability. In Appendix C, we perform an additional simulation study where covariates $\boldsymbol{X}$ are dependent and A has a more complex dependence on $\boldsymbol{X}$. Similar conclusions are obtained.

Table 1: Empirical MSEs for different estimators under three different simulation settings. Values in the parentheses are the standard errors of MSEs.

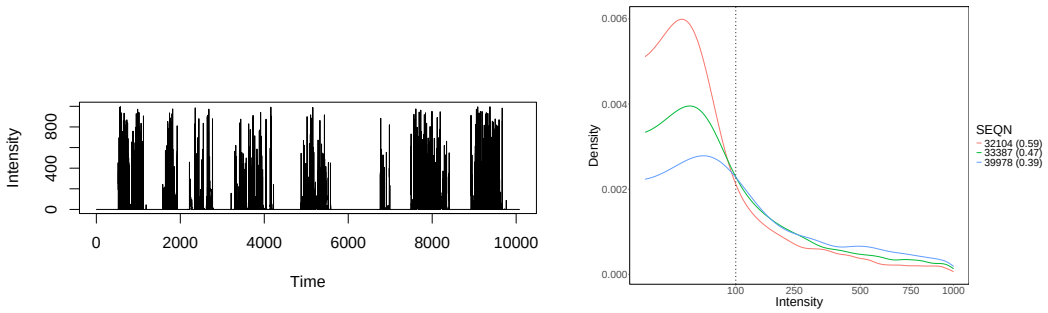
	Setting 1	Setting 2	Setting 3
`CFB`	34.82 (1.02)	130.54 (2.77)	138.15 (4.10)
`FCBPS`	92.91 (2.83)	201 (3.48)	194.57 (4.46)
`NPFCBPS`	105.17 (2.44)	210.44 (3.44)	228.26 (5.59)
`REG`	94.22 (4.46)	99.78 (4.11)	182.23 (8.56)
`FLM`	184.32 (10.73)	252.94 (8.46)	795.44 (70.13)
`FGAM`	350.48 (16.36)	221.67 (7.94)	2178.39 (254.33)
`NW`	666.78 (21.43)	306.06 (9.45)	2687.51 (259.31)

Table 2: Out-of-Sample MSEs for different estimators under three different simulation settings. Values in the parentheses are the standard errors of MSEs.

	Setting 1	Setting 2	Setting 3
`CFB`	34.48 (1.03)	133.41 (3.93)	137.87 (4.31)
`FCBPS`	91.86 (2.82)	205.00 (4.62)	198.17 (5.11)
`NPFCBPS`	105.39 (2.56)	215.78 (4.62)	231.27 (5.70)
`REG`	95.69 (4.80)	105.25 (4.26)	172.98 (7.60)
`FLM`	179.63 (10.23)	263.03 (8.65)	792.71 (64.85)
`FGAM`	345.62 (17.3)	223.74 (8.72)	2241.57 (301.22)
`NW`	516.04 (17.29)	252.73 (9.17)	1287.87 (92.64)

5.2 Real Data Application

We apply all seven estimators described in Section 5.1 to analyze a physical activity monitoring dataset. The dataset is extracted from the National Health and Nutrition Examination Survey (NHANES) 2005–2006. It contains the activity intensity values measured by activity monitors. For each participant, the physical activity intensity, ranging from 0 to 32767 cpm, was recorded every minute for 7 consecutive days. See Figure 1a for an illustration. More details on the physical activity measurements in this dataset can be found in



(a) The trajectory of activity intensity values. (b) Density curves.

Figure 1: Left plot: the trajectory of activity values during 7 consecutive days for a participant with subject ID 32104 in the physical monitor data. Right plot: Three medoid density curves by performing a k-medoid cluster algorithm with $k = 3$. A square root transformation is performed on the x-axis. Proportion of time spent in sedentary activities (<100 cpm) are given in parentheses.

`https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/PAXRAW_D.htm`. This dataset has been extensively studied in the literature to explore the relationship between physical activity and health-related variables. For example, Parker and Holan (2023) and Fishman et al. (2016) use physical activity data to predict mortality with additional covariates in their models. Their result suggests an inverse causal relationship between activity level and mortality rate. Maher et al. (2013) examine the associations between physical activity and obesity in male and female groups, concluding that moderate-to-vigorous physical activity is inversely associated with obesity. However, in the aforementioned studies, the analyses rely on either (non-data-adaptive) summary statistics of physical activity or (unsupervised) functional principal components, rather than directly exploiting the full information available from the activity profiles for causal estimation. Furthermore, Parker and Holan (2023) and Fishman et al. (2016) employ linear regression models, which may lead to model misspecification. Maher et al. (2013), on the other hand, does not account for other potential confounders. As a result, the causal relationship between physical activity and outcomes established in Maher et al. (2013) relies on potentially overly stringent unconfoundedness assumption, given the observational nature of the data. In our analysis, we aim to investigate the causal effect of activity profiles on BMI values based on this dataset. Other variables collected in this dataset are available from `https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.htm`, from which we have extracted the covariates and outcomes. The covariates $\boldsymbol{X}$ include age, education level and family poverty income ratio, and the outcome Y is the body mass index (BMI). The data set was also studied recently in Chang and McKeague (2022) and Lin et al. (2023). However, they treat physical activity as a functional outcome instead of a functional exposure.

In our analysis, we focus on white male subjects of age 20–50. Prior to the analysis, we follow the exact same pre-processing steps described in Lin et al. (2023) to process the observations of intensity values. These steps are described as follows. First, the observations with questionable reliability according to NHANES protocol are discarded. Then, for every subject, the observations with intensity values higher than 1000 or equal to 0 are removed.

In the dataset, we found that most ($\sim$90%) positive intensity values range between 0 and 1000. Observations with zero intensity values could correspond to activities with varying intensities, such as sleeping and bathing. Since one cannot accurately distinguish between these different activities, Lin et al. (2023) excluded these observations from the study. Again, following Lin et al. (2023), we remove those subjects who have less than 100 remaining minutes with positive intensity values. This step was introduced to avoid subjects with small number of observations (i.e., small N) which could lead to inaccurate estimation of the corresponding functional treatment. In Appendix D, we present a sensitivity analysis, by varying the cutoff minutes to 50, 75, and 200. The results were consistent with our conclusions for the cutoff of 100 minutes. Lastly, we removed observations with many missing covariates. After the above pre-processing steps, the sample size is 427.

Apart from following the pre-processing steps in Lin et al. (2023), we focus on the causal effect of physical activities within intensity range (0,1000] which includes sedentary to light intensity physical activities, as the benefit of light-intensity activities is less clear compared to the ample evidence for the benefit of moderate-to-vigorous activities (Fuezeki et al., 2017).

Note that the raw activity intensity profiles across different subjects are not aligned and thus generally incomparable. To address this problem, one may use their distribution functions to represent them. But these intensity distributions lie in a manifold space. Instead, we apply the kernel mean embedding (Muandet et al., 2017) with a Gaussian kernel to generate distributional representations in a Hilbert space, as discussed in Remark 2. Eventually, the treatments A are taken as the kernel mean embeddings of the intensity distributions. We follow the same procedures in Section 5.1, including the choices of kernels and tuning parameters, to obtain all the five estimators.

To provide a clear look at how different the estimated causal effects are provided by different estimators, we present the estimated BMI values for three representative density curves chosen by performing a cluster analysis using a k-medoid cluster algorithm with $k = 3$ on the density curves in the dataset. Roughly speaking, the three clusters correspond to a high, medium and lower proportion of time spent in sedentary activities (<100 cpm), with representative observations ID 32104, 33387 and 39978 respectively. Figure 1b shows three medoid density curves while Table 3 shows the corresponding fitted BMI values produced by different estimators for these three curves. All seven estimators provide the same order of BMI values for these three curves, indicating that a more active person tends to have a lower BMI. The result of `NPFCBPS` , `FLM` and `FGAM` are rather similar to that of `NW`, while `CFB`, `REG` and `FCBPS` are more similar. The proposed method `CFB` shows the greatest estimated difference in BMI values between the representative activity profiles from the high and medium sedentary time groups.

We compare the performances of the seven methods in categories of estimated average BMI values with the observed treatment patterns. According to the US Centers for Disease Control and Prevention, in terms of the BMI value, an adult may be categorized as: underweight (BMI ≤ 18.5), healthy ($18.5 <$ BMI ≤ 25), overweight ($25 <$ BMI ≤ 30) and obese (BMI > 30). We combine the underweight and healthy categories in our dataset because it only has three underweight observations. We visualize every unique activity profiles and their estimated BMI values ($\hat{\tau}(A)$) in Figure 2. For a better visualization, we stratify the collection of curves by the estimated BMI categories. For `NPFCBPS`, `FLM` and `FGAM`, there is

less variation in the estimated outcome across different density curves compared to other methods. Regardless of physical activity profiles, almost all the subjects have an estimated BMI in the overweight or obese range. For instance, the densities estimated by `FGAM` in the obese group do not exhibit a distinct pattern compared to the other two groups. In fact, the density shapes show considerable variation, making it difficult to identify a consistent pattern. The results from `NW`, which does not adjust for the confounders, are counter-intuitive since there are clearly two subgroups within the designated obese group, with one subgroup of individuals who spend considerable less time in sedentary activities. Although the results of `CFB` are similar to those of `REG` and `FCBPS`, all indicating a steady positive relationship between the proportion of time spent on sedentary activities and BMI, `CFB` shows a clearer effect compared to other methods, which is consistent with the findings in Table 3 shows a clear effect of changing from high to medium sedentary time spent on BMI. Our analysis affirms the recommendation in Fuezeki et al. (2017) that "Currently inactive or insufficiently active people should be encouraged to engage in physical activity of any intensity".

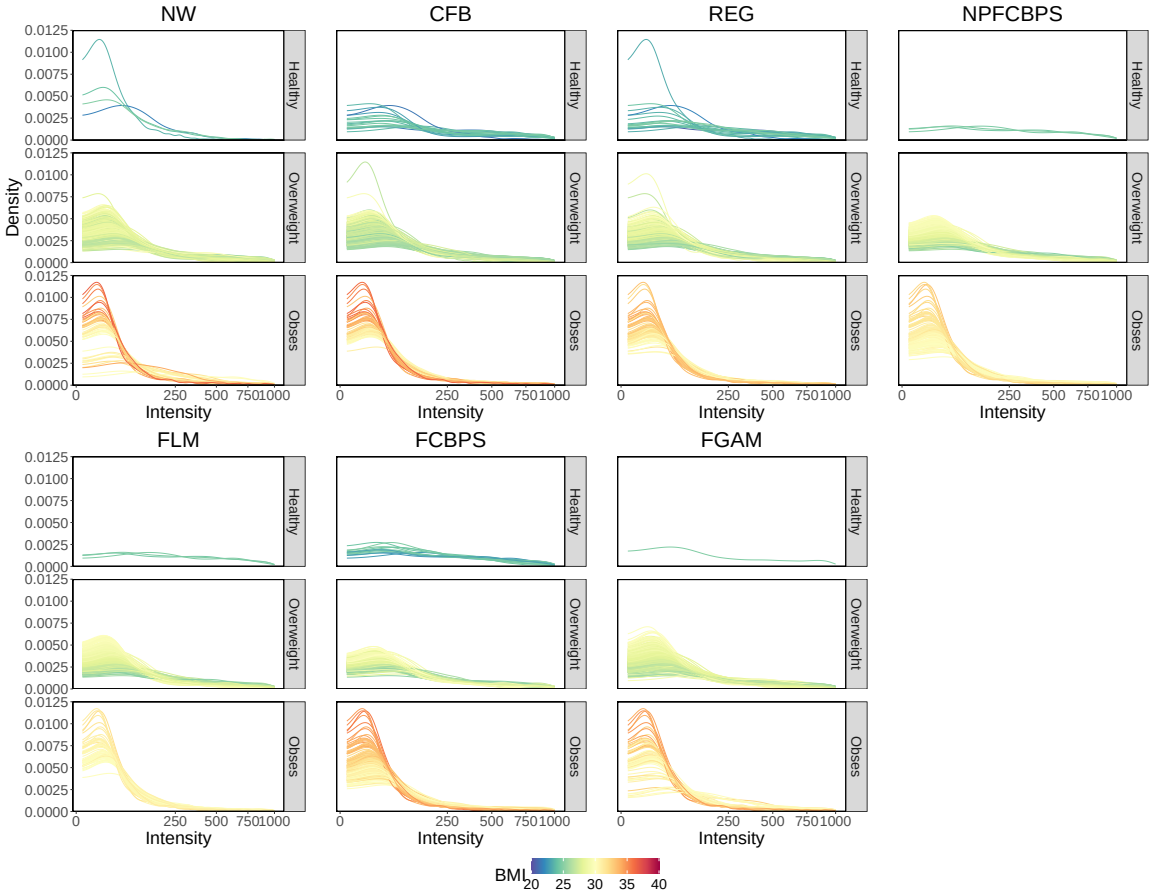


Figure 2: Density curves divided into different categories by their corresponding estimated BMI values using different estimators (`NW`, `CFB`, `REG`, `FCBPS`, `NPFCBPS`, `FLM` and `FGAM`). A square root transformation is performed on the x-axis.

Table 3: Estimated BMI values for the three representative density curves in Figure 1b by different estimators. SEQN indicates the respondent sequence number, i.e., subject ID.

SEQN	CFB	FCBPS	NPFCBPS	REG	FLM	FGAM	NW
32104	30.38	31.75	30.55	31.31	29.86	29.87	29.36
33387	27.16	29.26	28.96	28.43	28.49	27.95	28.78
39978	26.10	27.23	27.56	26.96	27.20	26.77	26.57

6. Discussion

In this paper, we establish a novel covariate balancing framework for FTE estimation. Our framework adopts the highly flexible weight-modified kernel ridge regression to characterize the FTE on the outcome. The proposed weights are obtained by balancing an RKHS of the functional treatment and can be computed efficiently. The proposed FTE estimator is guaranteed to achieve the optimal rate of convergence without any smoothness assumptions of the oracle weight function. Its appealing empirical performance is demonstrated in an extensive simulation study and a real data application.

In the following, we outline several directions for future work. Assumption 1 can be restrictive in practice as it requires all the covariates $\boldsymbol{X}$ that adjust the dependence between $\{Y(a), a \in \mathcal{A}\}$ and A are observed. However, this is often not guaranteed in practice, which leads to the presence of unmeasured confounding. Inspired by the recent development in causal inference that tackles unmeasured confounding using instrumental or proxy variables, we will investigate FTE estimation while relaxing Assumption 1 to allow for unmeasured confounding. In addition, while Theorem 3 provides a convergence result for $\hat{\tau}$ with respect to the empirical norm, there are fundamental difficulties in calculating the $\mathcal{L}_2$ norm for functions with functional inputs numerically, and obtaining the $\mathcal{L}_2$ norm convergence rate. Obtaining such results probably requires a modification of our method to smooth the weights in order to establish the convergence of the estimated weight function. In Section 3.2, we propose a practical way or tuning the hyperparameters η and λ, leveraging an initial weighted estimator from Zhang et al. (2021) and the estimated kernel ridge regression model $\hat{m}$. However, one may want to avoid using these auxiliary estimators. A better tuning parameter selection method is left as a future direction. Last but not least, statistical inference remains a challenge for nonparametric function estimations. In our setting, τ is a function defined over a function space $\mathcal{A}$, which significantly complicates the task of developing valid inference tools. Bootstrap ideas could be potentially used to perform inference. We will leave a thorough study and development of inference as a future direction.

Acknowledgements

The work of Jiayi Wang is partly supported by the National Science Foundation (NSF-2401272). The work of Raymond K. W. Wong is partly supported by the National Science Foundation (DMS-1711952 and CCF-1934904). Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. The work of Xiaoke Zhang is partly supported by the George Washington

University University Facilitating Fund and Columbian College of Arts and Sciences Impact Award. The work of Kwun Chuen Gary Chan is partly supported by the National Science Foundation (DMS-1711952). The authors would like to thank the two referees and the action editor for their helpful and constructive comments. We are also grateful to the action editor and editors-in-chief for their editorial efforts in handling the paper.

Appendix A. Algorithm

Algorithm 1: Outlines for obtaining $\hat{\tau}$.

Input: Observed confounders $\boldsymbol{X}_i \in \mathbb{R}^p$ and approximated treatments $\hat{A}_i \in \mathcal{A}$, $i = 1, \ldots, n$; smoothing parameter $\lambda > 0$; the fitted outcome regression model $\hat{m}$ and a sequence of tuning parameters η_k, $k = 1, \ldots, K$.

Calculate $\boldsymbol{G}_A$, $\boldsymbol{G}_X$, $\bar{\boldsymbol{G}}$ and $\bar{g}_X$ according to (15).

Decompose $\boldsymbol{G}_F$ and obtain $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}$.

for $k = 1, \ldots, K$ **do**

  Optimize (18) by L-BFGS-B algorithm with $\eta = \eta_k$ and obtain solution $\hat{\boldsymbol{w}}^{(\eta_k)}$.

  Compute the value of $V(\eta_k)$ from (19).

end

Select $\tilde{\eta}$ such that $V(\tilde{\eta})$ is the smallest among all $V(\eta_k)$, $k = 1, \ldots, K$.

Construct the adjusted response $Z_i = \hat{w}_i^{(\tilde{\eta})} Y_i$, $i = 1, \ldots, n$.

Fix the response as Z_i, $i = 1, \ldots, n$ and obtain

$$\hat{\tau} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} + \lambda \mathcal{I} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{\hat{A}_i} Z_i \right).$$

Output: $\hat{\boldsymbol{w}}^{\tilde{\eta}}$ and $\hat{\tau}$.

Appendix B. Derivations and Proofs

Proof [Derivation of (3)]

$$\begin{aligned}
&\mathbb{E}\{w^*(A, \boldsymbol{X}) Y \mid A = a\} = \mathbb{E}\left[w^*(a, \boldsymbol{X}) \mathbb{E}(Y \mid A = a, \boldsymbol{X}) \mid A = a\right] \\
=&\mathbb{E}\left[w^*(a, \boldsymbol{X}) \mathbb{E}(Y(a) \mid A = a, \boldsymbol{X}) \mid A = a\right] = \mathbb{E}\left[w^*(a, \boldsymbol{X}) \mathbb{E}(Y(a) \mid D(a) = 1, \boldsymbol{X}) \mid A = a\right] \\
=&\mathbb{E}\left[w^*(a, \boldsymbol{X}) \mathbb{E}(Y(a) \mid \boldsymbol{X}) \mid A = a\right] = \mathbb{E}\left\{ \frac{\rho_{\boldsymbol{X}}(\boldsymbol{X})}{\rho_{\boldsymbol{X}|A}(\boldsymbol{X} \mid a)} \mathbb{E}(Y(a) \mid \boldsymbol{X}) \mid A = a \right\} \\
=&\int_{\boldsymbol{x}} \frac{\rho_{\boldsymbol{X}}(\boldsymbol{x})}{\rho_{\boldsymbol{X}|A}(\boldsymbol{x} \mid a)} \mathbb{E}(Y(a) \mid \boldsymbol{X} = \boldsymbol{x}) \rho_{\boldsymbol{X}|A}(\boldsymbol{x} \mid a) \mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{x}} \rho_{\boldsymbol{X}}(\boldsymbol{x}) \mathbb{E}(Y(a) \mid \boldsymbol{X} = \boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{E}\{Y(a)\}.
\end{aligned}$$

■

Proof [Proof of Proposition 1] Given the discussion of Assumptions in Section 4, under Assumption 5, the spectral theorem gives

$$\mathcal{S} = \sum_{l=1}^{L} t_l \langle \cdot, e_l \rangle_{\mathcal{H}_A} e_l,$$

where $0 < t_{l+1} \leq t_l$, $\{e_l\}_{l=1}^{\infty}$ is a basis of $\mathrm{Ker}\mathcal{S}^{\perp}$. Define the operator $T : \mathcal{L}_2(\rho_A) \to \mathcal{L}_2(\rho_A)$ to be the integral operator of kernel K_A,

$$(T\rho)(a) = \mathbb{E}K_A(a, A)\psi(A) = \int_{\mathcal{A}} K_A(a, a')\psi(a')d\rho_A(a')$$

for $\psi \in \mathcal{L}_2(\rho_A)$. Based on Remark 2 in Caponnetto and De Vito (2007), it can be shown that

$$T = \sum_{l=1}^{K} \nu_l^A \langle \cdot, \phi_l^A \rangle_{\rho_A} \phi_l^A,$$

where $\{\phi_l^A\}_{l=1}^{L}$ is basis of $\mathrm{Ker}T^{\perp}$, $T^{1/2}\phi_l^A = t_l^{1/2}\phi_l^A = e_l$ and $\lambda_A^l = t_l$, $l = 1, \ldots, L$. And we have

$$K_A(a_1, a_2) = \sum_{l=1}^{L} \nu_l^A \phi_l^A(a_1)\phi_l^A(a_2).$$

As K_X is a bounded continuous positive definite kernel, Mercer Theorem yields

$$K_X(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{l=1}^{L} \nu_l^X \phi_l^X(\boldsymbol{x}_1)\phi_l^X(\boldsymbol{x}_2),$$

where $0 < \phi_{l+1}^X \leq \phi_l^X$ are eigenvalues, $\{(\phi_l^X\}_{l=1}^{\infty}$ is a set of orthonormal basis in $\mathcal{L}_2(\rho_X)$ and $\{(\nu_l^X)^{1/2}\phi_l^X\}_{l=1}^{\infty}$ is a set of orthonormal basis in $\mathcal{H}_X$.

Next, we prove the direction that $\mathbb{E}_{\boldsymbol{X}\sim\rho_X} g(\cdot, \boldsymbol{X}) \in \mathcal{H}_A$ for any $g \in \mathcal{H}$.

Given these two sets of basis, for any function $g \in \mathcal{H}$, it can be expressed as

$$g(a, \boldsymbol{x}) = \sum_{l_1, l_2} \beta_{l_1, l_2} \sqrt{\nu_{l_1}^A} \sqrt{\nu_{l_2}^X} \phi_l^A(a)\phi_l^X(\boldsymbol{x}),$$

for some coefficients $\{\beta_{l_1,l_2}\}_{l_1,l_2=1}^{\infty}$ with

$$\|g\|_{\mathcal{H}}^2 = \sum_{l_1, l_2} \beta_{l_1, l_2}^2 < \infty.$$

Then we have

$$\|\mathbb{E}_{\boldsymbol{X}\sim\rho_X} g(\cdot, \boldsymbol{X})\|_{\mathcal{H}_A}^2 = \sum_{l_1} (\beta_{l_1,l_2})^2 \left\{ \mathbb{E}\left(\sum_{l_2} \sqrt{\nu_{l_2}^X} \phi_{l_2}^X(\boldsymbol{X}) \right) \right\}^2 \leq \sum_{l_1} (\beta_{l_1,l_2})^2 \mathbb{E}\left(\sum_{l_2} \sqrt{\nu_{l_2}^X} \phi_{l_2}^X(\boldsymbol{X}) \right)^2$$
$$\leq \sum_{l_1} (\beta_{l_1,l_2})^2 \sum_{l_2} \nu_{l_2}^X (\beta_{l_2}^X)^2 \leq \left(\max_{l_2} \nu_{l_2}^X \right) \sum_{l_1} \sum_{l_2} (\beta_{l_1,l_2}^A)^2 = \nu_1^X (\sum_{l_1,l_2} \beta_{l_1,l_2}^2) = \nu_1^X \|g\|_{\mathcal{H}}^2,$$

the second inequality is due to the fact that $\{\phi_l^X\}_{l=1}^{\infty}$ is orthonormal in $\mathcal{L}_2(\rho_X)$. Note that K_X is bounded, thus ν_1^X is bounded. The conclusion follows.

Now, we prove the direction that for any function $f \in \mathcal{H}_A$, there exists a function $g \in \mathcal{H}$, such that $\mathbb{E}_{\boldsymbol{X}\sim\rho_X} g(\cdot, \boldsymbol{X}) \in \mathcal{H}$.

First, note that these exists a function $u \in \mathcal{H}_X$, such that

$$\mathbb{E}u(\boldsymbol{X}) = \int_{\boldsymbol{x}} u(\boldsymbol{x})\rho_X(\boldsymbol{x})d\boldsymbol{x} \neq 0.$$

Take

$$g(a, \boldsymbol{x}) = f(a)\frac{u(\boldsymbol{x})}{\mathbb{E}u(\boldsymbol{X})}.$$

One can verify that $\mathbb{E}_{\boldsymbol{X}\sim\rho_X} g(\cdot, \boldsymbol{X}) = f(\cdot)$. Since $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_X$, $f \in \mathcal{H}_A$ and $u(\boldsymbol{x})/\mathbb{E}u(\boldsymbol{X}) \in \mathcal{H}_X$, therefore $g \in \mathcal{H}$. Therefore, the conclusion is verified. ■

Proof [Proof of Theorem 1] Take $\mathcal{H}_n^{\perp}$ as the orthogonal space of $\mathcal{H}_n$. For any function $u \in \mathcal{H}$, we can decompose it into two orthogonal parts u_1 and u_2 such that $u_1 \in \mathcal{H}_n$ and $u_2 \in \mathcal{H}_n^{\perp}$. Then for any $u \in \mathcal{H}$,

$$\begin{aligned}
&Q(\boldsymbol{w}, \lambda, u) \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{\hat{A}_i} w_i u(\hat{A}_i, \boldsymbol{X}_i)\right)\right\}(A_k) - \frac{1}{n}\sum_{j=1}^{n} u(A_k, \boldsymbol{X}_j)\right]^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{\hat{A}_i} w_i \langle K((\hat{A}_i, \boldsymbol{X}_i), (\cdot,\cdot)), u\rangle\right)\right\}(A_k)\right. \\
&\left. -\frac{1}{n}\sum_{j=1}^{n}\langle K(\hat{A}_k, \boldsymbol{X}_j), (\cdot,\cdot)), u\rangle\right]^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{\hat{A}_i} w_i \langle K_A(\cdot, \hat{A}_i)K_X(\cdot, \boldsymbol{X}_i), (\cdot,\cdot)), u_1 + u_2\rangle\right)\right\}(A_k)\right. \\
&\left. - \left\langle K_A(\cdot, \hat{A}_k)\left\{\frac{1}{n}\sum_{j=1}^{n} K_X(\cdot, \boldsymbol{X}_j)\right\}, (\cdot,\cdot)), u_1 + u_2\right\rangle\right]^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left[\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{\hat{A}_i} w_i \langle K_A(\cdot, \hat{A}_i)K_X(\cdot, \boldsymbol{X}_i), (\cdot,\cdot)), u_1\rangle\right)\right\}(A_k)\right. \\
&\left. - \left\langle K_A(\cdot, \hat{A}_k)\left\{\frac{1}{n}\sum_{j=1}^{n} K_X(\cdot, \boldsymbol{X}_j)\right\}, (\cdot,\cdot)), u_1\right\rangle\right]^2 = Q(\boldsymbol{w}, \lambda, u_1).
\end{aligned}$$

On the other hand, $\|u\|_{\mathcal{H}} = \|u_1\|_{\mathcal{H}} + \|u_2\|_{\mathcal{H}}$. For any u such that $\|u\|_{\mathcal{H}} = 1$ and $\|u_1\|_{\mathcal{H}} > 0$, we can always find another $\tilde{u} = \frac{\|u\|_{\mathcal{H}}}{\|u_1\|_{\mathcal{H}}} u_1 \in \mathcal{H}_n$. It is easy to verify that $\|\tilde{u}\|_{\mathcal{H}} = 1$ and

$$Q(\boldsymbol{w}, \lambda, \tilde{u}) = Q\left(\boldsymbol{w}, \lambda, \frac{\|u\|_{\mathcal{H}}}{\|u_1\|_{\mathcal{H}}} u_1\right) = \left(\frac{\|u\|_{\mathcal{H}}}{\|u_1\|_{\mathcal{H}}}\right)^2 Q\left(\boldsymbol{w}, \lambda, u_1\right) = \left(\frac{\|u\|_{\mathcal{H}}}{\|u_1\|_{\mathcal{H}}}\right)^2 Q\left(\boldsymbol{w}, \lambda, u\right) \geq Q\left(\boldsymbol{w}, \lambda, u\right).$$

The conclusion follows. ■

Proof [Proof of Lemma 1] Consider any vectors $\boldsymbol{w}_1 \in \mathbb{R}^n$ and $\boldsymbol{w}_2 \in \mathbb{R}^n$, and $t \in [0, 1]$. For $\boldsymbol{\beta} \in \mathbb{R}^q$, we have

$$\begin{aligned}
&\|[\boldsymbol{A}\mathrm{diag}\{t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2\}\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta}\|_2^2 \\
=&\|t\,[\boldsymbol{A}\mathrm{diag}(\boldsymbol{w}_1)\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta} + (1-t)\,[\boldsymbol{A}\mathrm{diag}(\boldsymbol{w}_2)\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta}\|_2^2 \\
\leq& t\,\|[\boldsymbol{A}\mathrm{diag}(\boldsymbol{w}_1)\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta}\|_2^2 + (1-t)\,\|[\boldsymbol{A}\mathrm{diag}(\boldsymbol{w}_2)\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta}\|_2^2.
\end{aligned}$$

The inequality is due to the convexity of $\|\cdot\|_2^2$. Suppose that $\boldsymbol{\beta}$ is the right singular vector of $\boldsymbol{A}\mathrm{diag}\{t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2\}\boldsymbol{B} - \boldsymbol{D}$ that corresponds to the largest singular value. Then

$$\begin{aligned}
&[\sigma_{\max}\{\boldsymbol{A}\mathrm{diag}(\boldsymbol{w})\boldsymbol{B} - \boldsymbol{D}\}]^2 \\
=&\|[\boldsymbol{A}\mathrm{diag}\{t\boldsymbol{w}_1 + (1-t)\boldsymbol{w}_2\}\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta}\|_2^2 \\
\leq& t\,\|[\boldsymbol{A}\mathrm{diag}(\boldsymbol{w}_1)\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta}\|_2^2 + (1-t)\,\|[\boldsymbol{A}\mathrm{diag}(\boldsymbol{w}_2)\boldsymbol{B} - \boldsymbol{D}]\,\boldsymbol{\beta}\|_2^2 \\
\leq& t\,[\sigma_{\max}\{\boldsymbol{A}\mathrm{diag}(\boldsymbol{w})\boldsymbol{B} - \boldsymbol{D}\}]^2 + (1-t)\,[\sigma_{\max}\{\boldsymbol{A}\mathrm{diag}(\boldsymbol{w}_2)\boldsymbol{B} - \boldsymbol{D}\}]^2.
\end{aligned}$$

The second inequality is due to the definition of the largest singular value. Then conclusion follows. ■

Proof [Proof of Theorem 2]

First, we derive the bounds for $\sup_{u\in\mathcal{H}(1)} Q(\boldsymbol{w}^*, \lambda, u)$ and $R(\boldsymbol{w}^*, \lambda)$

Take the function $z(a; u) = \mathbb{E}_{\boldsymbol{X}\sim\rho_X} u(a, X)$ for $u \in \mathcal{H}(1)$. Take $\mathcal{S}_A = \frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i}$, $\mathcal{S}_{\hat{A}} = \frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}$. It's easy to see that $\sup_{u\in\mathcal{H}(1)} Q(w, \lambda, u)$ can be bounded by the following

components:

$$
\begin{aligned}
&\sup_{u\in\mathcal{H}(1)} Q(\boldsymbol{w}^*,\lambda,u)\\
&\lesssim \sup_{u\in\mathcal{H}(1)}\left\|\left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}+\lambda\mathcal{I}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n\left(\mathcal{K}_{\hat{A}_i}\left\{w_i^* u(\hat{A}_i,\boldsymbol{X}_i)-z(\hat{A}_i;u)\right\}\right.\right.\right.\\
&\qquad \left.\left.\left.-\mathcal{K}_{A_i}\left\{w_i^* u(A_i,\boldsymbol{X}_i)-z(A_i;u)\right\}\right)\right]\right\|_n^2 \qquad (24)\\
&+\sup_{u\in\mathcal{H}(1)}\left\|\left(\mathcal{S}_{\hat{A}}+\lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{A_i}-\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{\hat{A}_i}\right)\left(\mathcal{S}_A+\lambda\mathcal{I}\right)^{-1}\right.\\
&\qquad \left.\left[\frac{1}{n}\sum_{i=1}^n\mathcal{K}_{A_i}\left\{w_i^* u(A_i,\boldsymbol{X}_i)-z(A_i;u)\right\}\right]\right\|_n^2 \qquad (25)\\
&+\sup_{u\in\mathcal{H}(1)}\left\|\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{A_i}+\lambda\mathcal{I}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n\mathcal{K}_{A_i}\left\{w_i^* u(A_i,\boldsymbol{X}_i)-z(A_i;u)\right\}\right]\right\|_n^2 \qquad (26)\\
&+\sup_{u\in\mathcal{H}(1)}\left\|\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{A_i}+\lambda\mathcal{I}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n\mathcal{K}_{A_i}z(A_i;u)\right]-z(\cdot;u)\right\|_n^2 \qquad (27)\\
&+\sup_{u\in\mathcal{H}(1)}\left\|z(\cdot;u)-\frac{1}{n}\sum_{j=1}^n u(\cdot,\boldsymbol{X}_j)\right\|_n^2 \qquad (28)
\end{aligned}
$$

Next, we consider to bound (24), (25), (26), (27) and (28) one by one. Note that under the fully observed case (Assumption 6 1), $\hat{A}_i = A_i$ for $i = 1,\ldots,n$, and we only need to focus on (26) - (28).

First, follow the proof in Section A.1.11 of Szabó et al. (2015), by Assumption 6, we have

$$
\left\|\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{\hat{A}_i}\right)-\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{A_i}\right)\right\|_{\mathcal{L}_{\mathcal{H}_A}}^2 \le \frac{1}{n}\sum_{i=1}^n\left\|\mathcal{S}_{\hat{A}_i}-\mathcal{S}_{A_i}\right\|_{\mathcal{L}_{\mathcal{H}_A}}^2 = \mathcal{O}_{\mathrm{p}}\left(H^2C_2^h\kappa^{2h}\right).
$$

And note that for any function $f\in\mathcal{H}_A$,

$$
\|f\|_n^2 = \left\|\left(\frac{1}{n}\sum_{i=1}^n\mathcal{S}_{\hat{A}_i}\right)^{1/2} f\right\|_{\mathcal{H}_A}^2.
$$

- For (24).

$$
\begin{aligned}
&(24)\\
&\leq \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} + \lambda \mathcal{I} \right)^{-1} \right\|^2_{\mathcal{L}_{\mathcal{H}_A}} \\
&\quad \times \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\mathcal{K}_{\hat{A}_i} \left\{ w_i^* u(\hat{A}_i, \boldsymbol{X}_i) - z(\hat{A}_i; u) \right\} - \mathcal{K}_{A_i} \left\{ w_i^* u(A_i, \boldsymbol{X}_i) - z(A_i; u) \right\} \right) \right\|^2_{\mathcal{H}_A}.
\end{aligned}
$$

By the spectral theorem,

$$
\left\| \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} + \lambda \mathcal{I} \right)^{-1} \right\|^2_{\mathcal{L}_{\mathcal{H}_A}} \leq \frac{1}{\lambda}.
$$

And by Hölder continuous in Assumption 6,

$$
\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \left(\mathcal{K}_{\hat{A}_i} \left\{ w_i^* u(\hat{A}_i, \boldsymbol{X}_i) - z(\hat{A}_i; u) \right\} - \mathcal{K}_{A_i} \left\{ w_i^* u(A_i, \boldsymbol{X}_i) - z(A_i; u) \right\} \right) \right\|^2_{\mathcal{H}_A} \\
\lesssim &\frac{H^2}{n}\sum_{i=1}^{n} \left[d(A_i, \hat{A}_i)^{2h} C_1^2 C_2^4 \right] = \mathcal{O}_{\mathrm{p}}\left(H^2 C_1^2 C_2^4 \kappa^{2h} \right).
\end{aligned}
$$

Then

$$
(24) = \mathcal{O}_{\mathrm{p}}\left(\frac{H^2 C_1^2 C_2^4 \kappa^{2h}}{\lambda} \right).
$$

- For (25), by similar proof, we can show that

$$
\begin{aligned}
&(25)\\
&\leq \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} + \lambda \mathcal{I} \right)^{-1} \right\|^2_{\mathcal{L}_{\mathcal{H}_A}} \left\| \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{A_i} - \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right) \right\|^2_{\mathcal{L}_{\mathcal{H}_A}} \\
&\quad \times \sup_{u \in \mathcal{H}(1)} \left\| (\mathcal{S}_A + \lambda \mathcal{I})^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{A_i} \left\{ w_i^* u(A_i, \boldsymbol{X}_i) - z(A_i; u) \right\} \right] \right\|^2_{\mathcal{H}_A} \\
&= \mathcal{O}_{\mathrm{p}}\left(\frac{H^2 C_1^2 C_2^4 C_2^h \kappa^{2h}}{\lambda} \right) \frac{1}{\lambda} \\
&\quad \times \sup_{u \in \mathcal{H}(1)} \left\| (\mathcal{S}_A + \lambda \mathcal{I})^{-\frac{1}{2}} \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{A_i} \left\{ w_i^* u(A_i, \boldsymbol{X}_i) - z(A_i; u) \right\} \right] \right\|^2_{\mathcal{H}_A}.
\end{aligned}
$$

By later argument (see the proof for bounding (26)), we can prove that

$$\sup_{u\in\mathcal{H}(1)}\left\|(\mathcal{S}_A+\lambda\mathcal{I})^{-\frac{1}{2}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}\left\{w_i^*u(A_i,\boldsymbol{X}_i)-z(A_i;u)\right\}\right]\right\|_{\mathcal{H}_A}^2=\mathcal{O}_{\mathrm{p}}\left(\frac{\mathcal{N}(\lambda)}{n}\right).$$

Then by the condition $\kappa^{2h}=\mathcal{O}(\lambda^2)$, we have

$$(25)=\mathcal{O}_{\mathrm{p}}\left(\frac{H^2C_1^2C_2^4C_2^h\kappa^{2h}}{\lambda}\frac{\mathcal{N}(\lambda)}{n\lambda}\right)=\mathcal{O}_{\mathrm{p}}\left(\frac{\mathcal{N}(\lambda)}{n}\right).$$

- Next, we focus on controlling term (26).

  We have

$$\begin{aligned}
&\sup_{u\in\mathcal{H}(1)}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i}+\lambda\mathcal{I}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}\left\{w_i^*u(A_i,\boldsymbol{X}_i)-z(A_i;u)\right\}\right]\right\|_n^2\\
&=\sup_{u\in\mathcal{H}(1)}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i}+\lambda\mathcal{I}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}\left\{w_i^*u(A_i,\boldsymbol{X}_i)-z(A_i;u)\right\}\right]\right\|_{\mathcal{H}_A}^2\\
&\le\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i}+\lambda\mathcal{I}\right)^{-1}(\mathcal{S}+\lambda)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H}_A)}^2\\
&\qquad\times\sup_{u\in\mathcal{H}(1)}\left\|(\mathcal{S}+\lambda)^{-\frac{1}{2}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}\left\{w_i^*u(A_i,\boldsymbol{X}_i)-z(A_i;u)\right\}\right]\right\|_{\mathcal{H}_A}^2. \qquad (29)
\end{aligned}$$

  We start with bounding the first term:

$$\begin{aligned}
&\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i}\right)^{\frac{1}{2}}(S_{\boldsymbol{A}}+\lambda\mathcal{I})^{-1}(S+\lambda\mathcal{I})^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H}_A)}\\
=&\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i}\right)^{\frac{1}{2}}(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\left\{I-(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\left(\mathcal{S}-\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i}\right)\right)(\mathcal{S}+\lambda\mathcal{I})^{\frac{1}{2}}\right\}^{-1}\right\|_{\mathcal{L}(\mathcal{H}_A)}\\
\le&\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{\hat{A}_i}\right)^{\frac{1}{2}}(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H}_A)}\\
&\qquad\times\left\|\left\{I-(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\left(\mathcal{S}-\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i}\right)\right)(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\right\}^{-1}\right\|_{\mathcal{L}(\mathcal{H}_A)}.
\end{aligned}$$

From Caponnetto and De Vito (2007), we can show that if $n \geq 2C_\eta \kappa \mathcal{N}(\lambda)/\lambda$, where $C_\eta = 32\log^2(6/\eta)$, and $\lambda \leq \|\mathcal{S}\|_{\mathcal{L}(\mathcal{H}_A)}$,

$$\left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{A_i} \right) \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \leq \frac{1}{2}, \tag{30}$$

with probability at least $1 - 2\eta/3$. And therefore under the same condition, we have

$$\left\| \left\{ I - (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{A_i} \right) \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\}^{-1} \right\|_{\mathcal{L}(\mathcal{H}_A)} \leq 2. \tag{31}$$

with probability at least $1 - 2\eta/3$.

Then we bound $\|(\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i}/n)^{1/2} (\mathcal{S}+\lambda\mathcal{I})^{-1/2} \|_{\mathcal{L}(\mathcal{H}_A)}$. Notice that

$$\begin{aligned}
&\left\| \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right)^{\frac{1}{2}} (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|^2_{\mathcal{L}(\mathcal{H}_A)} = \left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
\leq &\left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \mathcal{S} (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} + \left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
= &\left\| \mathcal{S}^{\frac{1}{2}} (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|^2_{\mathcal{L}(\mathcal{H}_A)} + \left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)}
\end{aligned}$$

For the first component, it's easy to see that

$$\left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \mathcal{S} (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} = \left\| \mathcal{S}^{\frac{1}{2}} (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|^2_{\mathcal{L}(\mathcal{H}_A)} \leq 1$$

because of the spectral theorem.

And by (30), we have

$$\begin{aligned}
&\left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
\leq &\left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{A_i} \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
&\quad + \left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} - \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{A_i} \right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
\leq &\mathcal{O}_{\mathrm{p}}(1) + \frac{1}{\lambda}\mathcal{O}_{\mathrm{p}}\Big(HC_2^{h/2}\kappa^h \Big) \leq \mathcal{O}_{\mathrm{p}}(1).
\end{aligned}$$

The last inequality is due to the condition for κ that $\kappa^h = \mathcal{O}(\lambda)$.

Then we have

$$\left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right)^{\frac{1}{2}} (\mathcal{S} + \lambda \mathcal{I})^{-\frac{1}{2}} \right\|^2_{\mathcal{L}(\mathcal{H}_A)} = \mathcal{O}_{\mathrm{p}}(1) \,.$$

And overall we show that

$$\left\| \left(\frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\hat{A}_i} \right)^{\frac{1}{2}} (S_{\boldsymbol{A}} + \lambda \mathcal{I})^{-1} (S + \lambda \mathcal{I})^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} = \mathcal{O}_{\mathrm{p}}(1) \,. \tag{32}$$

Next, we bound the second term in (29). Take r_i as independent Rademacher random variables. It's easy to see that $\mathbb{E}\{w_i^* u(A_i, X_i) - z(A_i; u)\} = 0$ for every i and $u \in \mathcal{H}(1)$. Due to symmetrization inequality, we have

$$\begin{aligned}
&\mathbb{E} \sup_{u \in \mathcal{H}(1)} \left\| (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{A_i} \left\{ w_i^* u(A_i, \boldsymbol{X}_i) - z(A_i; u) \right\} \right] \right\|^2_{\mathcal{H}_A} \\
\leq & 4\mathbb{E} \sup_{u \in \mathcal{H}(1)} \left\| (\mathcal{S} + \lambda \mathcal{I})^{1/2} \frac{1}{n} \sum_{i=1}^{n} K_{A_i} \{ r_i w_i^* u(A_i, \boldsymbol{X}_i) \} \right\|^2_{\mathcal{H}_A} \\
& \qquad + 4\mathbb{E} \sup_{u \in \mathcal{H}(1)} \left\| (\mathcal{S} + \lambda \mathcal{I})^{1/2} \frac{1}{n} \sum_{i=1}^{n} K_{A_i} r_i z(A_i; u) \right\|^2_{\mathcal{H}_A}
\end{aligned} \tag{33}$$

Let's focus on the first term in (33).

$$\begin{aligned}
&\mathbb{E} \sup_{u \in \mathcal{H}(1)} \left\| (\mathcal{S} + \lambda \mathcal{I})^{1/2} \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{A_i} \{ r_i w_i^* u(A_i, \boldsymbol{X}_i) \} \right\|^2_{\mathcal{H}_A} \\
= & \mathbb{E} \sup_{u \in \mathcal{H}(1)} \left\langle (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{A_i} \{ r_i w_i^* u(A_i, \boldsymbol{X}_i) \}, \right. \\
& \qquad \left. (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_{A_i} \{ r_i w_i^* u(A_i, \boldsymbol{X}_i) \} \right\rangle_{\mathcal{H}_A} \\
= & \mathbb{E} \sup_{u \in \mathcal{H}(1)} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} r_i r_j w_i^* w_j^* u(A_i, \boldsymbol{X}_i) u(A_j, \boldsymbol{X}_j) \\
& \qquad \times \left\langle (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \mathcal{K}_{A_i}, (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \mathcal{K}_{A_j} \right\rangle_{\mathcal{H}_A} \\
\leq & \frac{1}{n^2} \mathbb{E} \sup_{u \in \mathcal{H}(1)} \sum_{i=1}^{n} (w_i^*)^2 u^2(A_i, \boldsymbol{X}_i) \left\langle (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \mathcal{K}_{A_i}, (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \mathcal{K}_{A_i} \right\rangle_{\mathcal{H}_A} \\
& + \frac{1}{n^2} \mathbb{E} \sup_{u \in \mathcal{H}(1)} \sum_{i \neq j} r_i r_j w_i^* w_j^* u(A_i, \boldsymbol{X}_i) u(A_j, \boldsymbol{X}_j) \left\langle (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \mathcal{K}_{A_i}, (\mathcal{S} + \lambda \mathcal{I})^{-1/2} \mathcal{K}_{A_j} \right\rangle_{\mathcal{H}_A} \\
& = (i) + (ii)
\end{aligned} \tag{34}$$

We first deal with the (i). For every $u \in \mathcal{H}(1)$, $\|u\|_\infty = \sup_{(a,\boldsymbol{x})} |\langle K((a,\boldsymbol{x}),(\cdot,\cdot)), u\rangle_{\mathcal{H}}| \leq \sup_{(a,x)} |K((a,\boldsymbol{x}),(a,\boldsymbol{x}))| \leq C_2\kappa$. By contraction inequality and symmetrization inequality, we have

$$
\begin{aligned}
&\mathbb{E} \sup_{u\in\mathcal{H}(1)} \sum_{i=1}^{n} (w_i^*)^2 u^2(A_i, \boldsymbol{X}_i) \left\langle (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_i} \right\rangle_{\mathcal{H}_A} \\
\leq & nC_2^2\kappa^2 \mathbb{E} \left\langle (\mathcal{S}+\lambda\mathcal{I})^{-1/2}K_{A_1}, (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_1} \right\rangle_{\mathcal{H}_A} \\
\leq & nC_2^2\kappa^2 \mathbb{E}\|\mathcal{K}_{A_1}^*(\mathcal{S}+\lambda\mathcal{I})^{-1}\mathcal{K}_{A_1}\|_{\mathcal{L}(\mathcal{H}_A)} \\
\leq & nC_2^2\kappa^2 \mathbb{E}\{\mathrm{Tr}(\mathcal{K}_{A_1}^*(\mathcal{S}+\lambda\mathcal{I})^{-1}\mathcal{K}_{A_1})\} \\
= & nC_2^2\kappa^2 \mathbb{E}\{\mathrm{Tr}((\mathcal{S}+\lambda\mathcal{I})^{-1}\mathcal{K}_{A_1}\mathcal{K}_{A_1}^*)\} = nC_2^2\kappa^2 \int_a \mathrm{Tr}\left\{(\mathcal{S}+\lambda\mathcal{I})^{-1}\mathcal{S}_a\right\} d\rho_A(a) \\
\leq & nC_2^2\kappa^2 \mathrm{Tr}\left\{(\mathcal{S}+\lambda\mathcal{I})^{-1}\mathcal{S}\right\} = nC_2^2\kappa^2\mathcal{N}(\lambda).
\end{aligned}
$$

Next, we study the term (ii). Based on the proof in Proposition 1, we have $K_A(\cdot,\cdot) = \sum_{l_1=1}^{\infty} \nu_{l_1}^A \phi_{l_1}^A(\cdot)\phi_{l_1}^A(\cdot)$, where $\nu_{l_1}^A$ are eigenvalues and $\phi_{l_1}^A(\cdot)$ are eigenfunctions (orthonormal basis of $\mathcal{L}_2(\rho_A)$). $K_X(\cdot,\cdot) = \sum_{l_2=1}^{\infty} \nu_{l_2}^X \phi_{l_2}^X(\cdot)\phi_{l_2}^X(\cdot)$, where $\nu_{l_2}^X$ are eigenvalues and $\phi_{l_2}^X(\cdot)$ are eigenfunctions (orthonormal basis of $\mathcal{L}_2(\rho_X)$. Since the reproducing kernel for $\mathcal{H}$ is $K((\cdot,\star),(\cdot,\star)) = K_A(\cdot,\cdot)K_X(\star,\star)$, we have $K((\cdot,\star),(\cdot,\star)) = \sum_{l=1}^{\infty} \nu_l\phi_l(\cdot,\star)\phi_l(\cdot,\star)$, where $\nu_l = \nu_{l_1}^A\nu_{l_2}^X$ and $\phi_l(\cdot,\star) = \phi_{l_1}^A(\cdot)\phi_{l_2}^X(\star)$ for some l_1, l_2 such that ν, $l = 1,\ldots,\infty$ is nonincreasing. Take $\Phi(\cdot,\star) = \{\sqrt{\nu}\phi_l(\cdot,\star)\}_{l=1}^{\infty}$. $\mathcal{H}(1) = \{u(\cdot,\star) = \langle\beta, \Phi(\cdot,\star)\rangle : \sum_{l=1}^{\infty}\beta_l^2 \leq 1\}$. Take $\mathcal{E}(1) = \{\beta : \sum_{l=1}^{\infty}\beta_l^2 \leq 1\}$, then

$$
\begin{aligned}
(ii) \leq & \mathbb{E} \sup_{\beta\in\mathcal{E}(1)} \left\{\sum_{l=1}^{\infty}(\beta_l)^2 \sum_{l'=1}^{\infty}(\beta_{l'})^2\right\}^{\frac{1}{2}} \\
& \times \left[\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty}\left\{\sum_{i\neq j}\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\mathcal{K}_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\mathcal{K}_{A_j}\right\rangle_{\mathcal{H}_A} r_i r_j w_i^* w_j^* \sqrt{\nu_l}\phi_l(A_i,\boldsymbol{X}_i)\sqrt{\nu_{l'}}\phi_{l'}(A_j,\boldsymbol{X}_j)\right\}^2\right]^{\frac{1}{2}} \\
\leq & \left[\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty}\sum_{i\neq j}\mathbb{E}\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\mathcal{K}_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\mathcal{K}_{A_j}\right\rangle_{\mathcal{H}_A}^2 (w_i^*)^2(w_j^*)^2\nu_l\phi_l^2(A_i,\boldsymbol{X}_i)\nu_{l'}\phi_{l'}^2(A_j,\boldsymbol{X}_j)\right]^{\frac{1}{2}} \\
\leq & \left[\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty}\sum_{i\neq j}\mathbb{E}\left\{(w_i^*)^2\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_i}\right\rangle_{\mathcal{H}_A}\nu_l\phi_l^2(A_i,\boldsymbol{X}_i)\right\}\right. \\
& \left.\mathbb{E}\left\{(w_j^*)^2\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_j}, (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_j}\right\rangle_{\mathcal{H}_A}\nu_{l'}\phi_{l'}^2(A_j,\boldsymbol{X}_j)\right\}\right]^{\frac{1}{2}} \\
= & \left[\sum_{i\neq j}\mathbb{E}\left\{(w_i^*)^2\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_i}\right\rangle_{\mathcal{H}_A}\left(\sum_{l=1}^{\infty}\nu_l\phi_l^2(A_i,\boldsymbol{X}_i)\right)\right\}\right. \\
& \left.\mathbb{E}\left\{(w_j^*)^2\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_j}, (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}K_{A_j}\right\rangle_{\mathcal{H}_A}\left(\sum_{l'=1}^{\infty}\nu_{l'}\phi_{l'}^2(A_j,\boldsymbol{X}_j)\right)\right\}\right]^{\frac{1}{2}}.
\end{aligned}
$$

The first inequality by adopting Cauchy Schwarz inequality. The second inequality is due to that $r_i,\ i = 1,\ldots,n$ are all independent Rademacher random variables. The

third inequality is because that $(A_i, X_i),\ i = 1, \ldots, n$ are independent pairs and

$$\begin{aligned}
&\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_j}\right\rangle_{\mathcal{H}_A}^2 \\
\leq &\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_i}\right\rangle_{\mathcal{H}_A} \left\langle (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_j}, (\mathcal{S}+\lambda\mathcal{I})^{-1/2}\mathcal{K}_{A_j}\right\rangle_{\mathcal{H}_A}
\end{aligned}$$

Note that

$$\sum_{l=1}^{\infty} \nu_l \phi_l^2(A_i, \boldsymbol{X}_i) = K_A(A_i, A_i) K_X(\boldsymbol{X}_i, \boldsymbol{X}_i) \leq \kappa C_2.$$

Then

$$\begin{aligned}
&\mathbb{E}\left\{ (w_i^*)^2 \left\langle (\mathcal{S}+\lambda\mathcal{I})^{-1/2}K_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-1/2}K_{A_i}\right\rangle_{\mathcal{H}_A} \left(\sum_{l=1}^{\infty}\nu_l\phi_l^2(A_i, X_i)\right)\right\} \\
\leq &C_1^2\kappa C_2 \mathbb{E}\left\langle (\mathcal{S}+\lambda\mathcal{I})^{-1/2}K_{A_i}, (\mathcal{S}+\lambda\mathcal{I})^{-1/2}K_{A_i}\right\rangle_{\mathcal{H}_A} \\
\leq &C_1^2\kappa C_2 \mathcal{N}(\lambda)
\end{aligned}$$

Now we prove that

$$(ii) \leq \frac{C_1^2\kappa C_2}{n^2}\left\{\sqrt{n(n-1)}\mathcal{N}(\lambda)\right\} \leq \frac{C_1^2\kappa C_2}{n}\mathcal{N}(\lambda).$$

Combine the bound of (i) and (ii) into (34), we have

$$\mathbb{E}\sup_{u\in\mathcal{H}(1)}\left\|(\mathcal{S}+\lambda\mathcal{I})^{1/2}\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}\{r_i w_i^* u(A_i, \boldsymbol{X}_i)\}\right\|_{\mathcal{H}_A}^2 \leq \frac{C_2^2\kappa^2\mathcal{N}(\lambda)}{n} + \frac{C_1^2\kappa C_2}{n}\mathcal{N}(\lambda) \tag{35}$$

Next we deal with the second term in (33). For $u \in \mathcal{H}(1)$, by previous construction of $\mathcal{H}(1)$, we can express $u(a,x) = \sum_{l_1}\sum_{l_2}\beta_{l_1}^A\beta_{l_2}^X\sqrt{\nu_{l_1}^A\nu_{l_2}^X}\phi_{l_1}^A(a)\phi_{l_2}^X(x)$ for some coefficients $\beta_{l_1}^A$ and $\beta_{l_2}^X$. Then

$$\begin{aligned}
\|z(\cdot;u)\|_{\mathcal{H}_A}^2 = \sum_{l_1}(\beta_{l_1}^A)^2\left\{\mathbb{E}\left(\sum_{l_2}\sqrt{\nu_{l_2}^X}\beta_{l_2}^X\phi_{l_2}^X(\boldsymbol{X})\right)\right\}^2 &\leq \sum_{l_1}(\beta_{l_1}^A)^2\mathbb{E}\left(\sum_{l_2}\sqrt{\nu_{l_2}^X}\beta_{l_2}^X\phi_{l_2}^X(\boldsymbol{X})\right)^2 \\
&\leq \sum_{l_1}(\beta_{l_1}^A)^2\sum_{l_2}\nu_{l_2}^X(\beta_{l_2}^X)^2 \leq \left(\max_{l_2}\nu_{l_2}^X\right)\sum_{l_1}\sum_{l_2}(\beta_{l_1}^A)^2(\beta_{l_2}^X)^2 = \nu_1^X
\end{aligned}$$

Then $\{z(\cdot;u) : \|u\|_{\mathcal{H}} \le 1\} \subset \{z : \|z\|_{\mathcal{H}_A} \le \nu_1^X\}$, we can follow previous strategy and prove that

$$\begin{aligned}
&\mathbb{E} \sup_{u\in\mathcal{H}(1)} \left\| (\mathcal{S}+\lambda\mathcal{I})^{1/2} \frac{1}{n}\sum_{i=1}^n K_{A_i} r_i z(A_i;u) \right\|_{\mathcal{H}_A}^2 \qquad (36)\\
&\le (\nu_1^X)^2 \mathbb{E} \sup_{z\in\{z:\|z\|_{\mathcal{H}_A}\le 1\}} \left\| (\mathcal{S}+\lambda\mathcal{I})^{1/2} \frac{1}{n}\sum_{i=1}^n K_{A_i} r_i z(A_i) \right\|_{\mathcal{H}_A}^2\\
&\le \frac{(\nu_1^X)^2\kappa^2\mathcal{N}(\lambda)}{n} + \frac{(\nu_1^X)^2 C_1^2\kappa}{n}\mathcal{N}(\lambda)
\end{aligned}$$

And overall, there exists a constant $c_1 > 0$ depending on $\nu_1^X, \kappa, C_1, C_2$, such that

$$\mathbb{E} \sup_{u\in\mathcal{H}(1)} \left\| (\mathcal{S}+\lambda\mathcal{I})^{-1/2} \left[\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{A_i} \left\{ w_i^* u(A_i, \boldsymbol{X}_i) - z(A_i;u) \right\} \right] \right\|_{\mathcal{H}_A}^2 \le \frac{c_1\mathcal{N}(\lambda)}{n}.$$

And combine the result with (32), we have

$$\sup_{u\in\mathcal{H}(1)} \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} + \lambda\mathcal{I} \right)^{-1} \left[\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{A_i} \left\{ w_i^* u(A_i, \boldsymbol{X}_i) - z(A_i;u) \right\} \right] \right\|_n^2 = \mathcal{O}_{\mathrm{p}}\left(\frac{\mathcal{N}(\lambda)}{n} \right).$$

- Next, we consider to bound (27). By the above arguments,

$$\begin{aligned}
&\sup_{u\in\mathcal{H}(1)} \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} + \lambda\mathcal{I} \right)^{-1} \left[\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{A_i} z(A_i;u) \right] - z(\cdot;u) \right\|_n^2\\
\le &\nu_1^X \sup_{z\in\{z:\|z\|_{\mathcal{H}_A}\le 1\}} \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} + \lambda\mathcal{I} \right)^{-1} \left[\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{A_i} z(A_i) \right] - z \right\|_n^2
\end{aligned}$$

Take $z^\lambda = (\mathcal{S}+\lambda\mathcal{I})^{-1} S z$. We can verify that

$$\begin{aligned}
\|z^\lambda - z\|_{\mathcal{H}_A}^2 &= \left\| \left[(\mathcal{S}+\lambda\mathcal{I})^{-1} S - \mathcal{I} \right] z \right\|_{\mathcal{H}_A}^2 = \sum_{l=1}^\infty \left[\frac{\nu_l^A}{\nu_l^A + \lambda} - 1 \right]^2 \langle z, \phi_l^A \rangle_{\mathcal{H}_A}^2\\
&\le \sum_{l=1}^\infty \langle z, \phi_l^A \rangle_{\mathcal{H}_A}^2 = \|z\|_{\mathcal{H}_A}^2.\\
\|z^\lambda - z\|_{\mathcal{L}_2(\rho_A)}^2 &= \left\| \sqrt{\mathcal{S}} \left[(\mathcal{S}+\lambda\mathcal{I})^{-1} S \right] z - z \right\|_{\mathcal{H}_A}^2 = \sum_{l=1}^\infty \nu_l^A \left[\frac{\nu_l^A}{\nu_l^A + \lambda} - 1 \right]^2 \langle z, \phi_l^A \rangle_{\mathcal{H}_A}^2\\
&= \sum_{l=1}^\infty \left[\frac{\lambda}{\sqrt{\nu_l^A} + \lambda/\sqrt{\nu_l^A}} \right]^2 \langle z, \phi_l^A \rangle_{\mathcal{H}_A}^2 \le \sum_{l=1}^\infty \left(\frac{\lambda}{\sqrt{2\lambda}} \right)^2 \langle z, \phi_l^A \rangle_{\mathcal{H}_A}^2 = \lambda/2 \|z\|_{\mathcal{H}_A}^2.
\end{aligned}$$

Next, we derive the bound for

$$\sup_{z\in\{z:\|z\|_{\mathcal{H}_A}\leq 1\}} \|z^\lambda - z\|_n^2.$$

Take $z' = z^\lambda - z$, given that $\|z'\|_{\mathcal{H}_A} \leq \sqrt{c_2}\|z\|_{\mathcal{H}_A} \leq \sqrt{c_2}$ and $\|z'\|_{\mathcal{L}_2(\rho_A)} \leq \sqrt{c_3\lambda}\|z\|_{\mathcal{H}_A} \leq \sqrt{c_3\lambda}$ for some positive constants c_2 and c_3. Follow the same proof of Lemma 42 in Mendelson (2002), we can show that the Rademacher complexity

$$\mathbb{E} \sup_{z'\in\{z:\|z\|_{\mathcal{H}_A}\leq\sqrt{c_2},\|z\|_{\mathcal{L}_2(\rho_A)}\leq\sqrt{c_3\lambda}\}} \left|\frac{1}{n}\sum_{i=1}^n r_i z'(A_i)\right|^2 \leq \frac{c_4}{n}\left(\sum_{l=1}^\infty \min\{\nu_l^A, \lambda\}\right)$$

for some constant $c_4 > 0$ depending on c_2 and c_3. Next, we apply Corollary 2.2 in Bartlett et al. (2005), we can verify there exists a constant $b > 0$ such that $\|z'\|_\infty \leq b$ for any $\|z'\|_{\mathcal{H}_A} \leq \sqrt{c_2}$. Then for any $x > 0$, if $\lambda \geq 10b\{c_4\sum_{l=1}^\infty \min\{\nu_l^A, \lambda\}/n\}^{1/2} + 11b^2x/n$, we have

$$\{z' \in \{z : \|z\|_{\mathcal{H}_A} \leq \sqrt{c_2}\} : \|z'\|^2_{\mathcal{L}_2(\rho_A)} \leq c_3\lambda\} \subseteq \{z' \in \{z : \|z\|_{\mathcal{H}_A} \leq \sqrt{c_2}\} : \|z'\|_n^2 \leq 2c_3\lambda\},$$

with probability at least $1 - \exp(-x)$. Note that $v_l^A = t_l$, then as long as $\sqrt{\sum_{l=1}^\infty \min\{t_l, \lambda\}}/(\sqrt{n}\lambda) = \mathcal{O}_{\mathrm{p}}(1)$, we have $\|z'\|_n^2 = \mathcal{O}_{\mathrm{p}}(\lambda)$. The above inequalities also holds for $\hat{A}_i$, $i = 1, \ldots, n$ as $\hat{A}_i$ are independent samples from $\mathcal{A}$. Then we have the following the inequality

$$\begin{aligned}
&\sup_{z\in\{z:\|z\|_{\mathcal{H}_A}\leq 1\}} \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1} \left[\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{A_i} z(A_i)\right] - z \right\|_n^2 \\
\leq & 2 \sup_{z\in\{z:\|z\|_{\mathcal{H}_A}\leq 1\}} \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1} \left(\mathcal{S} - \frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i}\right)\left(z^\lambda - z\right) \right\|_n^2 \\
& + 2 \sup_{z\in\{z:\|z\|_{\mathcal{H}_A}\leq 1\}} \|z^\lambda - z\|_n^2 \\
\leq & \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1} (S + \lambda\mathcal{I})^{\frac{1}{2}} \right\|^2_{\mathcal{L}(\mathcal{H}_A)} \\
& c_2 \sup_{v\in\{z:\|z\|_{\mathcal{H}_A}\leq 1\}} \left\| (\mathcal{S} + \lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i}\right) v \right\|^2_{\mathcal{H}_A} + c_3\lambda
\end{aligned} \tag{37}$$

The operator norm in (37) can be bounded via (32). It remains to bound

$$\sup_{v\in\{z:\|z\|_{\mathcal{H}_A}\leq 1\}} \left\| (\mathcal{S} + \lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S} - \frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i}\right) v \right\|^2_{\mathcal{H}_A}.$$

Note that $\mathbb{E}\mathcal{S}_{A_i} = \mathcal{S}$, then we can apply the symmetrization equality again

$$\sup_{v\in\{z:\|z\|_{\mathcal{H}_A}\le 1\}} \left\|(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\left(\mathcal{S}-\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i}\right)v\right\|_{\mathcal{H}_A}^2$$
$$\le 4\mathbb{E}\sup_{v\in\{z:\|z\|_{\mathcal{H}_A}\le 1\}} \left\|(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} r_i v\right\|_{\mathcal{H}_A}^2,$$

where r_i, $i = 1, \ldots, n$ are independent Rademacher random variables.

Follow the similar proof in proving term (26), we can show that

$$\mathbb{E}\sup_{v\in\{z:\|z\|_{\mathcal{H}_A}\le 1\}} \left\|(\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}}\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} r_i v\right\|_{\mathcal{H}_A}^2 \le c_5\frac{\mathcal{N}(\lambda)}{n}$$

for some constant $c_5 > 0$ depending on κ. And overall, we prove that

$$(27) = \mathcal{O}_{\mathrm{p}}\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda\right)$$

- Last, we bound the term (28). First, fixed any $a \in \mathcal{A}$, we derive the bound for $\mathbb{E}\sup_{u\in\mathcal{H}(1)}[z(a;u) - \sum_{j=1}^n u(a, \boldsymbol{X}_j)/n]^2$.

  By symmetrization inequality, we have

$$\mathbb{E}\sup_{u\in\mathcal{H}(1)}\left[z(a;u) - \frac{\sum_{j=1}^n u(a,\boldsymbol{X}_j)}{n}\right]^2 \le 4\mathbb{E}\sup_{u\in\mathcal{H}(1)}\left[\frac{1}{n}\sum_{j=1}^n r_j u(a,\boldsymbol{X}_j)\right]^2$$

  Again by previous construction of $\mathcal{H}(1)$, we have $\mathcal{H}(1) = \{u(\cdot,\star) = \langle\beta, \Phi(\cdot,\star)\rangle : \sum_{l=1}^\infty \beta_l^2 \le 1\}$. Take $\mathcal{E}(1) = \{\beta : \sum_{l=1}^\infty \beta_l^2 \le 1\}$.

$$\mathbb{E}\sup_{u\in\mathcal{H}(1)}\left[\frac{1}{n}\sum_{j=1}^n r_i u(a,\boldsymbol{X}_j)\right]^2 \le \mathbb{E}\sup_{\beta\in\mathcal{E}(1)}\left[\frac{1}{n}\sum_{j=1}^n r_j\left\{\sum_l \beta_l\sqrt{\nu_l}\phi_l(a,\boldsymbol{X}_j)\right\}\right]^2$$
$$\le \mathbb{E}\sup_{\beta\in\mathcal{E}(1)}\left(\sum_l \beta_l^2\right)\left[\frac{1}{n}\sum_{j=1}^n r_j\sqrt{\nu_l}\phi_l(a,\boldsymbol{X}_j)\right]^2 \le \sum_l \nu_l\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n r_j\phi_l(a,\boldsymbol{X}_j)\right]^2$$
$$\le \frac{1}{n}\sum_l \nu_l\mathbb{E}\phi_l^2(a,\boldsymbol{X}_1) = \frac{1}{n}K_A(a,a)\mathbb{E}K_X(\boldsymbol{X}_1,\boldsymbol{X}_1) \le \frac{1}{n}\kappa C_2 K_A(a,a).$$

  Then we can show that

$$\mathbb{E}\left\{\sup_{u\in\mathcal{H}(1)}\left\|z(\cdot;u) - \frac{1}{n}\sum_{j=1}^n u(\cdot,\boldsymbol{X}_j)\right\|_n^2\right\} \le \mathbb{E}\left(\mathbb{E}\left\{\sup_{u\in\mathcal{H}(1)}\left[z(A;u) - \frac{1}{n}\sum_{j=1}^n u(A,\boldsymbol{X}_j)\right]^2 \mid A\right\}\right)$$
$$\lesssim \frac{1}{n}\kappa K_A(A,A) \lesssim \frac{1}{n} = \mathcal{O}_{\mathrm{p}}\left(\frac{1}{\sqrt{n}}\right).$$

Combine the bounds derived for (26), (27) and (28), the bound for $\sup_{u\in\mathcal{H}(1)} Q(w^*,\lambda,\|\cdot\|_n,u)$ follows.

To bound the penalty term $R(w^*,\lambda)$, note that

$$
\begin{aligned}
R(\boldsymbol{w}^*,\lambda,) &= \frac{1}{n^2}\sum_{i=1}^n (w_i^*)^2 \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1} \mathcal{K}_{\hat{A}_i} \right\|_n^2 \\
&\le \frac{C_1}{n^2}\sum_{i=1}^n \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1} \mathcal{K}_{\hat{A}_i} \right\|_n^2 \\
&\le \frac{C_1}{n^2}\sum_{i=1}^n \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1} (\mathcal{S}+\lambda\mathcal{I})^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)}^2 \left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \mathcal{K}_{\hat{A}_i} \right\|_{\mathcal{H}_A}^2 .
\end{aligned}
$$

$$
\begin{aligned}
&\left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1} (\mathcal{S}+\lambda\mathcal{I})^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
\le &\left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i} + \lambda\mathcal{I}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i} - \frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i}\right) \right. \\
&\qquad \left. \times (\mathcal{S}_{\boldsymbol{A}} + \lambda\mathcal{I})^{-1} (\mathcal{S}+\lambda\mathcal{I})^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
&+ \left\| \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{A_i} + \lambda\mathcal{I}\right)^{-1} (\mathcal{S}+\lambda\mathcal{I})^{\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
&= \mathcal{O}_{\mathrm{p}}(\lambda^{-1}\kappa^h) + \mathcal{O}_{\mathrm{p}}(1) = \mathcal{O}_{\mathrm{p}}(1).
\end{aligned}
$$

The last two equalities are due to (32) and the condition of κ.

$$
\begin{aligned}
\left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \mathcal{K}_{\hat{A}_i} \right\|_{\mathcal{H}_A}^2 &= \left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \left(\mathcal{S}_{\hat{A}_i} - \mathcal{S}_{A_i}\right) (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
&\quad + \left\| (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \mathcal{S}_{A_i} (\mathcal{S}+\lambda\mathcal{I})^{-\frac{1}{2}} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
&\le \left\| \left(\mathcal{S}_{\hat{A}_i} - \mathcal{S}_{A_i}\right) (\mathcal{S}+\lambda\mathcal{I})^{-1} \right\|_{\mathcal{L}(\mathcal{H}_A)} + \left\| \mathcal{S}_{A_i} (\mathcal{S}+\lambda\mathcal{I})^{-1} \right\|_{\mathcal{L}(\mathcal{H}_A)} \\
&\le \frac{1}{\lambda} \left\| \mathcal{S}_{\hat{A}_i} - \mathcal{S}_{A_i} \right\|_{\mathcal{L}(\mathcal{H}_A)} + \left\| \mathcal{S}_{A_i} (\mathcal{S}+\lambda\mathcal{I})^{-1} \right\|_{\mathcal{L}(\mathcal{H}_A)}
\end{aligned}
$$

It's easy to verify that

$$
\mathbb{E}\|\mathcal{S}_{A_i}(\mathcal{S}+\lambda\mathcal{I})^{-1}\|_{\mathcal{L}(\mathcal{H}_A)} \le \mathbb{E}\left\{\mathrm{Tr}\left((\mathcal{S}+\lambda\mathcal{I})^{-1}\mathcal{K}_{A_i}\mathcal{K}_{A_i}^*\right)\right\} \le \mathrm{Tr}\{(\mathcal{S}+\lambda\mathcal{I})^{-1}\mathcal{S}\} = \mathcal{N}(\lambda).
$$

Then combine with all the bounds, we obtain

$$
R(\boldsymbol{w}^*,\lambda) = \mathcal{O}_{\mathrm{p}}\left(\frac{\mathcal{N}(\lambda)}{n}\right) + \mathcal{O}_{\mathrm{p}}\left(\frac{\kappa^h}{\lambda}\right) \le \mathcal{O}_{\mathrm{p}}\left(\frac{\mathcal{N}(\lambda)}{n}\right).
$$

The last inequality is due to the condition of κ.

Now we are ready to bound $\sup_{u\in\mathcal{H}(1)} Q(\hat{\boldsymbol{w}},\lambda,u)$ and $R(\hat{\boldsymbol{w}},\lambda)$. Since $\hat{\boldsymbol{w}}$ is the solution of (14), by the basic inequality, we have

$$\sup_{u\in\mathcal{H}(1)} Q(\hat{\boldsymbol{w}},\lambda,u)+\eta R(\hat{\boldsymbol{w}},\lambda)\leq \sup_{u\in\mathcal{H}(1)} Q(\boldsymbol{w}^*,\lambda,u)+\eta R(\boldsymbol{w}^*,\lambda).$$

Therefore, we have

$$\sup_{u\in\mathcal{H}(1)} Q(\hat{\boldsymbol{w}},\lambda,u)\leq \left\{\sup_{u\in\mathcal{H}(1)} Q(\boldsymbol{w}^*,\lambda,u)+\eta R(\boldsymbol{w}^*,\lambda)\right\}=\mathcal{O}_{\mathrm{p}}\left[(1+\eta)\frac{\mathcal{N}(\lambda)}{n}+\lambda\right],$$
$$R(\hat{\boldsymbol{w}},\lambda)\leq \eta^{-1}\left\{\sup_{u\in\mathcal{H}(1)} Q(\boldsymbol{w}^*,\lambda,u)+\eta R(\boldsymbol{w}^*,\lambda)\right\}=\mathcal{O}_{\mathrm{p}}\left[\frac{\mathcal{N}(\lambda)}{n}+\eta^{-1}\left(\lambda+\frac{\mathcal{N}(\lambda)}{n}\right)\right].$$

■

Proof [Proof of Theorem 3] By Theorem 2, we can derive that

$$Q(\hat{\boldsymbol{w}},\lambda,m)=\mathcal{O}_{\mathrm{p}}\left[\|m\|_{\mathcal{H}}^2\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda+\eta\frac{\mathcal{N}(\lambda)}{n}\right)\right],$$
$$R(\hat{\boldsymbol{w}},\lambda,)=\mathcal{O}_{\mathrm{p}}\left[\eta^{-1}\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda\right)+\frac{\mathcal{N}(\lambda)}{n}\right].$$

From the decomposition, we have

$$\begin{aligned}\|\hat{\tau}-\tau\|_n &= \left\|\left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}+\lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\hat{A}_i}\hat{w}_i Y_i\right)-\tau\right\|_n \\ &= \left\|\left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}+\lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\hat{A}_i}\hat{w}_i m(\hat{A}_i,\boldsymbol{X}_i)\right)-\frac{1}{n}\sum_{j=1}^n m(\cdot,\boldsymbol{X}_j)\right\|_n \qquad (38)\\ &\quad+\left\|\frac{1}{n}\sum_{j=1}^n m(\cdot,\boldsymbol{X}_j)-\tau\right\|_n \qquad (39)\\ &\quad+\left\|\left(\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\hat{A}_i}+\lambda\mathcal{I}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{K}_{\hat{A}_i}\hat{w}_i\epsilon_i\right)\right\|_n \qquad (40)\end{aligned}$$

$$(38)=\{Q(\hat{\boldsymbol{w}},\lambda,m)\}^{1/2}=\mathcal{O}_{\mathrm{p}}\left[\|m\|_{\mathcal{H}}\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda+\eta\frac{\mathcal{N}(\lambda)}{n}\right)^{1/2}\right].$$

From the bound of (28) in Theorem 2, we have

$$(39)=\mathcal{O}_{\mathrm{p}}\left(\frac{\|m\|_{\mathcal{H}}}{\sqrt{n}}\right).$$

Notice that under Assumption 3,

$$\mathbb{E}\left\{\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_{A_i}+\lambda\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_{A_i}\hat{w}_i\epsilon_i\right)\right\|_n^2 \mid A_i, i=1,\ldots,n\right\} \leq \sigma_0^2 R(\hat{\boldsymbol{w}},\lambda)=\sigma_0^2\,\mathcal{O}_{\mathrm{p}}\left[\eta^{-1}\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda\right)+\frac{\mathcal{N}(\lambda)}{n}\right].$$

Then

$$(40)=\mathcal{O}_{\mathrm{p}}\left(\sigma_0\eta^{-1/2}\left(\frac{\mathcal{N}(\lambda)}{n}+\lambda\right)^{1/2}+\left(\frac{\mathcal{N}(\lambda)}{n}\right)^{1/2}\right).$$

Under Assumption 7 and the conditions of η and λ stated in Theorem 3, we note that $\mathcal{N}(\lambda)\asymp\lambda^{-1/b}$. Follow the proof of Lemma S7 in Wang et al. (2020), to satisfy the condition $\sqrt{\sum_{l=1}^{\infty}\min\{t_l,\lambda\}/(n\lambda^2)}=\mathcal{O}(1)$, we need $n\lambda^{-(1+1/b)}=\mathcal{O}(1)$, which is the same condition for $\mathcal{N}(\lambda)(\lambda n)^{-1}=\mathcal{O}(1)$. Then we can see that when $\lambda\asymp n^{-b/(1+b)}$, the conditions are satisfied and $\lambda\asymp(\mathcal{N}(\lambda)/\sqrt{n}$.

The bound of $\|\hat{\tau}-\tau\|_n$ follows.

Appendix C. Additional Simulation

In this section, we conduct an additional simulation study where $\boldsymbol{X}$ are dependent and A has a more complex dependence structure on $\boldsymbol{X}$. More specifically, we introduce the latent variable $\boldsymbol{Z}=[Z^{(1)},Z^{(2)},Z^{(3)},Z^{(4)}]^\intercal\in\mathbb{R}^4$ follow the multivariate normal distribution with the zero mean vector and the identity covariance matrix. The functional treatment A is generated by $A(t)=\sum_{k=1}^{4}A^{(k)}\sqrt{2}\sin(2\pi kt)$, $t\in[0,1]$, where $A^{(1)}\mid\boldsymbol{Z}\sim N(4Z^{(1)},1)$, $A^{(2)}\mid\boldsymbol{Z}\sim N(2\sqrt{3}Z^{(2)},1)$, $A^{(3)}\mid\boldsymbol{Z}\sim N(2\sqrt{2}Z^{(3)},1)$ and $A^{(4)}\mid\boldsymbol{Z}\sim N(2Z^{(4)},1)$. Then we generate confounders $\boldsymbol{X}=[X^{(1)},X^{(2)},X^{(3)},X^{(4)}]^\intercal\in\mathbb{R}^4$ as $\boldsymbol{X}=[\exp(Z^{(1)}/2),Z^{(2)}/(1+exp(Z^{(1)})),(Z^{(1)}Z^{(3)}/25+0.6)/2,(Z^{(2)}+Z^{(4)})/2]$. The introduction of the latent variable $\boldsymbol{Z}$ is to ensure the orthogonality of the principal components $A^{(l)}$, $l=1,\ldots,4$, while preserving complex dependencies both among the covariates X and between A and X. The outcome Y is generated by $Y\mid(A,\boldsymbol{X})\sim N(m(A,\boldsymbol{X}),1)$, where we consider three choices for m as follows. Let $\Omega(\boldsymbol{x})=z^{(2)}(z^{(1)})^2+(z^{(4)})^2\sin(2z^{(3)})$ where $\boldsymbol{z}=[z^{(1)},z^{(2)},z^{(3)},z^{(4)}]^\intercal$, and $\mu(t)=2\sqrt{2}\sin(2\pi t)+\sqrt{2}\cos(2\pi t)+\sqrt{2}\sin(4\pi t)/2+\sqrt{2}\cos(4\pi t)/2$. Here $\Omega(\cdot)$ is a function of $\boldsymbol{X}$ as $\boldsymbol{Z}$ itself is a function of $\boldsymbol{X}$.

- Setting 1: We let $m(a,\boldsymbol{x})=15\Omega(\boldsymbol{x})+\int_{t=0}^{1}a(t)\mu(t)\mathrm{d}t$. In this case, the treatment effect $\tau(a)$ is linear in a in the sense

  $$\tau(a)=\int_{t=0}^{1}a(t)\mu(t)\mathrm{d}t,$$

  and m is additive in $\Psi(\boldsymbol{x})$ and a.

- Setting 2: We let $m(a,\boldsymbol{x})=10\Omega(\boldsymbol{x})+0.5(a^{(1)})^2+4\sin(a^{(1)})$. Here m is additive in $\Psi(\boldsymbol{x})$ and a. Then the treatment effect is

  $$\tau(a)=0.5(a^{(1)})^2+4\sin(a^{(1)}),$$

  which is nonlinear in a.

- Setting 3: We let $m(a, \boldsymbol{x}) = [1 + 2/3\Omega(\boldsymbol{x})][0.5(a^{(1)})^2 + 4\sin(a^{(1)})]$. In this case, the treatment effect $\tau(a)$ has the same form as in Setting 3, but $\Psi(\boldsymbol{z})$ interacts with a in m.

Same as the description in Section 5.1, we compare estimators mentioned in Section 5.1 and use two evaluation metrics introduced in Section 5.1 to present the simulation results.

Table 4: Empirical MSEs for different estimators under three different simulation settings described in Appendix C. Values in the parentheses are the standard errors of MSEs.

	Setting 1	Setting 2	Setting 3
CFB	53.97 (1.33)	130.75 (2.93)	154.65 (4.72)
FCBPS	274.1 (38.85)	296.77 (32.85)	1350.72 (455.9)
NPFCBPS	103.88 (3.06)	211.97 (3.72)	254.24 (10.72)
REG	83.38 (6.27)	95.32 (2.36)	164.92 (8.32)
FLM	243.06 (18.23)	603.85 (23.01)	1780.32 (219.2)
FGAM	362.37 (19.71)	238.03 (9.41)	2137.04 (266.4)
NW	682.64 (21.27)	313.74 (9.27)	2732.82 (278.42)

Table 5: Out-of-Sample MSEs for different estimators under three different simulation settings described in Appendix C. Values in the parentheses are the standard errors of MSEs.

	Setting 1	Setting 2	Setting 3
CFB	58.28 (2.06)	144.96 (5.57)	186.09 (11.48)
FCBPS	267.13 (32.91)	310.52 (27.9)	1304.48 (401.61)
NPFCBPS	103.17 (3.13)	229.3 (5.58)	270.22 (11.9)
REG	86.42 (7.96)	105.57 (3.48)	160.91 (8.45)
FLM	245.29 (17.65)	632.48 (25.52)	1821.75 (213.19)
FGAM	392.3 (26.8)	257.24 (12)	2890.63 (647.76)
NW	523.46 (15.75)	254.91 (7.58)	1215.51 (74.72)

Appendix D. Sensitivity Analysis for the Cutoff

In this section, we vary the cutoff values to 50, 75 and 200. For cutoff values of 50 and 75, we observe two and one density curves, respectively, that differ significantly in shape from the rest. The highest estimated densities for these curves exceed 0.015, while the rest remain below 0.0125. To avoid the influence of outliers, we exclude these anomalous samples from the analysis. After the pre-processing, the sample sizes are 430, 429 and 423 for cutoff values 50, 75, 200 respectively.

As with Figure 2, which was produced using a cutoff value of 100, we present the activity profiles along with their estimated BMI values in Figures 3, 4, and 5, corresponding to cutoff values of 50, 75, and 200, respectively. Similar patterns across all methods in Figures 3-5 are observed, consistent with those shown in Figure 2.

Tables 6, 7 and 8 provide the estimated BMI values for three representative density curves with cutoff values of 50, 75, and 200, respectively. We obtain the same conclusion that all seven estimators provide the same order of BMI values for these three curves, indicating that a more active person tends to have a lower BMI. The results from `CFB`, `REG` and `FCBPS` are more similar, and show a greater difference in BMI values between the curve with ID 33387 and 32104.

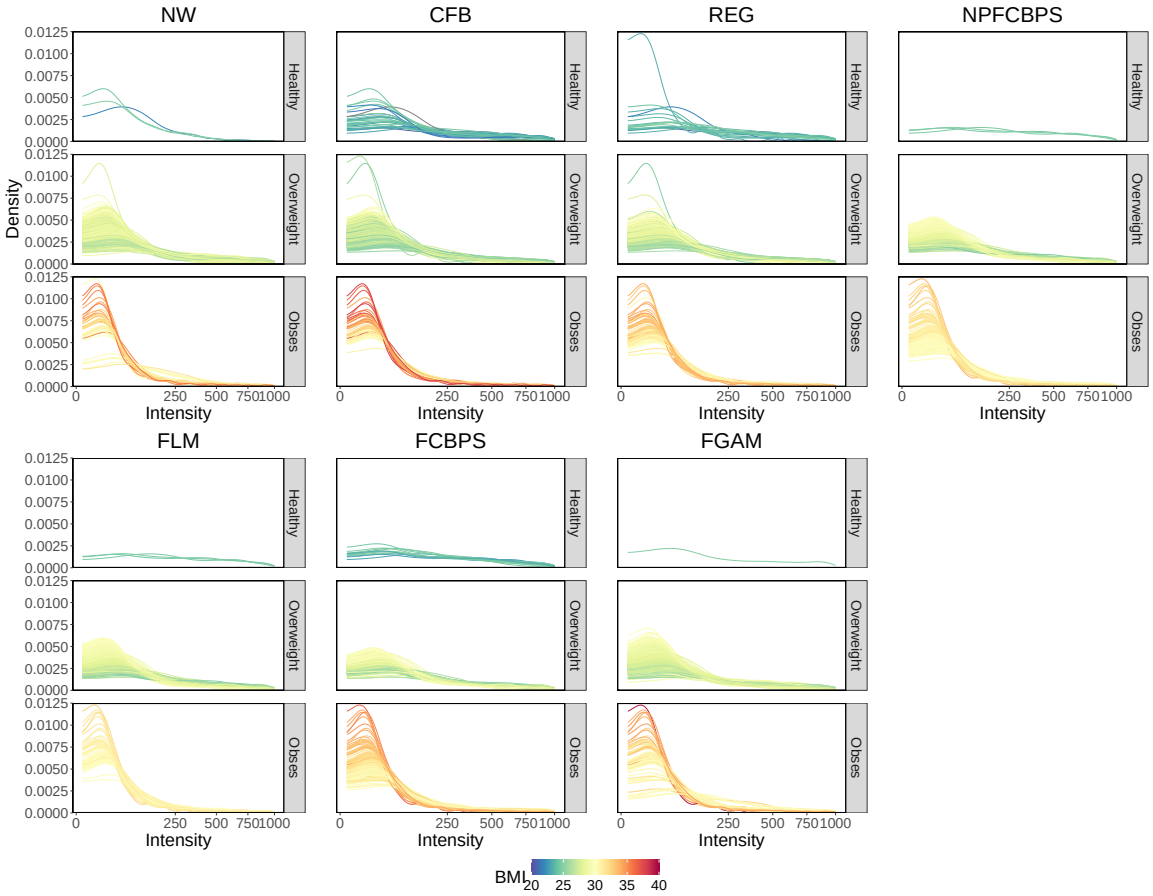


Figure 3: Density curves divided into different categories by their corresponding estimated BMI values using different estimators with cutoff value 50. See details in Figure 2.

Table 6: Estimated BMI values for the three representative density curves with cutoff value 50. See details in Table 3.

SEQN	`CFB`	`FCBPS`	`NPFCBPS`	`REG`	`FLM`	`FGAM`	`NW`
32104	30.30	31.50	30.52	31.25	29.92	29.81	29.39
33387	27.50	29.16	28.94	28.44	28.50	27.94	28.77
39978	25.81	27.24	27.55	26.99	27.18	26.83	26.72

■

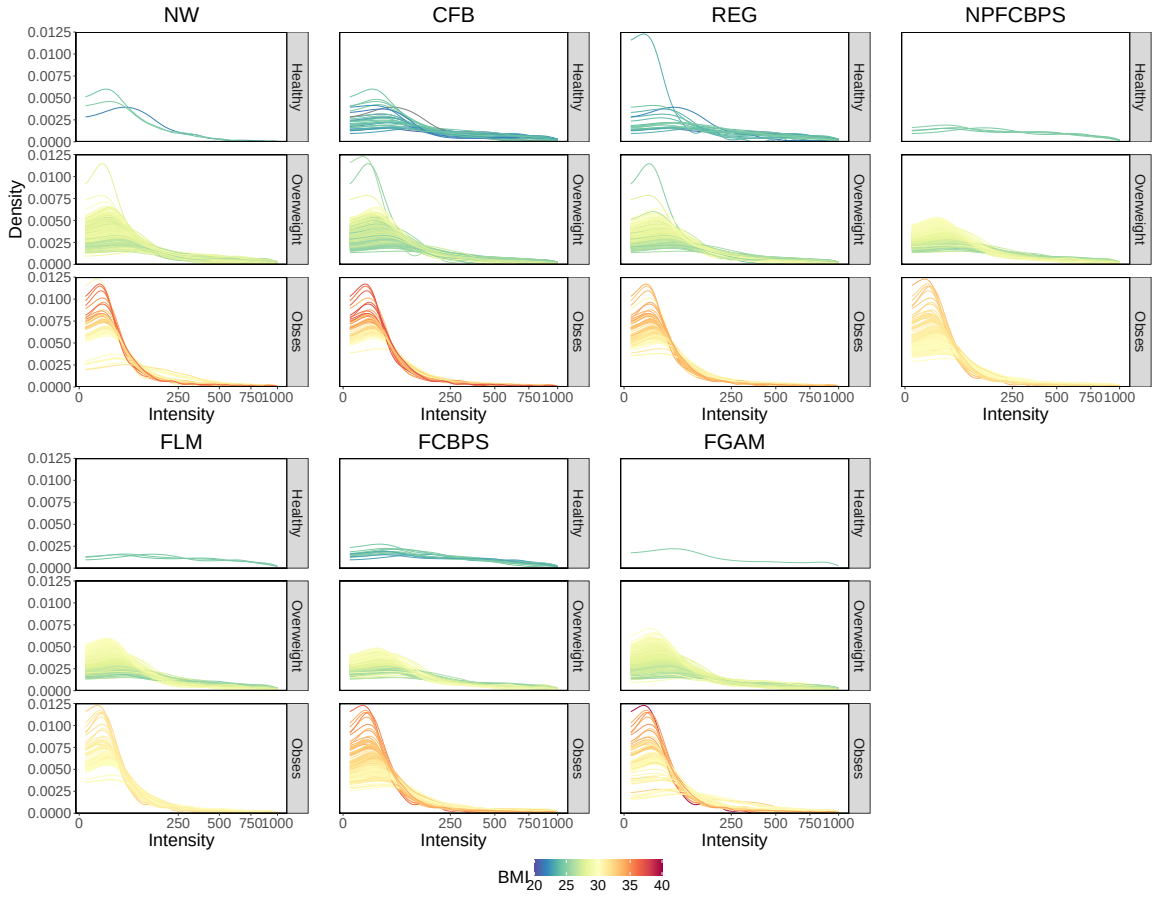


Figure 4: Density curves divided into different categories by their corresponding estimated BMI values using different estimators with cutoff value 75. See details in Figure 2.

Table 7: Estimated BMI values for the three representative density curves with cutoff value 75. See details in Table 3.

SEQN	CFB	FCBPS	NPFCBPS	REG	FLM	FGAM	NW
32104	30.12	31.49	30.56	31.27	29.94	29.84	29.39
33387	27.44	29.16	28.94	28.44	28.51	27.94	28.78
39978	25.75	27.24	27.53	26.99	27.18	26.83	26.71

References

Taha Bahadori, Eric Tchetgen Tchetgen, and David Heckerman. End-to-end balancing for causal continuous treatment-effect estimation. In *International Conference on Machine Learning*, pages 1313–1326. PMLR, 2022.

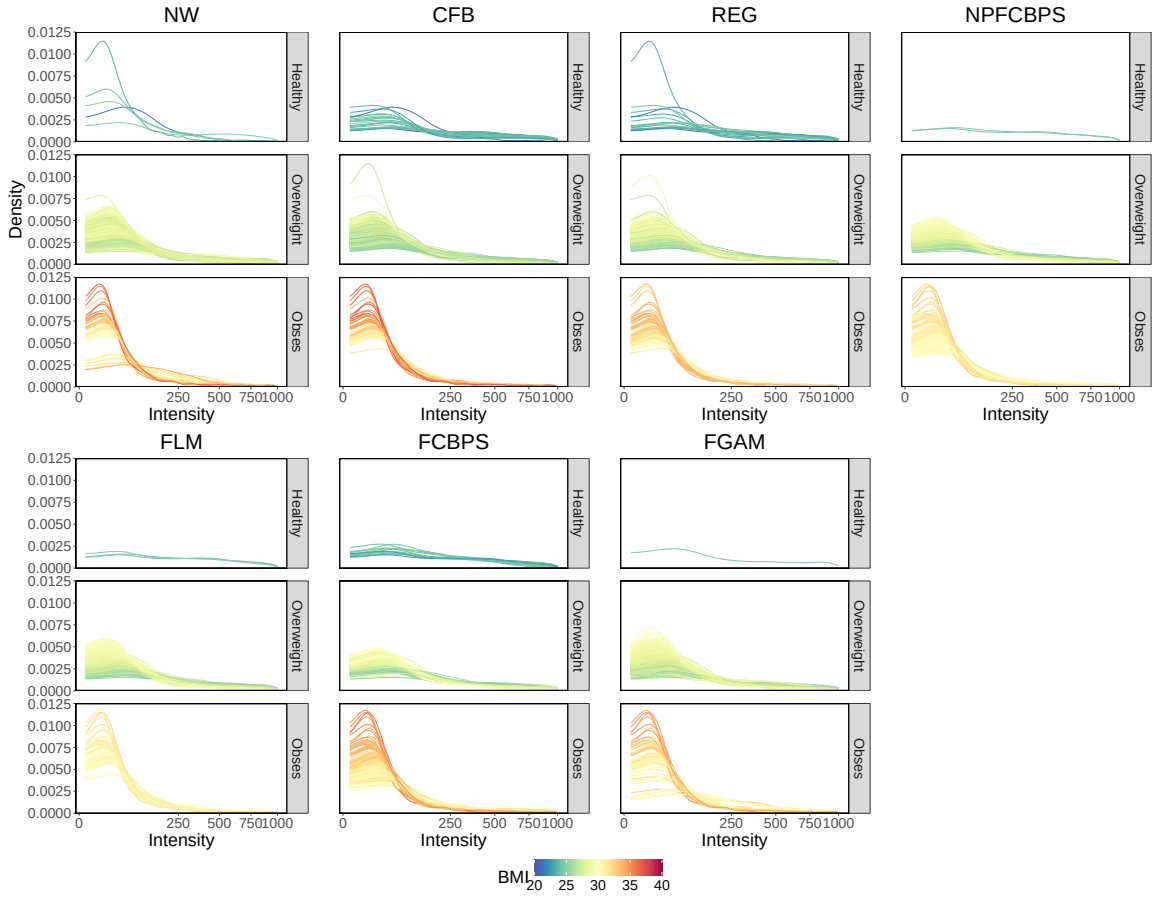


Figure 5: Density curves divided into different categories by their corresponding estimated BMI values using different estimators with cutoff value 200. See details in Figure 2.

Table 8: Estimated BMI values for the three representative density curves with cutoff value 200. See details in Table 3.

SEQN	CFB	FCBPS	NPFCBPS	REG	FLM	FGAM	NW
32104	30.55	31.79	30.49	31.43	29.93	29.92	29.42
33387	27.11	29.26	28.92	28.42	28.48	27.98	28.77
39978	26.02	27.19	27.53	26.97	27.12	26.73	26.57

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

Hsin-wen Chang and Ian W McKeague. Empirical likelihood-based inference for functional means with application to wearable device data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1947–1968, 2022.

Adam Ciarleglio, Eva Petkova, Todd Ogden, and Thaddeus Tarpey. Constructing treatment decision rules based on scalar and functional predictors when moderators of treatment effect are unknown. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1331–1356, 2018.

Gerda Claeskens, Tatyana Krivobokova, and Jean D Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544, 2009.

Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, 2010.

Peng Ding and Fan Li. Causal inference: a missing data perspective. *Statistical Science*, 33 (2):214–237, 2018.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121, 1996.

Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7):681–697, 2012.

Ezra I Fishman, Jeremy A Steeves, Vadim Zipunnikov, Annemarie Koster, David Berrigan, Tamara A Harris, and Rachel Murphy. Association between objectively measured physical activity and mortality in nhanes. *Medicine and science in sports and exercise*, 48(7):1303, 2016.

Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

Eszter Fuezeki, Tobias Engeroff, and Winfried Banzer. Health benefits of light-intensity physical activity: a systematic review of accelerometer data of the national health and nutrition examination survey (nhanes). *Sports medicine*, 47:1769–1793, 2017.

Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. *Advances in Neural Information Processing Systems*, 22, 2009.

Antonio F Galvao and Liang Wang. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 110(512):1528–1542, 2015.

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

Chong Gu and Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer, 2013.

Jens Hainmueller. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, pages 25–46, 2012.

Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 226164:73–84, 2004.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 243–263, 2014.

Kosuke Imai and David A van Dyk. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association*, 99 (467):854–866, 2004.

Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

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

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, and Manuel Davy. Nonlinear functional regression: a functional rkhs approach. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 374–380. JMLR Workshop and Conference Proceedings, 2010.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016.

Nathan Kallus and Michele Santacatterina. Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments. *arXiv preprint arXiv:1910.11972*, 2019.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

Eric B Laber and Ana-Maria Staicu. Functional feature construction for individualized treatment regimes. *Journal of the American Statistical Association*, 113(523):1219–1227, 2018.

Fan Li. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389–2415, 2019.

Yunzhe Li, Kun Kuang, Bo Li, Peng Cui, Jianrong Tao, Hongxia Yang, and Fei Wu. Continuous treatment effect estimation via generative adversarial de-confounding. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, pages 4–22. PMLR, 2020.

Zhenhua Lin, Dehan Kong, and Linbo Wang. Causal inference on distribution functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):378–398, 2023.

Michael J Lopez and Roee Gutman. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454, 2017.

Carol A Maher, Emily Mire, Deirdre M Harrington, Amanda E Staiano, and Peter T Katzmarzyk. The independent and combined associations of physical activity and sedentary behavior with obesity in adults: Nhanes 2003-06. *Obesity*, 21(12):E730–E737, 2013.

Ian W McKeague and Min Qian. Estimation of treatment policies based on functional predictors. *Statistica Sinica*, 24(3):1461, 2014.

Mathew W McLean, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert. Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269, 2014.

Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer, 2002.

Rui Miao, Wu Xue, and Xiaoke Zhang. Average treatment effect estimation in observational studies with functional covariates. *Statistics and Its Interface*, 15(2):237–246, 2022.

Rui Miao, Xiaoke Zhang, and Raymond KW Wong. A wavelet-based independence test for functional data with an application to meg functional connectivity. *Journal of the American Statistical Association*, 118(543):1876–1889, 2023.

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.

M Neovius, Y Linne, and S Rossner. BMI, waist-circumference and waist-hip-ratio as diagnostic tests for fatness in adolescents. *International Journal of Obesity*, 29(2):163–169, 2005.

Junier Oliva, William Neiswanger, Barnabás Póczos, Eric Xing, Hy Trac, Shirley Ho, and Jeff Schneider. Fast function to function regression. In *Artificial Intelligence and Statistics*, pages 717–725. PMLR, 2015.

Paul A Parker and Scott H Holan. A bayesian functional data model for surveys collected under informative sampling with application to mortality estimation using nhanes. *Biometrics*, 79(2):1397–1408, 2023.

Jing Qin and Biao Zhang. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):101–122, 2007.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(2), 2012.

John Rice and Murray Rosenblatt. Smoothing splines: regression, derivatives and deconvolution. *The Annals of Statistics*, pages 141–156, 1983.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

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

Paul R Rosenbaum, PR Rosenbaum, and Briskman. *Design of Observational Studies*, volume 10. Springer, 2010.

Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.

Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pages 948–957. PMLR, 2015.

Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Ruoxu Tan, Wei Huang, Zheng Zhang, and Guosheng Yin. Causal effect of functional treatment. *arXiv preprint arXiv:2210.00242*, 2022.

Stefan Tübbicke. Entropy balancing for continuous treatments. *Journal of Econometric Methods*, 11(1):71–89, 2022.

Jiayi Wang, Raymond K. W. Wong, and Xiaoke Zhang. Low-rank covariance function estimation for multidimensional functional data. *Journal of the American Statistical Association*, pages 1–14, 2020.

Jiayi Wang, Raymond KW Wong, Shu Yang, and Kwun Chuen Gary Chan. Estimation of partially conditional average treatment effect by double kernel-covariate balancing. *Electronic Journal of Statistics*, 16(2):4332–4378, 2022.

Yixin Wang and Jose R Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020.

Raymond K. W. Wong and Kwun Chuen Gary Chan. Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213, 2018.

Raymond K. W. Wong, Yehua Li, and Zhengyuan Zhu. Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114(525):406–418, 2019.

Shu Yang, Guido W Imbens, Zhanglin Cui, Douglas E Faries, and Zbigniew Kadziola. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065, 2016.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5): 1–46, 2021.

Haizhang Zhang, Yuesheng Xu, and Qinghui Zhang. Refinement of operator-valued reproducing kernels. *Journal of Machine Learning Research*, 13(4):91–136, 2012.

Jin-Ting Zhang and Jianwei Chen. Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079, 2007.

Xiaoke Zhang, Wu Xue, and Qiyue Wang. Covariate balancing functional propensity score for functional treatments in cross-sectional observational studies. *Computational Statistics & Data Analysis*, 163:107303, 2021.

Yi Zhao, Xi Luo, Martin Lindquist, and Brian Caffo. Functional mediation analysis with an application to functional magnetic resonance imaging data. *arXiv preprint arXiv:1805.06923*, 2018.

Yeying Zhu, Donna L Coffman, and Debashis Ghosh. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3(1):25–40, 2015.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

José R Zubizarreta, Caroline E Reinke, Rachel R Kelz, Jeffrey H Silber, and Paul R Rosenbaum. Matching for several sparse nominal variables in a case-control study of readmission following surgery. *The American Statistician*, 65(4):229–238, 2011.