Electronic Journal of Statistics
Vol. 16 (2022) 4332–4378
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2000

Estimation of partially conditional average treatment effect by double kernel-covariate balancing*

Jiayi Wang and Raymond K. W. Wong

Department of Statistics, Texas A&M University, College Station, TX 77843, USA
e-mail: jiayiwang@tamu.edu; raywong@tamu.edu

Shu Yang

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
e-mail: syang24@ncsu.edu

Kwun Chuen Gary Chan

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
e-mail: kcgchan@u.washington.edu

Abstract: We study nonparametric estimation for the partially conditional average treatment effect, defined as the treatment effect function over an interested subset of confounders. We propose a double kernel weighting estimator where the weights aim to control the balancing error of any function of the confounders from a reproducing kernel Hilbert space after kernel smoothing over the interested subset of variables. In addition, we present an augmented version of our estimator which can incorporate the estimation of outcome mean functions. Based on the representer theorem, gradient-based algorithms can be applied for solving the corresponding infinite-dimensional optimization problem. Asymptotic properties are studied without any smoothness assumptions for the propensity score function or the need for data splitting, relaxing certain existing stringent assumptions. The numerical performance of the proposed estimator is demonstrated by a simulation study and an application to the effect of a mother's smoking on a baby's birth weight conditioned on the mother's age.

Keywords and phrases: Augmented weighting estimator, causal inference, fully and partially conditional average treatment effect, treatment effect heterogeneity.

Received November 2021.

*Wong's research is partially 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. Yang's research is partially supported by the National Institute on Aging (1R01AG066883) and the National Science Foundation (DMS-1811245). Chan's research is partially supported by the National Heart, Lung, and Blood Institute (R01HL122212) and the National Science Foundation (DMS-1711952).

Contents

1. Introduction

Causal inference often concerns not only the average effect of the treatment on the outcome but also the conditional average treatment effect (CATE) given a set of individual characteristics, when treatment effect heterogeneity is expected or of interest. Specifically, let $T \in \{0,1\}$ be the treatment assignment, 0 for control and 1 for active treatment, $X \in \mathcal{X} \subset \mathbb{R}^d$ a vector of all pre-treatment confounders, and Y the outcome of interest. Following the potential outcomes framework, let $Y(t)$ be the potential outcome, possibly contrary to fact, had the unit received treatment $t \in \{0,1\}$. Then, the individual treatment effect is $Y(1) - Y(0)$, and the (fully) CATE can be characterized through

$\gamma(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\}$, $x \in \mathcal{X}$. Due to the fundamental problem in causal inference that the potential outcomes are not jointly observable, identification and estimation of the CATE in observational studies require further assumptions. A common assumption is the no unmeasured confounding (UNC) assumption, requiring X to capture all confounding variables that affect the treatment assignment and outcome. This often results in a multidimensional X. Given the UNC assumption, many methods have been proposed to estimate $\gamma(x)$ [30, 39, 24]. However, in clinical settings, researchers may only concern the variation of treatment effect over the change of a small subset of covariates $V \in \mathcal{V} \subseteq \mathcal{X}$, not necessarily the full set X. For example, researchers are interested in estimating the CATE of smoking (treatment) on birth weight (outcome) given mother's age, which is a function of a one-dimensional variable: age. While the target is a one-dimensional function, we still need to adjust for all confounders in addition to mother's age, such as mother's education attainment and numbers of prenatal care visits, to obtain a reasonable estimation. In this article, we focus on estimating $\tau(v) = \mathbb{E}\{\gamma(X) \mid V = v\}$ for $v \in \mathcal{V}$, which we refer to as the partially conditional average treatment effect (PCATE). When V is taken to be X, $\tau(v)$ becomes the fully conditional average treatment effect (FCATE) $\gamma(x)$. Despite our major focus on cases when $\mathcal{V}$ is a proper subset of $\mathcal{X}$, the proposed method in this paper does not exclude the setting with $\mathcal{V} = \mathcal{X}$, which results in the FCATE. However, existing results for FCATE may not be directly applicable to PCATE.

Without loss of generality, we focus on the setting with continuous V [3, 27, 13, 48, 11, 37] while the proposed method can be used to handle V that consists of continuous and discrete variables. When V contains discrete covariates, one can divide the whole sample into different strata by restricting the same values of discrete covariates of V in the same stratum. Then $\tau(v)$ can be obtained by estimating the PCATE over the remaining continuous covariates in V separately for every stratum. A typical estimation strategy involves two steps. The first step is to estimate nuisance parameters including the propensity score function and the outcome mean functions for the construction of adjusted responses (through weighting and augmentation) that are (asymptotically) unbiased for $\gamma(x)$ given $X = x$. The nuisance parameters can be estimated by parametric, nonparametric, and machine learning models. This step serves to adjust for confounding biases. In the second step, existing methods typically adopt nonparametric regression of the adjusted responses over V. However, these methods suffer from drawbacks. Firstly, all parametric methods are potentially sensitive to model misspecification especially when the CATE is complex. On the other hand, although nonparametric and machine learning methods are flexible, the first-step estimator of $\gamma(X)$ with high-dimensional X requires stringent assumptions for the possibly low-dimensional PCATE estimation to achieve the optimal convergence rate. For example, [3], [48], [13], [11] and [37] specify restrictive requirements for the convergence rate of the estimators of the nuisance parameters (see Remarks 3 and 6).

Instead of separating confounding adjustment and kernel smoothing in two steps, we propose a new framework that unifies the confounding adjustment and

kernel smoothing in one single weighting step. Two major contributions of our work are summarized as follows.

First, we generalize the idea of covariate balancing weighting in the average treatment effect (ATE) estimation literature [33, 16, 20, 49, 45] for estimating a scalar parameter to the PCATE estimation framework where the estimand is a function of v. This generalization, however, is non-trivial because we require covariate balancing in terms of flexible outcome models between the two treatment groups given all possible values of v. We assume that the outcome models lie in the reproducing kernel Hilbert space (RKHS, [40]). RKHS is a fairly general class of function space. Examples include many commonly seen spaces such as Sobolev space [e.g., 40, 15] and spline space [e.g., 32]. We then propose covariate function balancing (CFB) weights that are capable of controlling the balancing error with respect to the L_2-norm of any function with a bounded norm over the RKHS after kernel smoothing, see (3.4) and the detailed description in Section 3.2. The construction of the proposed weights specifically involves two kernels — the reproducing kernel of the RKHS and the kernel function used in kernel smoothing — and the goal of these weights can be understood as to balance covariate functions generated by these two kernels.

Second, asymptotic properties of the proposed estimator are derived under the complex dependency structure of weights and kernel smoothing without data splitting (see Remark 7). Our method *does not* require any smoothness assumptions on the propensity score model, in sharp contrast to existing methods, and only require mild smoothness assumptions for the outcome models to achieve (near) optimal convergence rate (see Sections 5.1 and 5.2). In addition, our proposed weighting estimator can be slightly modified to incorporate the estimation of the outcome mean functions, similar to the augmented inverse probability weighting (AIPW) estimator. We show that the augmentation of the outcome models relaxes the selection of tuning parameters theoretically.

The rest of the paper is organized as follows. Section 2 provides the basic setup for the CATE estimation. Section 3 introduces the proposed CFB weighting estimator, together with the computation method. Section 4 introduces an augmented version of the proposed estimator. The asymptotic properties of the proposed estimators are developed in Section 5. A simulation study and a real data application are presented in Sections 6 and 7, respectively. Additional computational and technical details are deferred to the Appendix.

2. Basic setup

Suppose $\{(T_i, Y_i(1), Y_i(0), X_i) : i = 1, \ldots, N\}$ are N independent and identically distributed copies of $\{T, Y(1), Y(0), X\}$. We assume that the observed outcome is $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ for $i = 1, \ldots, N$. Thus, the observed data $\{(T_i, Y_i, X_i) : i = 1, \ldots, N\}$ are also independent and identically distributed. For simplicity, we drop the subscript i when no confusion arises.

We focus on the setting satisfying treatment ignorability in observational studies [34].

Assumption 1 (No unmeasured confounding). $\{Y(1), Y(0)\} \perp\!\!\!\perp T \mid X$.

Assumption 1 rules out latent confounding between the treatment and outcome. In observational studies, its plausibility relies on whether or not the observed covariates X include all the confounders that affect the treatment as well as the outcome.

Most of the existing works [30, 39, 24, 37] focus on estimating the FCATE given the full set of X, i.e., $\gamma(x)$, $x \in \mathcal{X}$. However, to ensure Assumption 1 holds, X is often multidimensional, leading to a multidimensional CATE function $\gamma(x)$ that is challenging to estimate. Indeed, it is common that some covariates in X are simply confounders but not treatment effect modifiers of interest. Therefore, a more sensible way is to allow the conditioning variables to be an interested subset of confounders [3, 48, 13]. Instead of $\gamma(x)$, we focus on estimating the PCATE

$$\tau(v) = \mathbb{E}\{Y(1) - Y(0) \mid V = v\}, \quad v \in \mathcal{V} \subseteq \mathcal{X},$$

where V is a subset of X and is of dimension d_1 for $d_1 \leq d$. It is worth noting that $V = X$ is also allowed, and therefore $\gamma(x)$ can be estimated under our framework. For simplicity, we assume V is a continuous random vector for the rest of the paper. When V contains discrete random variables, one can divide the sample into different strata, of which the units have the same level of discrete covariates. Then $\tau(v)$ can be estimated by estimating the PCATE at every stratum.[1] For instance, suppose we are interested in estimating the PCATE of smoking on birth weights conditioned on mother's age and race. As race is a discrete variable, we could separate the sample into different strata based on races and estimate the PCATE conditioned on mother's age for each stratum. In the real data example (Section 7), we estimate the PCATE conditioned on mother's age for white and non-Hispanic mothers, where the sample size is 3754. Although it is not covered in Section 7, one could also estimate the PCATE conditioned on mother's age for other strata such as non-white mothers.

In addition to Assumption 1, we require sufficient overlap between the treatment groups. Let $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ be the propensity score. Throughout this paper, we also assume that the propensity score is strictly bounded above zero and below one to ensure overlap.

Assumption 2. *The propensity score $\pi(\cdot)$ is uniformly bounded away from zero and one. That is, there exist a constant $C_1 > 0$, such that $1/C_1 \leq \pi(x) \leq (1 - 1/C_1)$ for all $x \in \mathcal{X}$.*

Under Assumptions 1 and 2, $\tau(v)$ is identifiable based on the following formula

$$\tau(v) = \mathbb{E}\{Y(1) - Y(0) \mid V = v\} = \mathbb{E}\left\{ \frac{TY}{\pi(X)} - \frac{(1-T)Y}{1-\pi(X)} \;\middle|\; V = v \right\}.$$

First, suppose $\pi(X_i), i = 1, \ldots, N$, are known. Common procedures construct adjusted responses $Z_i = T_i Y_i / \pi(X_i) - (1 - T_i) Y_i / \{1 - \pi(X_i)\}$ and apply a kernel

[1] Kernel smoothing may fail to run if the sample size in some stratum is too small and compactly supported kernels are adopted. If such a numerical issue occurs, one can consider dropping or merging the stratum to its nearest neighbor to resolve the issue.

smoother to the data $\{(V_i, Z_i), i = 1, \ldots, N\}$ [e.g. 3]. Specifically, let $K(v)$ be a kernel function and $h > 0$ be a bandwidth parameter (with technical conditions specified in Section 5.1). The above strategy leads to the following estimator for $\tau(v)$:

$$\frac{1/(Nh^{d_1})\sum_{i=1}^{N} K\{(V_i - v)/h\} Z_i}{1/(Nh^{d_1})\sum_{j=1}^{N} K\{(V_j - v)/h\}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{K}_h(V_i, v) Z_i \tag{2.1}$$

where

$$\tilde{K}_h(v_1, v_2) = \frac{\frac{1}{h^{d_1}} K\{(v_1 - v_2)/h\}}{\frac{1}{N}\sum_{j=1}^{N} \frac{1}{h^{d_1}} K\{(V_j - v_2)/h\}}.$$

In observational studies, the propensity scores $\pi(X_i), i = 1, \ldots, N$, are often unknown. [3] proposes to estimate these scores using another kernel smoother, and construct the adjusted responses based on the estimated propensity scores. There are two drawbacks with this approach. First, it is well known that inverting the estimated propensity scores can result in instability, especially when some of the estimated propensity scores are close to zero or one [23]. Second, this procedure relies on the propensity score model to be correctly specified or sufficiently smooth to approximate well.

To overcome these issues, instead of obtaining the weights by inverting the estimated propensity scores, we focus on estimating the inverse propensity score weights directly. In the next section, we adopt the idea of covariate balancing weighting, which has been recently studied in the context of average treatment effect (ATE) estimation [e.g., 16, 20, 49, 9, 45, 46, 22, 43].

3. Covariate function balancing weighting for PCATE estimation

3.1. Motivation

To motivate the proposed estimator, suppose we are given two sets of the covariate balancing weights $\{\hat{w}_i : i = 1, \ldots, N\}$ and $\{\hat{w}'_i : i = 1, \ldots, N\}$ for the treated group and control group respectively. We express the adjusted response as

$$\hat{Z}_i = T_i \hat{w}_i Y_i - (1 - T_i)\hat{w}'_i Y_i, \quad i = 1, \ldots, N. \tag{3.1}$$

Without loss of generality, we can take $\hat{w}_i = 0$ if $T_i = 0$ and $\hat{w}'_i = 0$ if $T_i = 1$. Combining (2.1) and (3.1), the estimator of $\tau(v)$ is

$$\hat{\tau}(v) = \frac{1}{N}\sum_{i=1}^{N} T_i \hat{w}_i \tilde{K}_h(V_i, v) Y_i - \frac{1}{N}\sum_{i=1}^{N} (1 - T_i)\hat{w}'_i \tilde{K}_h(V_i, v) Y_i. \tag{3.2}$$

One can see that the estimator (3.2) is a difference between two terms, which are the estimates of $\mu_1(v) = \mathbb{E}\{Y(1) \mid V = v\}$ and $\mu_0(v) = \mathbb{E}\{Y(0) \mid V = v\}$, respectively. For simplicity, we focus on the first term and discuss the estimation

of the corresponding weights $\{w_i : T_i = 1\}$ in the treated group. The same procedure can be applied to estimate the second term, by swapping the values of indicators for the treated and controls. More specifically, $\{\hat{w}_i' : T_i = 0\}$ is the solution of (3.9) with T_i replaced by $(1 - T_i)$ for $i = 1, \ldots, N$.

We assume $Y_i(1) = m_1(X_i) + \varepsilon_i$ such that the ε_i's are independent random errors with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) \leq \sigma_0^2 < \infty$, and m_1 is considered as the outcome mean function for the treated group. Focusing on the first term of (3.2), we decompose $N^{-1}\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i, v)Y_i$ as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i, v)m_1(X_i) + \frac{1}{N}\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i, v)\varepsilon_i \\
&= \frac{1}{N}\sum_{i=1}^N (T_i\hat{w}_i - 1)\tilde{K}_h(V_i, v)m_1(X_i) + \frac{1}{N}\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i, v)\varepsilon_i \\
&\quad + \left\{\frac{1}{N}\sum_{i=1}^N \tilde{K}_h(V_i, v)m_1(X_i) - \mu_1(v)\right\} + \mu_1(v).
\end{aligned} \tag{3.3}$$

In the last equality, only the first two terms depend on the weights. The third term in the decomposition corresponds to the estimation error of a typical local constant regression (Nadaraya-Watson regression) and is well-studied in the literature [e.g. 44]. As ϵ_i's are mean-zero random variables that are independent of X_i's and T_i's., the second term $N^{-1}\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i, v)\varepsilon_i$ will be handled by controlling the variability of the weights (See the proof in Section E.2 for details). The primary challenge lies in controlling the first term, which requires the control of the (empirical) balance of a kernel-weighted function class because $m_1(X_i), i = 1, \ldots, N$, are unknown. This requirement makes achieving covariate balance significantly more challenging than those for estimating the ATE, *i.e.*, when V is deterministic [e.g., 16, 20, 49, 9, 45, 46, 22, 43], for multiple reasons: (i) covariate balance is required for all v in a continuum, and (ii) the bandwidth h in kernel smoothing is required to diminish with the sample size N.

3.2. Balancing via an empirical residual moment operator

Suppose $m_1 \in \mathcal{H}$, where $\mathcal{H}$ is an RKHS with reproducing kernel κ and norm $\|\cdot\|_{\mathcal{H}}$. Also, let the squared empirical norm be $\|u\|_N^2 = (1/N)\sum_{i=1}^N \{u(X_i)\}^2$ for any $u \in \mathcal{H}$. Our goal is to make the proposed estimator as close to μ_1 as possible. From the decomposition (3.3), we can achieve this goal by controlling the first three terms (i.e., excluding $\mu_1(v)$) in the decomposition. Intuitively, from the first term of (3.3), we aim to find weights $w = \{w_i : T_i = 1\}$ to ensure the following function balancing criteria:

$$\frac{1}{N}\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i, v)u(X_i) \approx \frac{1}{N}\sum_{i=1}^N u(X_i)\tilde{K}_h(V_i, v),$$

for all $u \in \mathcal{H}$, where the left and right hand sides are regarded as functions of v. To quantify such an approximation, we define the operator $\mathcal{M}_{N,h,w}$ mapping an element of $\mathcal{H}$ to a function on $\mathcal{V}$ by

$$\mathcal{M}_{N,h,w}(u,\cdot) = \frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)u(X_i)\tilde{K}_h(V_i,\cdot),$$

which we call the empirical residual moment operator with respect to the weights in w.

The approximation and hence the balancing error can be measured by

$$\|\mathcal{M}_{N,h,w}(u,\cdot)\|^2,$$

where $\|f\|$ is a generic metric applied to a function f defined on $\mathcal{V}$. Typical examples of a metric are L_∞-norm ($\|\cdot\|_\infty$), L_2-norm ($\|\cdot\|_2$) and empirical norm ($\|\cdot\|_N$). If one has non-uniform preference over $\mathcal{V}$, weighted L_2-norm and weighted empirical norm are also applicable. In the following, we focus on the balancing error based on L_2-norm:

$$S_{N,h}(w,u) = \|\mathcal{M}_{N,h,w}(u,\cdot)\|_2^2. \tag{3.4}$$

We will return to the discussion of other norms in Section 5. Ideally, our target is to minimize $\sup_{u\in\mathcal{H}} S_{N,h}(w,u)$ uniformly over a sufficiently complex space $\mathcal{H}$. As soon as one attempts to do this, one may find that $S_{N,h}(w,tu) = t^2 S_{N,h}(w,u)$ for any $t \geq 0$, which indicates a scaling issue about u. Therefore, we will standardize the magnitude of u and restrict the space to $\mathcal{H}_N = \{u \in \mathcal{H} : \|u\|_N^2 = 1\}$ as in [45]. Also, to overcome overfitting, we add a penalty on u in terms of $\|\cdot\|_{\mathcal{H}}$ that regularizes the complexity of u and focus on controlling the balancing error over smoother functions. An alternative strategy in the ATE estimation literature imposes a constraint $\mathcal{H}(1) = \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} = 1\}$, which, however, restricts to a pre-fixed function class. Our strategy allows data-driven tuning based on $\lambda_1\|\cdot\|_{\mathcal{H}}^2$ in (3.6) to adapt to a relevant function class. Inspired by the discussion for (3.3), we also introduce another penalty term

$$R_{N,h}(w) = \frac{1}{N}\sum_{i=1}^{N}\|T_i w_i \tilde{K}_h(V_i,\cdot)\|_2^2, \tag{3.5}$$

to control the variability of the weights. From the decomposition (3.3), we expect a careful control on $S_{N,h}(w,m_1)$ and $R_{N,h}(w)$ would lead to a bound on $\|\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i,\cdot)Y_i/N - \mu_1\|_2$ (See Section 5.2).

In summary, given any $h > 0$, our CFB weights $\hat{w}$ is constructed as follows:

$$\hat{w} = \operatorname*{argmin}_{w}\left[\sup_{u\in\mathcal{H}_N}\left\{S_{N,h}(w,u) - \lambda_1\|u\|_{\mathcal{H}}^2\right\} + \lambda_2 R_{N,h}(w)\right], \tag{3.6}$$

where λ_1 and λ_2 are tuning parameters ($\lambda_1 > 0$ and $\lambda_2 > 0$). Note that (3.6) does not depend on the weights $\{w_i : T_i = 0\}$ of the control group, and the optimization is only performed with respect to $\{w_i : T_i = 1\}$.

Remark 1. By standard representer theorem, we can show that the solution $\tilde{u} = \hat{u}/\|\hat{u}\|_N$ of the inner optimization satisfies that $\hat{u}$ belongs to $\mathcal{K}_N = \text{span}\{\kappa(X_i, \cdot) : i = 1, \ldots, N\}$ (See Section B.1 in the Appendix). Therefore, by the definition of $\mathcal{M}_{N,h,w}$, the weights are determined by achieving the balance of the covariate functions generated by two kernels: the reproducing kernel κ and the smoothing kernel K.

Remark 2. [45] adopts a similar optimization form as in (3.6) to obtain weights. The key difference between their estimator and ours is the choice of balancing error tailored to the target quantity. In [45], the choice of balancing error is $\{\sum_{i=1}^N (T_i w_i - 1)u(X_i)/N\}^2$, which is designed for estimating the *scalar* ATE. There is no guarantee that the resulting weights will ensure enough balance for the estimation of the PCATE, *a function of* v. Heuristically, one can regard the balancing error in [45] as the limit of $S_{N,h}$ as $h \to \infty$. For finite h, two fundamental difficulties emerge that do not exist in [45]. First, $\mathcal{M}_{N,h,w}(u, v)$ changes with v, and so the choice of $S_{N,h}$ involves a metric for a function of v in (3.4). This is directly related to the fact that our target is a function (PCATE) instead of a scalar (ATE). For reasonable metrics, the resulting balancing errors measure imbalances over all (possibly infinite) values of v, which is significantly more difficult than the imbalance control required for ATE. Second, for each v, the involvement of kernel function in $\mathcal{M}_{N,h,w}$ suggests that the effective sample size used in the corresponding balancing is much smaller than $\sum_{i=1}^N T_i$. There is no theoretical guarantee for the weights of [45] to ensure enough balance required for the PCATE, since the proposed weights are designed to balance a function instead of a scalar. We show that the proposed CFB weighting estimator achieves desirable properties both theoretically (Section 5) and empirically (Section 6).

3.3. Computation

In this section, we discuss the computation of the CFB weights and defer details and proof to the Appendix. For simplicity of exposition, we introduce more notations: $\circ$ is the element-wise product of two vectors, $J = (1, 1, \ldots, 1)^\intercal$, $\Omega(A)$ represents the maximum eigenvalue of a symmetric matrix A, $P \in \mathbb{R}^{N \times r}$ consists of the singular vectors of the Gram matrix $M := [\kappa(X_i, X_j)]_{i,j=1}^N \in \mathbb{R}^{N \times N}$ of rank r, $D \in \mathbb{R}^{r \times r}$ is the diagonal matrix such that

$$M = PDP^\intercal, \tag{3.7}$$

and

$$G_h = \begin{bmatrix} \int_{\mathcal{V}} \tilde{K}_h(V_1, v)\tilde{K}_h(V_1, v)\mathrm{d}v & \cdots & \int_{\mathcal{V}} \tilde{K}_h(V_1, v)\tilde{K}_h(V_N, v)\mathrm{d}v \\ \vdots & \ddots & \vdots \\ \int_{\mathcal{V}} \tilde{K}_h(V_N, v)\tilde{K}_h(V_1, v)\mathrm{d}v & \cdots & \int_{\mathcal{V}} \tilde{K}_h(V_N, v)\tilde{K}_h(V_N, v)\mathrm{d}v \end{bmatrix} \in \mathbb{R}^{N \times N}. \tag{3.8}$$

Applying the standard representer theory, (3.6) can be reformulated as

$$\hat{w} = \underset{w \geq 1}{\operatorname{argmin}} \left[\Omega \left\{ \frac{1}{N} P^{\intercal} \operatorname{diag}(T \circ w - J) G_h \operatorname{diag}(T \circ w - J) P - N\lambda_1 D^{-1} \right\} + \lambda_2 R_{N,h}(w) \right]. \tag{3.9}$$

The underlying optimization is convex as shown in Lemma 1.

Lemma 1. *The optimization* (3.6) *or equivalently* (3.9) *is convex.*

By Lemma 1, generic convex optimization algorithms are applicable. Also, because the corresponding gradient has a closed-form expression[2], gradient-based algorithms, such as the L-BFGS-B algorithm, can be applied efficiently to solve the optimization. Regarding the selection of the tuning parameters λ_1 and λ_2, we adopt a criterion similar to [45]. Based on our theoretical study in Section 5, we will see that the optimal order of λ_2 is $\lambda_2 \asymp \lambda_1 h^{d_1}$ (see Theorems 5.2 and 5.4.) For simplicity, we choose $\lambda_2 = \rho_1 h^{d_1} \lambda_1$, where $\rho_1 > 0$ is a fixed parameter. Notice that $\lambda_2 > 0$ is mainly imposed to control $R_{N,h}(w)$. From our experience, $R_{N,h}$ is usually stable and does not take a large value even if λ_2 is small. Therefore we are inclined to take a small λ_2. We fix $\rho_1 = 0.01$ in all our numerical applications. Next, we discuss how to tune λ_1. Roughly speaking, as λ_1 increases, $B_{N,h}(\hat{w}) := S_{N,h}(\hat{w}, u^{\lambda_1})$, where $u^{\lambda_1} = \operatorname{argmax}_{u \in \mathcal{H}_N} \{S_{N,h}(\hat{w}, u) - \lambda_1 \|u\|^2_{\mathcal{H}}\}$, decreases and approaches zero. This is because the smoother the function is, the easier it is to be balanced. The main idea is to select the smallest λ_1 such that $B_{N,h}(\hat{w})$ will not decrease much if we further enlarge λ_1. In practice, we compute the proposed weights with respect to a grid of λ_1 such that $\lambda_1^{(1)} < \cdots < \lambda_1^{(K)}$. Write $\hat{w}^{(k)}$ as the proposed weights with respect to $\lambda_1^{(k)}$. We select $\lambda_1^{(k^*)}$ as our choice if k^* is the smallest k such that $\{B_{N,h}(\hat{w}^{(k+1)}) - B_{N,h}(\hat{w}^{(k)})\}/(\lambda_1^{(k+1)} - \lambda_1^{(k)}) \geq \rho_2$, where ρ_2 is chosen as a negative constant of small magnitude. We set $\rho_2 = -10^{-6}$ in all numerical applications. Algorithm 1 outlines the optimization steps, together with the tuning parameter selection.

In the following, we discuss several practical strategies to speed up the optimization. First, Line 1 in Algorithm 1 computes G_h. Although the form of G_h may seem complicated, this does not change with w. Therefore, for each h, we can pre-compute G_h once at the beginning of an algorithm for the optimization (3.9). However, when the integral $g_h(v_1, v_2) = \int_{\mathcal{V}} \tilde{K}_h(v_1, v) \tilde{K}_h(v_2, v) dv$ does not possess a known expression, one generally has to perform a large number of numerical integration for the computation of G_h, when N is large. But, for smooth choices of K, g_h is also a smooth function. When N is large, we could evaluate $g_h(V_i, V_j)$, $i \in S_1, j \in S_2$ at smaller subsets S_1 and S_2. Then typical interpolation methods [17] can be implemented to approximate unevaluated integrals in G_h to ease the computation burden.

Second, Line 2 in Algorithm 1 computes the dominant eigen-pair of an $r \times r$ matrix to obtain the gradient and objective value. Since common choices of the reproducing kernel κ are smooth, the corresponding Gram matrix M can

[2] when the maximum eigenvalue in the objective function is of multiplicity 1

Algorithm 1: Optimization steps for solving (3.9) with selection of λ_1 and λ_2

Input: the data set$\{(X_i, V_i, T_i) :, i = 1, \ldots, n\}$; the bandwidth h, the grid for λ_1: $\{\lambda_1^{(k)} : k = 1, \ldots, K\}$, ρ_1, ρ_2

Calculate G_h according to (3.8).

Calculate the Gram matrix M and compute the eigen-decomposition to obtain P and D.

for $k = 0, 1, \ldots, K$ **do**

  Optimize (3.9) by L-BFGS-B algorithm with $\lambda_1 = \lambda_1^{(k)}$ and $\lambda_2 = \rho_1 \lambda_1^{(k)} h^{d_1}$ to obtain the solution $\hat{w}^{(k)}$.

  Calculate the balancing error

$$B_{N,h}(\hat{w}^{(k)}) = N^{-1} \zeta^\intercal P^\intercal \mathrm{diag}(T \circ \hat{w}^{(k)} - J) G_h \mathrm{diag}(T \circ \hat{w}^{(k)} - J) P \zeta,$$

  where ζ is the eigenvector that corresponds to the largest eigenvalue of $\{\frac{1}{N} P^\intercal \mathrm{diag}(T \circ \hat{w}^{(k)} - J) G_h \mathrm{diag}(T \circ \hat{w}^{(k)} - J) P - N \lambda_1^{(k)} D^{-1}\}$.

end

Select k^* as the smallest k such that $\{B_{N,h}(\hat{w}^{(k+1)}) - B_{N,h}(\hat{w}^{(k)})\}/(\lambda_1^{(k+1)} - \lambda_1^{(k)}) \geq \rho_2$

Output: $w^{(k^*)}$

usually be approximated well by a low-rank matrix. When N is large, to facilitate computation, one can choose P and D with a smaller dimension r such that (3.7) holds approximately. Due to the smaller r, this would significantly reduce the burden of computing the dominant eigen-pair of the $r \times r$ matrix.

4. Augmented estimator

Inspired by the augmented inverse propensity weighting (AIPW) estimators in the ATE literature [10, 5], we also propose an augmented estimator that directly adjusts for the outcome models $m_1(\cdot)$ and $m_0(\cdot)$. Augmented estimators combine estimations of weights (propensity scores) and outcome mean functions. And they have shown to be effective in the literature and practice. For the proposed augmented estimator, we also observe similar empirical benefits, due to leveraging both the weights and the outcome regression. In our theoretical study, we show that both the augmented and non-augmented estimators achieve the optimal convergence rate. Interestingly, the augmented estimator relaxes the requirement of the order of tuning parameters.

Recall that the outcome regression functions $m_1(\cdot)$ and $m_0(\cdot)$ are assumed to be in an RKHS $\mathcal{H}$, kernel-based estimators $\hat{m}_1(\cdot)$ and $\hat{m}_0(\cdot)$ can be employed. We then perform augmentation and obtain the adjusted response in (3.1) as

$$\hat{Z}_i = \hat{w}_i T_i \{Y_i - \hat{m}_1(X_i)\} + \hat{m}_1(X_i) - [\hat{w}_i'(1 - T_i)\{Y_i - \hat{m}_0(X_i)\} + \hat{m}_0(X_i)].$$

Correspondingly, the decomposition in (3.3) becomes

$$\frac{1}{N} \sum_{i=1}^{N} \tilde{K}_h(V_i, v) \hat{m}_1(X_i) + \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i \tilde{K}_h(V_i, v) \{Y_i - \hat{m}_1(X_i)\}$$

$$
\begin{aligned}
=&\frac{1}{N}\sum_{i=1}^{N}(1-T_i\hat{w}_i)\tilde{K}_h(V_i,v)\hat{m}_1(X_i)+\frac{1}{N}\sum_{i=1}^{N}T_i\hat{w}_i\tilde{K}_h(V_i,v)m_1(X_i)\\
&+\frac{1}{N}\sum_{i=1}^{N}T_i\hat{w}_i\tilde{K}_h(V_i,v)\epsilon_i\\
=&\frac{1}{N}\sum_{i=1}^{N}(T_i\hat{w}_i-1)\tilde{K}_h(V_i,v)\{m_1(X_i)-\hat{m}_1(X_i)\}+\frac{1}{N}\sum_{i=1}^{N}T_i\hat{w}_i\tilde{K}_h(V_i,v)\epsilon_i\\
&+\left\{\frac{1}{N}\sum_{i=1}^{N}\tilde{K}_h(V_i,v)m_1(X_i)-\mu_1(v)\right\}+\mu_1(v).
\end{aligned}
$$

Now, our goal is to control the difference between $N^{-1}\sum_{i=1}^{N}T_i\hat{w}_i\tilde{K}_h(V_i,v)\times\{m_1(X_i)-\hat{m}_1(X_i)\}$ and $N^{-1}\sum_{i=1}^{N}\tilde{K}_h(V_i,v)\{m_1(X_i)-\hat{m}_1(X_i)\}$. The weight estimators in Section 3.2 can be adopted similarly to control this difference. It can be shown that the term $S_{N,h}(\hat{w},m_1-\hat{m}_1)=\|N^{-1}\sum_{i=1}^{N}(T_i\hat{w}_i-1)\tilde{K}_h(V_i,\cdot)\times\{m_1(X_i)-\hat{m}_1(X_i)\}\|_2^2$ can achieve a faster rate of convergence than $S_{N,h}(\hat{w},m_1)$ does with the same estimated weights $\hat{w}$ as long as $\hat{m}_1$ is a consistent estimator. However, this property does not improve the final convergence rate of the PCATE estimation. This is because the term $\|N^{-1}\sum_{i=1}^{N}\tilde{K}_h(V_i,\cdot)m_1(X_i)-\mu_1\|_2^2$ dominates other terms, and thus the final rate can never be faster than the optimal non-parametric rate. See Remark 4 for more details. Our theoretical results reveal that the benefit of using the augmentations lies in the relaxed order requirement of the tuning parameters to achieve the optimal convergence rate. Therefore, the performance of the augmented estimator is expected to be more robust in the tuning parameter selection.

Unlike other AIPW-type estimators [27, 13, 48, 37] which often rely on data splitting for estimating the propensity score and outcome mean functions to relax technical conditions, our estimator does not require data splitting to facilitate the convergence with augmentation (see Remark 7). We defer the theoretical comparison between our estimator and the existing AIPW-type estimator (see Remark 6).

Lastly, we note that there are existing work using weights to balance the residuals [e.g. 5, 45], which appears similar to the proposed augmented estimator. These estimators are designed for ATE estimation and the balancing weights cannot be directly adopted here for PCATE estimation with theoretical guarantee.

5. Asymptotic properties

In this section, we conduct an asymptotic analysis for the proposed estimator. For simplicity, we assume $\mathcal{X}=[0,1]^d$. To facilitate our theoretical discussion in terms of smoothness, we assume the RKHS $\mathcal{H}$ is contained in a Sobolev space (see Assumption 3). Our results can be extended to other choices of $\mathcal{H}$ if the corresponding entropy result and boundedness condition for the unit ball

$\{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \leq 1\}$ are provided. Recall that we focus on $\mathbb{E}\{Y(1) \mid V = v\}$. Similar analysis can be applied to $\mathbb{E}\{Y(0) \mid V = v\}$ and finally the PCATE.

5.1. Regularity conditions

Let ℓ be a positive integer. For any function u defined on $\mathcal{X}$, the the Sobolev norm is $\|u\|_{\mathcal{W}^\ell} = \sqrt{\sum_{|\beta| \leq \ell} \|D^\beta u\|_2^2}$, where $D^\beta u(x_1, \ldots, x_d) = \frac{\partial^{|\beta|} u}{\partial x_1^{\beta_1} \ldots \partial x_d^{\beta_d}}$ for a multi-index $\beta = (\beta_1, \ldots, \beta_d)$. The Sobolev space $\mathcal{W}^\ell$ consists of functions with finite Sobolev norm. For $\epsilon > 0$, we denote by $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$ the ϵ-covering number of a set $\mathcal{F}$ with respect to some norm $\|\cdot\|$. Next, we list the assumptions that are useful for our asymptotic results.

Assumption 3. *The unit ball of $\mathcal{H}$ is a subset of a ball in the Sobolev space $\mathcal{W}^\ell$, with the ratio $\alpha := d/\ell$ less than 2.*

Assumption 4. *The regression function m_1 belongs to an RKHS $\mathcal{H}$.*

Assumption 5. ***(a)*** *K is symmetric, $\int K(s)ds = 1$, and there exists a constant C_2 such that $K(s) \leq C_2$ for all s. Moreover, $\int s^2 K(s)ds < \infty$ and $\int K^2(s)ds < \infty$.* ***(b)*** *Take $\mathcal{K} = \{K\{(v - \cdot)/h\} : h > 0, v \in [0,1]^{d_1}\}$. There exist constants $A_1 > 0$ and $\nu_1 > 0$ such that $\mathcal{N}(\varepsilon, \mathcal{K}, \|\cdot\|_\infty) \leq A_1 \varepsilon^{-\nu_1}$.*

Assumption 6. *The density function $g(\cdot)$ of the random variable $V \in [0,1]^{d_1}$ is continuous, differentiable, and bounded away from zero,* i.e.*, there exist constants $C_3 > 0$ and $C_4 > 0$ such that $C_3 \leq g(v) \leq C_4$.*

Assumption 7. *$h \to 0$ and $N^{\frac{2}{2+\alpha}} h^{d_1} \to \infty$, as $N \to \infty$.*

Assumption 8. *The joint density of $\{m_1(X), V\}$ and the conditional expectation $\mathbb{E}\{m_1(X) \mid V = v\}$ are continuous.*

Assumption 9. *The errors $\{\varepsilon_i, i = 1, \ldots, N\}$ are uncorrelated, with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) \leq \sigma_0^2$ for all $i = 1, \ldots, N$. Furthermore, $\{\varepsilon_i, i = 1, \ldots, N\}$ are independent of $\{T_i, i = 1, \ldots, N\}$ and $\{X_i, i = 1, \ldots, N\}$.*

Assumption 3 is a common condition in the literature of smoothing spline regression. Assumptions 5–8 comprise standard conditions for kernel smoother [e.g., 29, 12, 44] except that we require $N^{\frac{\alpha}{2+\alpha}} h^{d_1} \to \infty$ instead of $Nh^{d_1} \to \infty$ to ensure the difference between $\|u\|_N$ and $\|u\|_2$ is asymptotically negligible. Assumption 5(b) is satisfied whenever $K(\cdot) = \psi\{p(\cdot)\}$ with $p(\cdot)$ being a polynomial in d_1 variables and ψ being a real-valued function of bounded variation [38].

5.2. L_2-norm balancing

Given two sequences of positive real numbers $(A_1, A_2, \ldots)$ and $(B_1, B_2, \ldots)$, $A_N = \mathcal{O}(B_N)$ represents that there exists a positive constant M such that $A_N \leq MB_N$ as $N \to \infty$; $A_N = o(B_N)$ represents that $A_N/B_N \to 0$ as $N \to \infty$, and $A_N \asymp B_N$ represents $A_N = \mathcal{O}(B_N)$ and $B_N = \mathcal{O}(A_N)$.

Theorem 5.1. *Suppose Assumptions 1–7 hold. If* $\lambda_1^{-1} = o(Nh^{d_1})$*, we have* $S_{N,h}(\hat{w}, m) = \mathcal{O}_{\mathrm{p}}(\lambda_1 \|m\|_N^2 + \lambda_1 \|m\|_{\mathcal{H}}^2 + \lambda_2 h^{-d_1} \|m\|_N^2)$*. If we further assume* $\lambda_2^{-1} = \mathcal{O}(\lambda_1^{-1} h^{-d_1})$*, then* $R_{N,h}(\hat{w}) = \mathcal{O}_{\mathrm{p}}(h^{-d_1})$*.*

Theorem 5.1 specifies the control of the balancing error and the weight variability. From (3.3), we can bound $\| \sum_{i=1}^N T_i \hat{w}_i Y_i K_h(V_i, \cdot)/N - \mu_1 \|_2$ through the following decomposition

$$\begin{aligned}
\left\| \frac{1}{N} \sum_{i=1}^N T_i \hat{w}_i Y_i K_h(V_i, \cdot) - \mu_1 \right\|_2 \leq & \{ S_{N,h}(\hat{w}, m_1) \}^{\frac{1}{2}} \\
& + \left\{ \mathbb{E} \left\| \frac{1}{N} \sum_{i=1}^N T_i \hat{w}_i \tilde{K}_h(V_i, \cdot) \varepsilon_i \right\|_2^2 \right\}^{\frac{1}{2}} \\
& + \mathcal{O}_{\mathrm{p}}\left(N^{-\frac{1}{2}} \right) + \left\| \frac{1}{N} \sum_{i=1}^N \tilde{K}_h(V_i, \cdot) m_1(X_i) - \mu_1 \right\|_2 \\
\leq & \{ S_{N,h}(\hat{w}, m_1) \}^{\frac{1}{2}} + \sigma_0 \left\{ \frac{R_{N,h}(w)}{N} \right\}^{\frac{1}{2}} \\
& + \mathcal{O}_{\mathrm{p}}\left(N^{-\frac{1}{2}} \right) + \left\| \frac{1}{N} \sum_{i=1}^N \tilde{K}_h(V_i, \cdot) m_1(X_i) - \mu_1 \right\|_2 .
\end{aligned}$$

Then the results of Theorem 5.1 can be used to derive the convergence rate of the proposed estimator as shown in the following theorem.

Theorem 5.2. *Suppose Assumptions 1-9 hold. If* $\lambda_1^{-1} = o(Nh^{d_1})$*,* $\lambda_2^{-1} = \mathcal{O}(\lambda_1^{-1} h^{-d_1})$*, and* $h^2 = o\{(N^{-1} h^{-d_1})^{1/2}\}$*, we have*

$$\begin{aligned}
& \left\| \frac{1}{N} \sum_{i=1}^N T_i \hat{w}_i Y_i K_h(V_i, \cdot) - \mu_1 \right\|_2 \\
& \qquad = \mathcal{O}_{\mathrm{p}}(N^{-1/2} h^{-d_1/2} + \lambda_1^{1/2} \|m_1\|_{\mathcal{H}} + \lambda_2^{1/2} h^{-d_1/2} \|m_1\|_2).
\end{aligned}$$

The proof can be found in Section E.1 and E.2 in the Appendix. Since we require $\lambda_1^{-1} = o(Nh^{d_1})$, the best convergence rate that we can achieve in Theorem 5.2 is arbitrarily close to the optimal rate $N^{-1/2} h^{-d_1/2}$. It is unclear if this arbitrarily small gap is an artifact of our proof structure. However, in Theorem 5.4 below, we show that this gap can be closed by using the proposed augmented estimator.

Remark 3. [3] adopts an inverse probability weighting (IPW) method to estimate the PCATE, where the propensity scores are approximated parametrically or by kernel smoothing. They provide point-wise convergence result for their estimators, as opposed to L_2 convergence in our theorem. For their nonparametric propensity score estimator, their result is derived based on a strong smoothness assumption of the propensity score. More specifically, it requires high-order kernels (the order should not be less than d) in estimating both the propensity

score and the later PCATE in order to achieve the optimal convergence rate. Compared to their results, our proposed estimator does not involve such a strong smoothness assumption nor a parametric specification of the propensity score.

5.3. L_∞-norm balancing

In Section 3.2, we mention several choices of the metric in the balancing error (3.4). In this subsection, we provide a theoretical investigation of an important case with L_∞-norm. We note that efficient computation of the corresponding weights is challenging, and thus is not pursued in the current paper. Nonetheless, it is theoretically interesting to derive the convergence result for the proposed estimator with L_∞-norm. More specifically, the estimator of interest in this subsection is defined by replacing the L_2-norm in $S_{N,h}(w,u)$ and $R_{N,h}(w)$ with the L_∞-norm. Instead of the L_2 convergence rate (Theorem 5.2), we can obtain the uniform convergence rate of this estimator in the following theorem.

Theorem 5.3. *Suppose Assumptions 1–9 hold, Let $\tilde{w}$ be the solution to (3.6) but with $S_{N,h}(w,u) = \|\mathcal{M}_{N,h,w}(u,\cdot)\|_\infty$ and $R_{N,h}(w) = \|\frac{1}{N}\sum_{i=1}^N T_i w_i \tilde{K}_h(V_i,\cdot)\|_\infty$. If $\lambda_1^{-1} \asymp Nh^{d_1}\log(1/h)$, $\lambda_2 \asymp N^{-1}$, $\log(1/h)/(\log\log N) \to \infty$ as $N\to\infty$, and $h^2 = o\{(N^{-1}h^{-d_1}\log(1/h))^{1/2}\}$,*

$$\left\| \frac{1}{N}\sum_{i=1}^N T_i \hat{w}_i Y_i K_h(V_i,\cdot) - \mu_1 \right\|_\infty = \mathcal{O}_p\{N^{-1/2}h^{-d_1/2}\log^{1/2}(1/h)\}.$$

We provide the proof outline in Section E.3 in the Appendix.

Different from Theorem 5.2, the uniform convergence rate is optimal. Roughly speaking, this is because, compared to the optimal L_2 convergence rate, the optimal uniform convergence rate has an extra logarithmic order, which dominates the arbitrarily small gap mentioned in Section 5.2.

5.4. Augmented estimator

We also derive the asymptotic property of the augmented estimator.

Theorem 5.4. *Suppose Assumptions 1–9 hold. Take $e = m_1 - \hat{m}_1 \in \mathcal{H}$ such that $\|e\|_{\mathcal{H}} = o_p(1)$ and $\|e\|_2 = o_p(1)$. Suppose $\lambda_1^{-1} = o(Nh^{d_1})$, $\lambda_2^{-1} = \mathcal{O}(\lambda_1^{-1}h^{-d_1})$, and $h^2 = o\{(N^{-1}h^{-d_1})^{1/2}\}$, we have*

$$\left\| \frac{1}{N}\sum_{i=1}^N \tilde{K}_h(V_i,\cdot)\hat{m}_1(X_i) + \frac{1}{N}\sum_{i=1}^N T_i\hat{w}_i\tilde{K}_h(V_i,\cdot)\{Y_i - \hat{m}_1(X_i)\} - \mathbb{E}\{Y(1)|V=\cdot\} \right\|_2 = \mathcal{O}_p(N^{-1/2}h^{-d_1/2} + \lambda_1^{1/2}\|e\|_{\mathcal{H}} + \lambda_2^{1/2}h^{-d_1/2}\|e\|_2).$$

Remark 4. In Theorem 5.2, to obtain the best convergence rate that is arbitrarily close to $N^{-1/2}h^{-d_1/2}$, we require λ_1 and λ_2 to be arbitrarily close to $N^{-1}h^{-d_1}$ and N^{-1}, respectively. While in Theorem 5.4, as long as $\lambda_1 =$

$\mathcal{O}(N^{-1}h^{-d_1}\log(1/h)\|e\|_{\mathcal{H}}^{-2})$ and $\lambda_2 = \mathcal{O}(N^{-1}\log(1/h)\|e\|_N^{-2})$, the optimal convergence rate $N^{-1/2}h^{-d_1/2}$ is achievable. Therefore, with the help of augmentation, we can relax the order requirement of the tuning parameters for achieving the optimal rate. As a result, it is "easier" to tune λ_1 and λ_2 with augmentation.

Remark 5. Several existing works focus on estimating the FCATE $\gamma(\cdot)$ given the full set of covariates [24, 30]. While one could partially marginalize their estimate $\hat{\gamma}(\cdot)$ of $\gamma(\cdot)$ to obtain an estimate $\check{\tau}(\cdot)$ of $\tau(\cdot)$, it is not entirely clear whether the convergence rate of $\check{\tau}(\cdot)$ is optimal, even when $\hat{\gamma}(\cdot)$ is rate-optimal non-parametrically. The main reason is that the estimation error $\hat{\gamma}(x) - \gamma(x)$ are dependent across different values of x. Note that $\gamma(\cdot)$ is a d-dimensional function and the optimal rate is slower than the optimal rate that we achieve for $\tau(\cdot)$, a d_1-dimensional function, when $d_1 < d$. So the partially marginalizing step needs to be shown to speed up the convergence significantly, in order to be comparable to our rate result.

Remark 6. To directly estimate the PCATE $\tau(\cdot)$, a common approach is to apply smoothing methods to the adjusted responses with respect to V instead of X. Including ours, most papers follow this approach. The essential difficulty discussed in Remark 5 remains and hence the analyses are more challenging than those for the FCATE $\gamma(\cdot)$, if the optimal rate is sought. In the existing work [27, 36, 48, 13] that adopts augmentation, estimations of both propensity score and outcome mean functions, referred to as nuisance parameters in below, are required. [27] adopts parametric modeling for both nuisance parameters and achieve double robustness; *i.e.*, only one nuisance parameter is required to be consistent to achieve the optimal rate for $\tau(\cdot)$. However, parametric modeling is a strong assumption and may be restrictive. [36, 48, 13] adopt nonparametric nuisance modeling. Importantly, to achieve optimal rate of $\tau(\cdot)$, these works require consistency of *both* nuisance parameter estimations. In other words, the correct specification of both nuisance parameter models are required. [13] require both nuisance parameters to be estimated consistently with respect to L_∞ norm. While [36] and [48] implicitly require the product convergence rates from the two estimators to be faster than $N^{-1/2}$ to achieve the optimal rate of the PCATE estimation. In other words, if one nuisance estimator is not consistent, the other nuisance estimator has to converge faster than $N^{-1/2}$. [11] proposes a new orthogonal representation of $\tau(v)$ for a fixed v based on the outcome mean function and the weight function. Since their method targets at $\tau(v)$ for a fixed v, their weight function needs to be re-estimated when v changes. On the contrary, our weights are designed for the estimation of $\tau(\cdot)$ as a function and can be used for all v. In order to achieve the optimal convergence rate, [11] also require that models for both the outcome function and weight function are consistently estimated. Overall, unlike these existing estimators, our estimators do not rely on restrictive parametric modeling nor consistency of both nuisance parameter estimation.

Remark 7. Most existing work (discussed in Remark 6) require data-splitting or cross-fitting to remove the dependence between nuisance parameter estima-

TABLE 1
Models for simulation with two specifications for each of $\text{logit}\{\pi(X)\}$ *and* $m_t(X)$ $(t = 0, 1)$

Setting	$\pi(X)$	$m_t(X)$ $(t = 0, 1)$	$\tau(v)$
1	$1/(1 + \exp X_1 + X_3)$	$10 + X_1 + (2t - 1)(X_2 + X_4)$	$2v^2 + 2\sin(2v)$
2	$1/(1 + \exp Z_1 + Z_2 + Z_3)$	$10 + X_1 + (2t - 1)(X_2 + X_4)$	$2v^2 + 2\sin(2v)$
3	$1/(1 + \exp X_1 + X_3)$	$10 + (2t - 1)(Z_1^2 + 2Z_1 \sin(2Z_1)) + Z_2^2 + \sin(2Z_3)Z_4^2$	$2v^2 + 4v \sin(2v)$
4	$1/(1 + \exp Z_1 + Z_2 + Z_3)$	$10 + (2t - 1)(Z_1^2 + 2Z_1 \sin(2Z_1)) + Z_2^2 + \sin(2Z_3)Z_4^2$	$2v^2 + 4v \sin(2v)$

tions and the smoothing step for estimating $\tau(\cdot)$, which is crucial in their theoretical analyses. [47] first propose cross-fitting in the context of Target Maximum Likelihood Estimator and [10] subsequently apply to estimating equations. This technique can be used to relax the Donsker conditions required for the class of nuisance functions. [24] applies cross-fitting to FCATE estimation for similar purposes. While data-splitting and cross-fitting are beneficial in theoretical development, they are not generally a favorable modification, due to criticism of increased computation and fewer data for the estimation of different components (nuisance parameter estimation and smoothing). However, our estimators do not require data-splitting in both theory and practice. Our asymptotic analyses are non-standard and significantly different than these existing work since, without data-splitting, the estimated weights are intimately related with each others and an additional layer of smoothing further complicates the dependence structure.

6. Simulation study

We evaluate the finite-sample properties of various estimators with sample size $N = 100$. The covariate $X_i = [X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)}] \in \mathbb{R}^4$, $i = 1, \ldots, 100$ is generated by $X_i^{(1)} = Z_i^{(1)}$, $X_i^{(2)} = \{Z_i^{(1)}\}^2 + Z_i^{(2)}$, $X_i^{(3)} = \exp(Z_i^{(3)}/2) + Z_i^{(2)}$ and $X_i^{(4)} = \sin(2Z_i^{(1)}) + Z_i^{(4)}$ with $Z_i^{(j)} \sim \text{Uniform}[-2, 2]$ for $j = 1, \ldots, 4$. The conditioning variable of interest is set to be $V = X_1$. The treatment is generated by $T \mid X \sim \text{Bernoulli}\{\pi(X)\}$, and the outcome is generated by $Y \mid (T = t, X) \sim \text{N}\{m_t(X), 1\}$. To assess the estimators, we consider two different choices for each of $\pi(X)$ and $m_t(X)$, summarized in Table 1. In Settings 1 and 2, the outcome mean functions m_t are relatively easy to estimate, as they are linear with respect to covariates X. While in Settings 3 and 4, the outcome mean functions are nonlinear and more complex. Propensity score function $\pi(X)$ is set to be linear with respect to X in Settings 1 and 3, and nonlinear in Settings 2 and 4. The corresponding PCATEs are nonlinear and shown in Figure 1.

In our study, we compare the following estimators for $\tau(\cdot)$:

1. IPW: the inverse propensity weighting estimator from [3] with a logistic regression model for the propensity score. In Settings 1 and 3, the propensity score model is correctly specified.

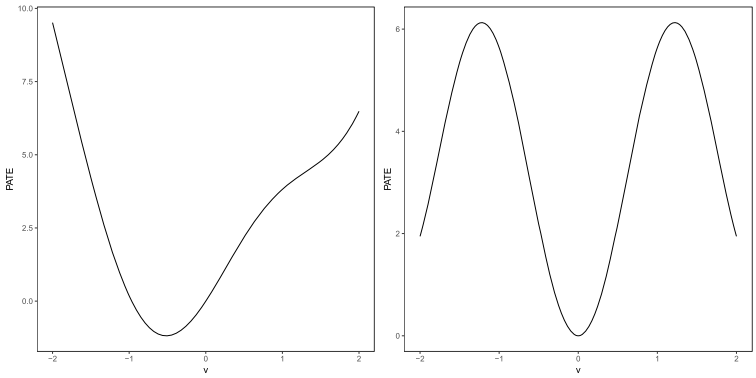

FIG 1. *The target PCATEs in the simulation study: the left panel plots the PCATE in Settings 1 and 2; the right panel plots the PCATE in Settings 3 and 4.*

2. ATE_{RKHS}: the weighted estimator described in Remark 2, whose weights are estimated based on the covariate balancing criterion in [45].
3. PROPOSED: the proposed estimator with the tensor product of second-order Sobolev kernels as the reproducing kernel κ.
4. Augmented estimators by augmenting the estimators in 1-3 by the outcome models. We consider two outcome models: linear regression (LM) and kernel ridge regression (KRR).
5. REG: the estimator that adopts outcome regressions [e.g., 21, 19]. To provide a better comparison between different methods, we directly smooth $\{(X_i, \hat{m}_1(X_i) - \hat{m}_0(X_i)) : i = 1, \ldots, N\}$ to estimate the PCATE, where $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$ are estimated with outcome models considered in 4.

For all estimators, a kernel smoother with Gaussian kernel is applied to the adjusted responses. For IPW, the bandwidth is set as $\tilde{h} = \hat{h} \times N^{1/5} \times N^{-2/7}$, where $\hat{h}$ is a commonly used optimal bandwidth in the literature such as the direct plug-in method [35, 42, 8]. Throughout our analysis, $\hat{h}$ is computed via the R package "nprobust". The same bandwidth formula $\tilde{h}$ is also considered by [27] and [13] to estimate the PCATE. For the proposed estimator, a bandwidth should be given prior to estimate the weights. We first compute the adjusted response by using weights from [45], and then obtain the bandwidth $\tilde{h}$ as the input to our proposed estimator.

Table 2 shows the average integrated squared error (AISE) and median integrated squared error (MeISE) of above estimators over 500 simulated datasets. Without augmentation, PROPOSED has significantly smaller AISE and MeISE than other methods among all four settings. All methods are improved by augmentations. In Settings 1 and 2, REG has the best performance. In these two settings, the outcome models are linear and thus can be estimated well by both LM and KRR. However, the differences between REG and PROPOSED are relatively small. As for Settings 3 and 4 where outcome mean functions are more complex, PROPOSED achieves the best performance and shows a significant im-

TABLE 2

Simulation results for the four settings, where the average integrated squared errors (AISE) with standard errors (SE) in parentheses and median integrated squared error (MeISE) are provided.

		Setting 1		Setting 2		Setting 3		Setting 4	
Augmentation	Method	AISE	MeISE	AISE	MeISE	AISE	MeISE	AISE	MeISE
No	IPW	164.202 (31.06)	49.842	92.218 (14.12)	46.916	314.273 (114.3)	77.336	121.435 (18.36)	63.109
	ATE_{RKHS}	70.236 (3.01)	50.961	57.551 (3.77)	39.123	92.154 (3.68)	71.184	70.091 (3.33)	48.937
	PROPOSED	15.737 (1.94)	6.500	5.993 (0.59)	2.33	17.874 (1.36)	9.074	8.797 (0.97)	3.265
LM	IPW	1.524 (0.06)	1.229	1.393 (0.04)	1.184	7.711 (2.57)	2.919	4.240 (0.32)	2.542
	ATE_{RKHS}	1.462 (0.04)	1.295	1.424 (0.04)	1.275	4.255 (0.19)	3.008	3.170 (0.10)	2.677
	PROPOSED	1.245 (0.03)	1.089	1.166 (0.03)	1.079	3.244 (0.18)	2.251	2.176 (0.06)	1.937
	REG	1.055 (0.03)	0.909	0.974 (0.02)	0.861	5.496 (0.19)	4.388	4.260 (0.05)	4.097
KRR	IPW	1.578 (0.08)	1.225	1.304 (0.04)	1.112	3.758 (0.21)	2.433	2.909 (0.17)	2.066
	ATE_{RKHS}	1.526 (0.04)	1.361	1.404 (0.04)	1.256	3.600 (0.14)	2.735	2.710 (0.08)	2.315
	PROPOSED	1.329 (0.04)	1.141	1.159 (0.03)	1.032	2.858 (0.12)	2.126	2.103 (0.06)	1.761
	REG	1.239 (0.04)	1.052	0.999 (0.02)	0.899	3.787 (0.12)	3.159	2.816 (0.06)	2.654

provement over REG, especially when outcome models are misspecified (See Settings 3 and 4 with LM augmentation). As ATE_{RKHS} is only designed for marginal covariate balancing, its performance is worse than PROPOSED across all scenarios.

7. Application

We apply the estimators in Section 6 to estimate the effect of maternal smoking on birth weight as a function of mother's age, by re-analyzing a dataset of mothers in Pennsylvania in the USA (http://www.stata-press.com/data/r13/cattaneo2.dta). Following [27], we focus on white and non-Hispanic mothers, resulting in the sample size $N = 3754$. The outcome Y is the infant birth weight measured in grams and the treatment indicator T is whether the mother is a smoker. For the treatment ignorability, we include the following covariates: mother's age, an indicator variable for alcohol consumption during pregnancy, an indicator for the first baby, mother's educational attainment, an indicator for the first prenatal visit in the first trimester, the number of prenatal care visits, and an indicator for whether there was a previous birth where the newborn died. Due to the boundary effect of the kernel smoother, we focus on $\tau(v)$ for $v \in [18, 36]$, which ranges from 0.05 quantile to 0.95 quantile of mothers' ages in the sample.

We compute various estimators of the PCATE in Section 6. For all the following IPW related estimators, logistic regression is adopted to estimate propensity scores. Following [3], we include IPW: the IPW estimator with no augmentation. Following [27], we include AIPW(LM): the IPW estimator with LM augmentation. We include Proposed: the proposed estimators with KRR augmentation here as it performs the best in the simulation study and aligns with our assumption for the outcome mean functions. For completeness, we also include AIPW(KRR): the IPW estimator with KRR augmentation; REG(KRR): the REG estimator where the outcome mean functions are estimated by KRR; REG(LM): the REG estimator where the outcome mean functions are estimated by LM. For both the KRR augmentation and the weights estimation in Proposed, we consider a tensor product RKHS, with the second-order Sobolev

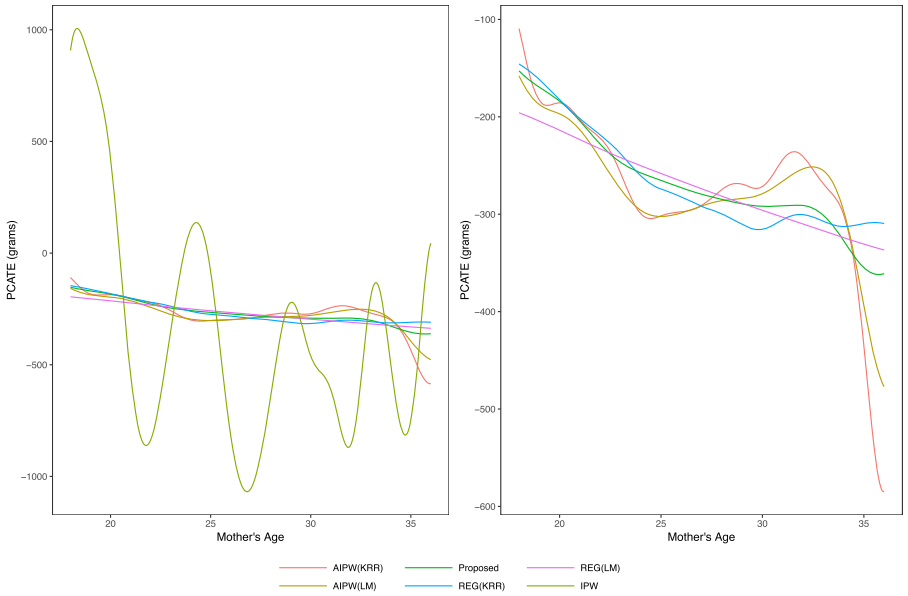

FIG 2. *The estimated PCATEs of maternal smoking on birth weight as a function of mother's age: the left panel includes all estimators and the right panel excludes the IPW estimator.*

kernel for continuous covariates and the identity kernel for binary covariates. For all the estimators, a kernel smoother with Gaussian kernel is applied to the adjusted responses.

Figure 2 shows the estimated PCATEs from different methods. From the left panel in Figure 2, IPW has large variations compared to other estimators. The significantly positive estimates before age 20 conflict with the results from various established research works indicating that smoking has adverse effect on birth weights [26, 4, 1, 2]. From the right panel in Figure 2, the remaining four estimators show a similar pattern that the effect becomes more severe as mother's age increases, which aligns with the existing literature [14, 41]. The REG(LM) estimator shows a linearly decreasing pattern, while the REG(KRR) estimator stops decreasing after age 30. AIPW(LM) and AIPW(KRR) show an increase pattern around age 27 to 32 and decrease quickly after age 32. Compared to AIPW(LM) and AIPW(KRR), Proposed shows stable effects between age 27 and 32 and the decrease after age 32 is relatively smoother.

8. Discussions

The PCATE characterizes subgroup treatment effects and provides insights about how treatment effect varies across the characteristics of interest. We develop a novel nonparametric estimator for the PCATE under treatment ignorability. The proposed double kernel weighting is a non-trivial extension of

covariate balancing weighting in the ATE estimation literature in that it aims to achieve approximate covariate balancing for all flexible outcome mean functions and for all subgroups defined based on continuous variables. In contrast to existing estimators, we do not require any smoothness assumption on the propensity score, and thus our weighting approach is particularly useful in studies when the treatment assignment mechanism is quite complex.

We conclude with several interesting and important extensions of the current estimator as future research directions. First, an improved data-adaptive bandwidth selection procedure is worth investigating as it plays an important role in smoothing. In addition, instead of local constant regression, other alternatives such as linear or spline smoothers can be considered. Third, given the appealing theoretical properties, we will investigate efficient computation of the proposed weighting estimators with L_∞-norm. Furthermore, the asymptotic distribution of the proposed estimator is worth studying so that inference procedures can be developed. Although some existing works [e.g., 31, 11, 48] provide asymptotic distributions for their proposed estimators, their analyses usually require more stringent assumptions. Also, the asymptotic distributional results under our setting are more challenging to obtain, due to the complex dependency structure resulting from the weights.

Appendix A: Comparisons with existing works

A.1. Comparisons with [48] and [13]

All three methods (the proposed method, [48] and [13]) perform non-parametric regression of the adjusted response on V to derive the corresponding PCATE estimations. They can all incorporate the outcome mean functions estimations into the adjusted response. Here, we list some differences between the proposed estimator with estimators provided in [48] and [13] as follows:

- Both [48] and [13] adopt inverse propensity weights which are known to be unstable. In contrast, the proposed estimation uses more stable balancing weights.
- Both [48] and [13] require nuisance parameters (propensity function and outcome mean function) to be consistently estimated, in order to achieve the optimal convergence rate. More specifically, they require the product convergence rates from two estimator to be faster than $N^{-1/2}h^{d_1/2}$. As for the proposed estimator, consistency of both nuisance parameter estimations are not required to achieve optimal convergence rate. See also Remark 6.
- [48] requires a data-splitting procedure to perform their estimation. While data-splitting is beneficial in theoretical development, it is not generally a favorable modification, due to criticism of increased computation and fewer data. In addition, the practical performance can be unstable. See the simulation study in Appendix C.1. In contrast, our estimator does not require data-splitting in both theory and practice. See also Remark 7.

Last, we note that, in addition to a data-splitting procedure, [13] also proposed a full-sample estimation procedure. But the corresponding analysis of the full-sample estimation requires stronger assumptions to achieve the optimal convergence rate.

A.2. Comparison with the weights in [45]

In this section, we provide more details about the comparison between the proposed weights constructed through (3.6) with the weights from [45].

First, as in Remark 2, we would like to emphasize that the proposed weights in [45] only control the balancing error with respect to the global mean (not the FCATE), which is a **scalar**. More specifically, their work aim to find weights $\{w_i\}$ such that

$$\frac{1}{N}\sum_{i=1}^{N} T_i w_i u(X_i) \approx \frac{1}{N}\sum_{i=1}^{N} u(X_i), \tag{A.1}$$

for all u in an appropriate class of functions. (Note that both LHS and RHS are scalars.) There is no guarantee that the resulting weights will ensure enough balance for the estimation of PCATE, which is a **function** of v. In contrast, our work focuses on the following balancing criterion that ensures the resulting weighted kernel smoothing estimator to remove confounding at all level of v:

$$\frac{1}{N}\sum_{i=1}^{N} T_i w_i u(X_i)\tilde{K}_h(V_i,\cdot) \approx \frac{1}{N}\sum_{i=1}^{N} u(X_i)\tilde{K}_h(V_i,\cdot), \tag{A.2}$$

for all u in an appropriate class of functions. Note that both the LHS and the RHS are functions. To quantify the balancing error, we need to quantify the differences between the functions in the LHS and the RHS. Toward this end, we proposed to use the L_2-distance which works nicely in both computation and theory.

Clearly, one could argue that the idea in [45] can be generalized to control the balancing error with respect to the function evaluated at a single point v, which is also a scalar setting. More precisely, one could extend (A.1) to

$$\frac{1}{N}\sum_{i=1}^{N} T_i w_i u(X_i)\tilde{K}_h(V_i,v) \approx \frac{1}{N}\sum_{i=1}^{N} u(X_i)\tilde{K}_h(V_i,v), \tag{A.3}$$

and construct weights $\tilde{w}(v)$ as follows:

$$\tilde{w}(v) = \underset{w\geq 1}{\operatorname{argmin}} \left[\sup_{u\in\mathcal{H}_N} \left(\left\{ \frac{1}{N}\sum_{i=1}^{N} T_i w_i u(X_i)\tilde{K}_h(V_i,v) - \frac{1}{N}\sum_{i=1}^{N} u(X_i)\tilde{K}_h(V_i,v) \right\}^2 - \lambda_1\|u\|_{\mathcal{H}}^2 \right) + \lambda_2 \frac{1}{N}\sum_{i=1}^{N} T_i w_i^2 \tilde{K}_h^2(V_i,v) \right]. \tag{A.4}$$

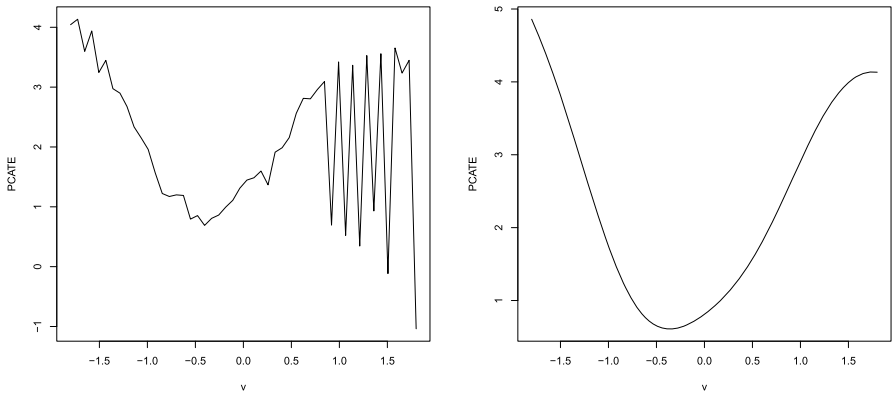

FIG 3. *Estimated PCATEs from one simulation where data is generated under Setting 1: the left panel is the result using weights constructed from* (A.4) *and the right panel is the result using the proposed weights constructed from* (3.6).

However, along this line of thinking, the weights need to be re-estimated for every single value of v. In other words, whenever we would like to estimate the PCATE evaluated at a new value of v, we have to estimate the weights based on (A.3) for that v again. As v changes, the estimated weights will also change accordingly. There are three issues with this approach. First, the weights are introduced to balance the covariate distribution. Therefore, like inverse propensity scores, they are expected to be the same for all v. Second, it is computationally expensive to re-estimate the weights whenever a new value of v is interested. Third, the changes of weights lead to non-smooth estimation of the PCATE function. In this section, we provide numerical experiment to illustrate this point. We randomly generated a sample under Setting 1, and estimated weights through (A.3) for 50 discrete evaluation points v. Then for every evaluation points v, we performed the local constant regression using adjusted response $Y_i\tilde{w}_i(v)$. The resulting PCATE function can be found in the left panel of Figure 3. As we can see, it is quite non-smooth, while the PCATE estimated by our proposed estimator enjoys smoothing property. Besides, we would like to highlight that the asymptotic theory of [45] would not be applicable here because in (A.3), the bandwith h would need to diminishes asymptotically to avoid bias.

Therefore, similarly as in the theory of kernel regression, the effective sample size in (A.3) would not be of order N. In fact, due to the nonparametric nature, one would expect a nonparametric convergence rate for tne PCATE estimator, instead of the $\sqrt{N}$ rate obtained by [45]. In a sharp contrast, the key feature for the proposed weights is that the corresponding balance is in the function sense, and therefore these weights work for every v **simultaneously**.

Appendix B: Computation

B.1. Reparametrization

To solve (3.9), we focus on the inner optimization of (3.6): $\sup_{u\in\mathcal{H}_N}\{S_{N,h}(w,u)-\lambda_1\|u\|_{\mathcal{H}}^2\}$, which is equivalent to

$$\sup_{u\in\mathcal{H}}\left\{\frac{S_{N,h}(w,u)}{\|u\|_N}-\lambda_1\frac{\|u\|_{\mathcal{H}}^2}{\|u\|_N}\right\}. \tag{B.1}$$

By the representer theorem [40], the solution to this infinite dimensional optimization (B.1) can be shown to lie in a finite dimensional subspace of $\mathcal{H}$: $\text{span}\{\kappa(X_i,\cdot) : i = 1,\dots,N\}$. Letting $M = [\kappa(X_i,X_j)]_{i,j=1}^N \in \mathbb{R}^{N\times N}$, (B.1) becomes

$$\sup_{\alpha\in\mathbb{R}^N}\left[\frac{S_{N,h}\{w,\sum_{j=1}^N\alpha_j\kappa(X_j,\cdot)\}}{\alpha^\intercal M^2\alpha/N}-\lambda_1\frac{\alpha^\intercal M\alpha}{\alpha^\intercal M^2\alpha/N}\right]. \tag{B.2}$$

By the definition of $S_{N,h}(w,u)$ in (3.4), we have

$$S_{N,h}\left\{w,\sum_{j=1}^N\alpha_j\kappa(X_j,\cdot)\right\}=\frac{1}{N^2}\alpha^\intercal M\text{diag}(T\circ w-J)G_h\text{diag}(T\circ w-J)M\alpha,$$

where $\circ$ represents the element-wise product of two vectors, $J=(1,1,\dots,1)^\intercal\in\mathbb{R}^N$, and

$$G_h=\begin{bmatrix}\int_{\mathcal{V}}\tilde{K}_h(V_1,v)\tilde{K}_h(V_1,v)\mathrm{d}v & \cdots & \int_{\mathcal{V}}\tilde{K}_h(V_1,v)\tilde{K}_h(V_N,v)\mathrm{d}v\\ \vdots & \ddots & \vdots\\ \int_{\mathcal{V}}\tilde{K}_h(V_N,v)\tilde{K}_h(V_1,v)\mathrm{d}v & \cdots & \int_{\mathcal{V}}\tilde{K}_h(V_N,v)\tilde{K}_h(V_N,v)\mathrm{d}v\end{bmatrix}.$$

Because M is positive semi-definite, we consider the eigen-decomposition of M as

$$M=PDP^\intercal \tag{B.3}$$

where $D\in\mathbb{R}^{r\times r}$ is a diagonal matrices with nonzero diagonal entries, and $P\in\mathbb{R}^{N\times r}$ is an orthonormal matrix. Letting $\beta=N^{-1/2}DP^\intercal\alpha$, (B.2) becomes

$$\sup_{\beta\in\mathbb{R}^r:\|\beta\|_2\le1}\beta^\intercal\left\{\frac{1}{N}P^\intercal\text{diag}(T\circ w-J)G_h\text{diag}(T\circ w-J)P-N\lambda_1D^{-1}\right\}\beta.$$

Therefore, (3.9) can be reparameterized as

$$\hat{w}=\underset{w\ge1}{\text{argmin}}\left[\Omega\left\{\frac{1}{N}P^\intercal\text{diag}(T\circ w-J)G_h\text{diag}(T\circ w-J)P-N\lambda_1D^{-1}\right\}+\lambda_2R_{N,h}(w)\right]. \tag{B.4}$$

B.2. Proof of Lemma 1

By the definition (3.5), $R_{N,h}(w)$ is a convex function of w. Also, $P^\intercal(T \circ w - J)$ is an affine transformation of w. Then it suffices to show that $\Omega\{\text{diag}(y)G_h\text{diag}(y)+B\}$ is a convex function of y for any symmetric matrix $B \in \mathbb{R}^{r\times r}$.

First, we show that G_h is a positive semi-definite matrix. For any vector $a \in \mathbb{R}^N$, we have

$$\begin{aligned} a^\intercal G_h a &= \int_{\mathcal{V}} a^\intercal \begin{bmatrix} \tilde{K}_h(V_1,v)\tilde{K}_h(V_1,v) & \cdots & \tilde{K}_h(V_1,v)\tilde{K}_h(V_N,v) \\ \vdots & \ddots & \vdots \\ \tilde{K}_h(V_N,v)\tilde{K}_h(V_1,v) & \cdots & \tilde{K}_h(V_N,v)\tilde{K}_h(V_N,v) \end{bmatrix} a \, dv \\ &= \int_{\mathcal{V}} \left\{ \sum_{j=1}^N \tilde{K}_h(V_j,v)a_j \right\}^2 dv \geq 0. \end{aligned}$$

Therefore there exists a matrix L such that $G_h = LL^\intercal$.

Consider any vector $y_1, y_2 \in \mathbb{R}^r$, and $t \in [0,1]$. For $\beta \in \mathbb{R}^r$, we have

$$\begin{aligned} &\beta^\intercal \left[\text{diag}\{ty_1+(1-t)y_2\}G_h\text{diag}\{ty_1+(1-t)y_2\}+B\right]\beta \\ &= \beta^\intercal \left[\text{diag}\{ty_1+(1-t)y_2\}LL^\intercal\text{diag}\{ty_1+(1-t)y_2\}+B\right]\beta \\ &= \|L^\intercal\text{diag}\{ty_1+(1-t)y_2\}\beta\|_2^2 + \beta^\intercal B\beta \\ &= \|tL^\intercal\text{diag}(y_1)\beta + (1-t)L^\intercal\text{diag}(y_2)\beta\|_2^2 + \beta^\intercal B\beta \\ &\leq t\|L^\intercal\text{diag}(y_1)\beta\|_2^2 + (1-t)\|L^\intercal\text{diag}(y_2)\beta\|_2^2 + \beta^\intercal B\beta \\ &= t\beta^\intercal\{\text{diag}(y_1)G_h\text{diag}(y_1)+B\}\beta + (1-t)\beta^\intercal\{\text{diag}(y_2)G_h\text{diag}(y_1)+B\}\beta, \end{aligned}$$

where the above inequality is due to the fact that $\|y\|_2^2$ is a convex function of y. Therefore, we have

$$\begin{aligned} &\Omega\left(\text{diag}\{ty_1+(1-t)y_2\}G_h\text{diag}\{ty_1+(1-t)y_2\}+B\right) \\ &\quad \leq t\Omega\left(\text{diag}(y_1)G_h\text{diag}(y_1)+B\right) + (1-t)\Omega\left(\text{diag}(y_2)G_h\text{diag}(y_2)+B\right), \end{aligned}$$

which completes the proof.

Appendix C: Simulation

C.1. Additional simulation results for AIPW estimators

In this section, we provide additional simulation results for the AIPW estimator considered in [13] and [48]. Both nuisance parameters are estimated nonparametrically. In particular, we consider the support vector machine (SVM), which is a commonly adopted nonparametric binary classification method, for estimating the propensity scores. For the estimation of outcome mean functions, we still adopt the kernel ridge regression. As [48] and [13] both require data-splitting

TABLE 3

Simulation results for AIPW *estimators when propensity scores are estimated by different nonparametric models. The average integrated squared errors (AISE) with standard errors in parentheses and median integrated squared error (MeISE) are provided.*

		Setting 1		Setting 2		Setting 3		Setting 4	
Data-splitting	Method	AISE	MeISE	AISE	MeISE	AISE	MeISE	AISE	MeISE
Yes	SVM	2.878 (0.34)	1.650	3.621 (1.17)	1.396	61.546 (34.71)	4.432	6.414 (0.82)	3.476
No	SVM	1.464 (0.08)	1.182	1.161 (0.03)	1.037	3.551 (0.33)	2.184	2.151 (0.07)	1.774
	PROPOSED	1.329 (0.04)	1.141	1.159 (0.03)	1.032	2.858 (0.12)	2.126	2.103 (0.06)	1.761

for estimations, and [13] also provide a full-sample version of their estimator, so here we present the results both with and without data-splitting.

For the implementation, we adopt `svm()` function from R package `e1071` with radial basis kernel for SVM. The tuning parameters for nonparametric models are all tuned by 5-fold cross-validation. The results can be found in Table 3.

As we can see, the results with data-splitting are significantly worse than those without data-splitting. The performance of the full-sample SVM estimator is worse than the proposed method in all four settings.

C.2. Sensitivity analysis for tuning parameters λ_1 and λ_2

In Theorem 5.4, we found that the order requirement of the tuning parameter is relaxed when augmentation is imposed. Therefore it is easier to achieve such requirement, and hence to tune the parameters. See the discussion in Remark 4. In practice, we expect the performance of the estimator would not vary much for a range of values of tuning parameters close to the optimal choice. To provide numerical evidence, we analyze the performance of our proposed estimator when changing the value of the tuning parameters. Here, we provide the simulation results for the proposed method when doubling and halving the values of selected tuning parameters λ_1 and λ_2. The results can be found in Table 4. We still use PROPOSED to represent the proposed method when the values of tuning parameters are decided by the strategy described in Section 3.3. Denote the selected tuning parameters as $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$. We use PROPOSED(2) to represent the method when the values of tuning parameters are doubled (i.e., $2\tilde{\lambda}_1$ and $2\tilde{\lambda}_2$) and use PROPOSED(0.5) to represent the method when the values of tuning pa-

TABLE 4

Simulation results of PROPOSED*(0.5),* PROPOSED *and* PROPOSED*(2) for four settings, where the average integrated squared errors (AISE) with standard errors in parentheses and median integrated squared error (MeISE) are provided.*

		Setting 1		Setting 2		Setting 3		Setting 4	
Augmentation	Method	AISE	MeISE	AISE	MeISE	AISE	MeISE	AISE	MeISE
No	PROPOSED(0.5)	14.246 (0.82)	7.025	5.73 (0.42)	2.693	18.426 (1.62)	9.473	8.510 (0.86)	3.568
	PROPOSED	15.814 (1.94)	6.502	5.993 (0.59)	2.333	17.874 (1.36)	9.074	8.797 (0.97)	3.277
	PROPOSED(2)	14.111 (0.88)	6.889	5.322 (0.38)	2.318	17.345 (0.97)	9.517	7.594 (0.53)	3.341
LM	PROPOSED(0.5)	1.237 (0.03)	1.081	1.158 (0.03)	1.065	3.195 (0.18)	2.199	2.123 (0.06)	1.906
	PROPOSED	1.245 (0.03)	1.089	1.166 (0.03)	1.079	3.244 (0.18)	2.251	2.176 (0.06)	1.937
	PROPOSED(2)	1.235 (0.03)	1.075	1.155 (0.03)	1.082	3.151 (0.17)	2.235	2.087 (0.05)	1.915
KRR	PROPOSED(0.5)	1.326 (0.04)	1.128	1.154 (0.03)	1.039	2.831 (0.12)	2.102	2.073 (0.06)	1.746
	PROPOSED	1.329 (0.04)	1.141	1.159 (0.03)	1.032	2.858 (0.12)	2.126	2.103 (0.06)	1.761
	PROPOSED(2)	1.315 (0.04)	1.117	1.150 (0.03)	1.033	2.825 (0.12)	2.092	2.067 (0.06)	1.738

rameters are halved (i.e., $0.5\tilde{\lambda}_1$ and $0.5\tilde{\lambda}_2$). As we can see, by varying the tuning parameters, the change in terms of AISE for estimators without augmentation is larger than those with augmentations.

Appendix D: Uncertainty quantification

In this section, we outline two possible ad-hoc strategies to construct the confidence interval, which can be used for practical purpose.

For the AIPW, the estimation error of the estimated weights and outcome means are shown to be negligible in its asymptotic (point-wise) variance [13]. For large samples, the proposed weights are expected to behave like propensity scores, as propensity scores are the solution to population balancing conditions. Therefore, we expect the effect of estimation errors of the weights and the outcome means to be small, if not negligible, in our estimation. In this regard, we could get the corresponding point-wise confidence interval from existing results for the typical local constant regression estimator, see [18], [48] and [27]. Mainly, the asymptotic variance can be constructed as

$$\hat{\sigma}^2(v) = \frac{1}{Nh^{d_1}} \frac{(\int K^2(v)dv) \sum_{i=1}^N \tilde{K}_h(V_i, v)(\hat{Z}_i - \hat{\tau}(V_i))^2/N}{(Nh^{d_1})^{-1} \sum_{i=1}^N K\{(V_i - v)/h\}}, \tag{D.1}$$

where $\hat{Z}_i$ is the adjusted response $\hat{Z}_i = \hat{w}_i T_i\{Y_i - \hat{m}_1(X_i)\} + \hat{m}_1(X_i) - [\hat{w}_i(1 - T_i)\{Y_i - \hat{m}_0(X_i)\} + \hat{m}_0(X_i)]$. Note that we adopted the bandwidth formula $\tilde{h} = \hat{h} \times N^{1/5} \times N^{-2/7}$ in all the numerical experiments, which is to undersmooth the data and make the asymptotic bias negligible. As an illustration, we have provided this point-wise confidence interval in the real-data example. See the 95% confidence bands for AIPW(KRR) and Proposed in Figure 4. In addition, we also evaluated the performance of this proposed confidence interval in the simulation study for IPW and Proposed weighted estimators with LM and KRR augmentations. Figures 5 and 6 show the results for $N = 100$ and $N = 200$ respectively. As we can see, Proposed provides better coverage but shorter intervals compared to AIPW in most cases. When $N = 200$, the empirical coverage probability of Proposed is more than 90% in Settings 1 and 2. In Settings 3 and 4 where the outcome mean functions are more complex, the empirical coverage probability of Proposed is more than 85%. We also provide the simulation results with respect to the L_2 loss for $N = 200$ in Table 5. The results lead to the same conclusion as that in Table 2.

Another way is to use the bootstrap. More specifically, let $(X_1^*, T_1^*, Y_1^*), \ldots, (X_N^*, T_N^*, Y_N^*)$ be the bootstrap sample. Construct the weights based on this bootstrap sample and obtain a corresponding kernel regression estimator, denoted $\hat{\tau}^*$. Now repeat the bootstrap procedure B times, this yields B estimators $\hat{\tau}^*_{(1)}, \ldots, \hat{\tau}^*_{(B)}$. Then we can estimate the variance of $\tau(v)$ by the sample variance of the B replicated estimators.

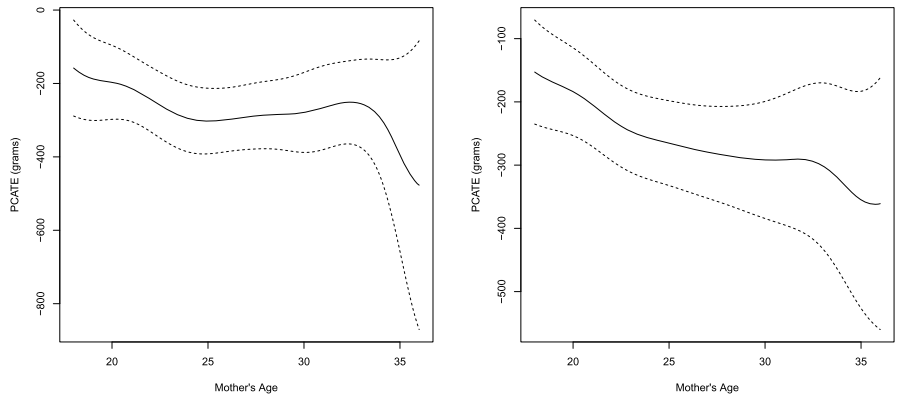

FIG 4. *Estimated PCATEs of maternal smoking on birth weight as a function of mother's age with their corresponding 95% confidence bands presented in dashed lines: the left panel is the result for AIPW(KRR) and the right panel is the result for Proposed.*

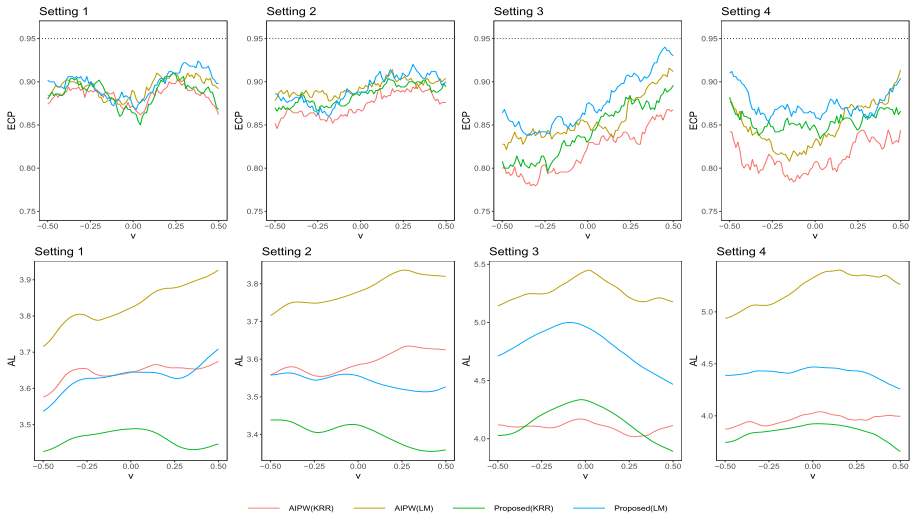

FIG 5. *Empirical coverage probability (ECP) and average length (AL) of point-wise* 95% *confidence intervals for four estimators (AIPW(LM), AIPW(KRR), Proposed(LM) and Proposed(KRR)) within the middle region of the domain of variable* v *in four simulation settings when* $N = 100$.

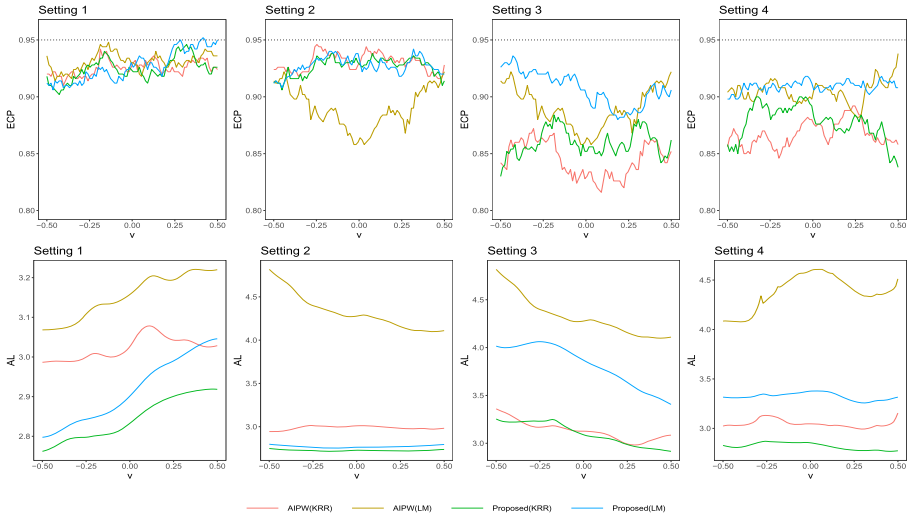

FIG 6. *Simulation results for the confidence interval when* $N = 200$. *See detailed description in Figure 5.*

TABLE 5
Simulation results for $N = 200$. *See detailed description in Table 2.*

		Setting 1		Setting 2		Setting 3		Setting 4	
Augmentation	Method	AISE	MeISE	AISE	MeISE	AISE	MeISE	AISE	MeISE
No	IPW	295.059 (191.48)	40.148	47.912 (3.88)	26.544	100.085 (10.22)	49.672	66.008 (5.43)	37.677
	ATE_{RKHS}	53.154 (2.63)	38.109	29.547 (1.42)	19.077	62.47 (2.39)	47.714	41.315 (1.73)	29.371
	PROPOSED	6.153 (0.42)	2.445	2.048 (0.46)	1.041	5.439 (0.36)	2.484	2.383 (0.15)	1.454
LM	IPW	1.129 (0.1)	0.771	0.882 (0.04)	0.705	3.574 (0.53)	1.918	2.809 (0.22)	1.734
	ATE_{RKHS}	1.069 (0.03)	0.92	0.926 (0.02)	0.808	2.916 (0.08)	2.476	2.526 (0.08)	2.167
	PROPOSED	0.742 (0.02)	0.645	0.656 (0.02)	0.59	1.581 (0.04)	1.391	1.155 (0.03)	1.015
	REG	0.61 (0.01)	0.562	0.574 (0.01)	0.536	4.376 (0.05)	4.214	3.965 (0.03)	3.906
KRR	IPW	0.952 (0.03)	0.794	0.833 (0.04)	0.684	2.267 (0.15)	1.492	1.64 (0.07)	1.248
	ATE_{RKHS}	1.077 (0.03)	0.947	0.897 (0.02)	0.803	2.456 (0.1)	1.878	1.81 (0.05)	1.575
	PROPOSED	0.777 (0.02)	0.688	0.652 (0.02)	0.579	1.87 (0.09)	1.320	1.331 (0.04)	1.116
	REG	0.692 (0.02)	0.624	0.57 (0.01)	0.538	2.06 (0.09)	1.562	1.463 (0.04)	1.250

Appendix E: Proofs of Theorems

Through out the proof, we use $\check{x}$ to represent a generic vector in $\mathcal{X}$, and use $\check{v} \in \mathcal{V}$ to represent the sub-vector of $\check{x}$ that is of interest.

E.1. Proof of Theorem 5.1

For simplicity, we introduce additional notations. Define $\gamma_i := T_i w_i^* - 1$, where $w_i^* = 1/\pi(X_i)$ for $i = 1, \ldots, N$, and $\mathcal{H}(1) := \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \leq 1\}$. By Lemma 2.1 of [28], there exists a constant b such that $\sup_{u \in \mathcal{H}(1)} |u|_\infty \leq b$.

We replace $\frac{1}{Nh^{d_1}} \sum_{j=1}^N K(\frac{V_j - v}{h})$ in $S_{N,h}(w^*, u)$ with its expectation $g_h(v)$

and obtain

$$\tilde{S}_{N,h}(w^*,u) := \left\| \frac{1}{g_h(\cdot)} \left\{ \frac{1}{Nh^{d_1}} \sum_{i=1}^{N} \gamma_i u(X_i) K\left(\frac{V_i - \cdot}{h}\right) \right\} \right\|_2^2. \tag{E.1}$$

We first show that g_h is lower bounded. By Assumption 7, without loss of generality, we consider $h \leq 1$. By Assumption 6, there exists a constant c_1 such that

$$\begin{aligned} g_h(v) &= \mathbb{E}\left\{ \frac{1}{h^{d_1}} K\left(\frac{V_i - v}{h}\right) \right\} = \frac{1}{h^{d_1}} \int_I K\left(\frac{V - v}{h}\right) g(V) dV \\ &= \int_{(zh+v)\in[0,1]^{d_1}} K(z) g(zh+v) dz \\ &\geq C_3 \int_{(zh+v)\in[0,1]^{d_1}} K(z) dz \geq C_3 \int_{(z+v)\in[0,1]^{d_1}} K(z) dz \\ &\geq C_3 \min\left\{ \int_{[0,1/2]^{d_1}} K(z) dz, \int_{[-1/2,0]^{d_1}} K(z) dz \right\} \geq c_1. \end{aligned} \tag{E.2}$$

Then, we have

$$\begin{aligned} \tilde{S}_{N,h}(w^*,u) &\leq \frac{1}{\inf_{v\in[0,1]^{d_1}} g_h^2(v)} \frac{1}{h^{2d_1}} \left\| \frac{1}{N} \sum_{i=1}^{N} \gamma_i u(X_i) K\left(\frac{V_i - \cdot}{h}\right) \right\|_2^2 \\ &\leq \frac{1}{c_1^2 h^{2d_1}} \left\| \frac{1}{N} \sum_{i=1}^{N} \gamma_i u(X_i) K\left(\frac{V_i - \cdot}{h}\right) \right\|_2^2. \end{aligned} \tag{E.3}$$

Below, we will establish the bound of $\left| N^{-1} \sum_{i=1}^{N} \gamma_i u(X_i) K\left\{(V_i - v)/h\right\} \right|$ uniformly for any $u \in \mathcal{H}_N$ for a given v. To start with, we define

$$\mathcal{F}_{h,v} := \left\{ f : f(\check{x}) = u(\check{x}) K\left(\frac{\check{v} - v}{h}\right); u \in \mathcal{H}(1) \right\}, \quad \|f\|_N := \sqrt{\frac{1}{N} \sum_{i=1}^{N} f^2(X_i)},$$

$$\mathcal{K}_h := \left\{ K\left(\frac{\cdot - v}{h}\right) : v \in [0,1]^{d_1} \right\}, \sigma_{\mathcal{K}_h,N} := \sup_{\tilde{v}\in[0,1]^{d_1}} \sqrt{\frac{1}{N} \sum_{i=1}^{N} K^2\left(\frac{V_i - \tilde{v}}{h}\right)}.$$

The next lemma provides an entropy bound for the space $\mathcal{F}_{h,v}$.

Lemma B.1. *For any fixed h and v, there exists a constant $A > 0$, such that*

$$H(\delta, \mathcal{F}_{h,v}, \|\cdot\|_N) \begin{cases} = 0 & \text{if } \delta > 2b\sigma_{\mathcal{K}_h,N} \\ \leq A\sigma_{\mathcal{K}_h,N}^{\alpha} \delta^{-\alpha} & \text{otherwise} \end{cases}.$$

Proof. For any f_1, $f_2 \in \mathcal{F}_{h,v}$, we have $\|f_1 - f_2\|_N \le \|f_1\|_N + \|f_2\|_N \le 2b\sigma_{\mathcal{K}_h,N}$. Therefore, $H\{\delta, \mathcal{F}_{h,v}, L^2(\mathbb{P}_N)\} = 0$, when $\delta > 2b\sigma_{\mathcal{K}_h,N}$. By [7], we have $H\{\epsilon, \mathcal{H}(1), \|\cdot\|_\infty\} \le A\epsilon^{-\alpha}$ for some constant $A > 0$. Therefore, the covering number $\mathcal{N}\{\epsilon, \mathcal{H}(1), \|\cdot\|_\infty\} \le \exp(A\epsilon^{-\alpha})$. Consider $\mathcal{N} \subset \mathcal{H}(1)$ as the ϵ-net of $\mathcal{H}(1)$ with respect to $\|\cdot\|_\infty$. By definition, for any $u \in \mathcal{H}(1)$, there exists a $u_0 \in \mathcal{N}$, such that

$$\sup_{x\in[0,1]^d} |u(x) - u_0(x)| \le \epsilon. \tag{E.4}$$

Consider $\mathcal{N}_v := \{f : f(\check{x}) = u(\check{x})K\{(\check{v} - v\}/h); u \in \mathcal{N}\}$. Then, for any $f \in \mathcal{F}_{h,v}$, there exists a $f_0 \in \mathcal{N}_v$, such that

$$\begin{aligned}
\|f - f_0\|_N^2 &= \frac{1}{N}\sum_{i=1}^N \left| u(X_i)K\left(\frac{V_i - v}{h}\right) - u_0(X_i)K\left(\frac{V_i - v}{h}\right)\right|^2 \\
&= \frac{1}{N}\sum_{i=1}^N K^2\left(\frac{V_i - v}{h}\right)|u(X_i) - u_0(X_i)|^2 \\
&\le \sup_{x\in[0,1]^d} |u(x) - u_0(x)|^2 \frac{1}{N}\sum_{i=1}^N K^2\left(\frac{V_i - v}{h}\right) \\
&\le \epsilon^2\sigma^2_{\mathcal{K}_h,N}.
\end{aligned}$$

The last inequality is due to (E.4) and $N^{-1}\sum_{i=1}^N K^2\{(V_i - v)/h\} \le \sigma^2_{\mathcal{K}_h,N}$. Therefore, we have

$$\mathcal{N}(\epsilon\sigma_{\mathcal{K}_h,N}, \mathcal{F}_{h,v}, \|\cdot\|_N) \le \mathcal{N}\{\epsilon, \mathcal{H}(1), \|\cdot\|_\infty\} \le \exp\left(A\epsilon^{-\alpha}\right).$$

The conclusion follows by taking $\delta = \epsilon\sigma_{\mathcal{K}_h,N}$. □

Then, we study the concentration property of the terms $\sigma_{\mathcal{K}_h,N}$ and $\sum_{i=1}^N K\{(V_i - \tilde{v})/h\}/(Nh^{d_1})$.

Lemma B.2. *Under Assumptions 5–7, there exist constants $c_2, c_3, c_4 > 0$ depending on C_2, C_3, A_1 and ν_1, such that, for all sufficiently large N, the following hold:*

$$\mathbb{E}\sigma^2_{\mathcal{K}_h,N} \le c_3 h^{d_1}, \tag{E.5}$$

$$\mathbb{P}(\sigma^2_{\mathcal{K}_h,N} \ge 2tc_3h^{d_1}) < c\exp\left\{-c_2tNh^{d_1}\right\}, \quad t \ge 1, \tag{E.6}$$

$$\mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}} \left|\frac{1}{Nh^{d_1}}\sum_{i=1}^N K\left(\frac{V_i - \tilde{v}}{h}\right) - g_h(\tilde{v})\right| \ge tc_1\right) \le c\exp\left\{-c_4tNh^{d_1}\right\}, \tag{E.7}$$

for $\frac{1}{2} \le t < 1$.

Proof. Take r_i, $i = 1, \ldots, n$, as independent Rademacher random variables. We have

$$\begin{aligned}
\mathbb{E}\sigma^2_{\mathcal{K}_h,N} &= \mathbb{E} \sup_{v\in[0,1]^{d_1}} \frac{1}{N}\sum_{i=1}^{N} K^2\left(\frac{V_i-\tilde{v}}{h}\right) \\
&\le \mathbb{E} \sup_{\tilde{v}\in[0,1]^{d_1}} \mathbb{E}K^2\left(\frac{V_i-\tilde{v}}{h}\right) + \mathbb{E} \sup_{\tilde{v}\in[0,1]^{d_1}} \left|\frac{1}{N}\sum_{i=1}^{N} K^2\left(\frac{V_i-\tilde{v}}{h}\right) - \mathbb{E}K^2\left(\frac{V_i-\tilde{v}}{h}\right)\right| \\
&= \sup_{\tilde{v}\in[0,1]^{d_1}} \mathbb{E}K^2\left(\frac{V_i-\tilde{v}}{h}\right) + \mathbb{E} \sup_{\tilde{v}\in[0,1]^{d_1}} \left|\frac{1}{N}\sum_{i=1}^{N} K^2\left(\frac{V_i-\tilde{v}}{h}\right) - \mathbb{E}K^2\left(\frac{V_i-\tilde{v}}{h}\right)\right| \\
&\le \sup_{\tilde{v}\in[0,1]^{d_1}} \mathbb{E}K^2\left(\frac{V_i-\tilde{v}}{h}\right) + 2\mathbb{E} \sup_{\tilde{v}\in[0,1]^{d_1}} \left|\frac{1}{N}\sum_{i=1}^{N} r_i K^2\left(\frac{V_i-\tilde{v}}{h}\right)\right| \\
&\le \sup_{\tilde{v}\in[0,1]^{d_1}} \mathbb{E}K^2\left(\frac{V_i-\tilde{v}}{h}\right) + 8C_2\mathbb{E} \sup_{\tilde{v}\in[0,1]^{d_1}} \left|\frac{1}{N}\sum_{i=1}^{N} r_i K\left(\frac{V_i-\tilde{v}}{h}\right)\right|.
\end{aligned}$$

The second last inequality is due to the symmetrization inequality from Theorem 2.1 in [25], while the last inequality is due to the contraction inequality from Theorem 2.3 in [25]. Next, we bound the Rademacher complexity

$$\mathbb{E}\|R_N\|_{\mathcal{K}_h} := \mathbb{E} \sup_{\tilde{v}\in[0,1]^{d_1}} \left|\frac{1}{N}\sum_{i=1}^{N} r_i K\left(\frac{V_i-\tilde{v}}{h}\right)\right|.$$

Since $\mathcal{K}_h \subset \mathcal{K}$, from the entropy bound in Assumption 5 for $\mathcal{K}$, we have $\mathcal{N}(\varepsilon, \mathcal{K}_h, \|\cdot\|_N) \le A_1\varepsilon^{-\nu_1}$. Define $\sigma^2_{\mathcal{K}_h} := \sup_{\tilde{v}\in[0,1]^{d_1}} \mathbb{E}K^2\{(V_i-\tilde{v})/h\}$. By applying Theorem 3.12 in [25], we have

$$\mathbb{E}\|R_N\|_{\mathcal{K}_h} \le c\left[\sqrt{\frac{\nu_1}{N}}\sigma_{\mathcal{K}_h}\sqrt{\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}} + \frac{\nu_1C_2}{N}\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}\right], \tag{E.8}$$

where $c > 0$ is an universal constant. Next, we have

$$\begin{aligned}
\sigma^2_{\mathcal{K}_h} &= \sup_{\tilde{v}\in[0,1]^{d_1}} \int_0^1 K^2\left(\frac{v-\tilde{v}}{h}\right) g(v)dv \\
&= h^{d_1} \sup_{\tilde{v}\in[0,1]^{d_1}} \int_{(zh+\tilde{v})\in[0,1]^{d_1}} K^2(z)g(zh+\tilde{v})dz, \\
&\le C_3h^{d_1} \sup_{\tilde{v}\in[0,1]^{d_1}} \int_{(zh+\tilde{v})\in[0,1]^{d_1}} K^2(z)dz \qquad \text{(E.9)} \\
&\le C_2^2h^{d_1}, \qquad \text{(E.10)}
\end{aligned}$$

where (E.9) and (E.10) are due to $g(\cdot) \le C_3$ and $K(\cdot) \le C_2$; (E.10) is valid for $h \le 1$. Since $\int_{[0,1]^{d_1}} K^2(z)dz > 0$, we have $\sigma^2_{\mathcal{K}_h} \asymp h^{d_1}$.

Therefore, there exists a constant $c_3 > 0$ depending on C_2, C_3, ν_1 and A_1, such that

$$\begin{aligned}\mathbb{E}\sigma^2_{\mathcal{K}_h,N} &\le \sigma^2_{\mathcal{K}_h} + 8C_2\mathbb{E}\|R_N\|_{\mathcal{K}_h} \\ &\le \sigma^2_{\mathcal{K}_h} + 8C_2 c\left[\sqrt{\frac{\nu_1}{N}}\sigma_{\mathcal{K}_h}\sqrt{\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}} + \frac{\nu_1 C_2}{N}\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}\right] \\ &\le c_3 h^{d_1}.\end{aligned}$$

The last inequality is due to Assumption 7 and it is valid for all large enough N.

From Talagrand's inequality (Theorem 2.5 in [25]), and

$$\sup_{\tilde{v}\in[0,1]^{d_1}} \mathbb{E}K^4\left(\frac{V-\tilde{v}}{h}\right) \le C_2^4 h^{d_1},$$

we have for any $t \ge 1$,

$$\mathbb{P}(\sigma^2_{\mathcal{K}_h,N} \ge 2tc_3h^{d_1}) \le c\exp\left\{-c_2 tNh^{d_1}\right\},$$

where $c > 0$ is an universal constant and $c_2 > 0$ is a constant depending on C_2, C_3, ν_1 and A_1.

Also, by adopting symmetrization inequality again, there exists a constant $c_5 > 0$ depending on A_1, ν_1 and C_2 such that

$$\begin{aligned}&\mathbb{E}\sup_{\tilde{v}\in[0,1]^{d_1}}\left|\frac{1}{N}\sum_{i=1}^N K\left(\frac{V_i-\tilde{v}}{h}\right) - \mathbb{E}K\left(\frac{V_i-\tilde{v}}{h}\right)\right| \le 2\mathbb{E}\|R_N\|_{\mathcal{K}_h} \\ &\le 2c\left[\sqrt{\frac{\nu_1}{N}}\sigma_{\mathcal{K}_h}\sqrt{\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}} + \frac{\nu_1C_2}{N}\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}\right] \qquad \text{(E.11)} \\ &\le c_5 N^{-1/2}h^{d_1/2}\sqrt{\log(1/h^{d_1})},\end{aligned}$$

where the last inequality is due to Assumption 7, and the first term of (E.11) is dominant for large enough N.

By Talagrand's inequality, for any $t > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}}\left|\frac{1}{N}\sum_{i=1}^N K\left(\frac{V_i-\tilde{v}}{h}\right) - \mathbb{E}K\left(\frac{V_i-\tilde{v}}{h}\right)\right| \ge c_5\frac{h^{d_1/2}}{N^{1/2}}\sqrt{\log(1/h^{d_1})} + t\right) \\ &\le c\exp\left(-\frac{1}{c}\frac{N^2t^2}{\tilde{V}+ntC_2}\right),\end{aligned}$$

where $\tilde{V} := NC_2^2h^{d_1} + 16C_2c_5N^{1/2}h^{d_1/2}\sqrt{\log(1/h^{d_1})} \le 2NC_2^2h^{d_1}$, for all large enough N.

Take $t = t'c_1h^{d_1} - c_5N^{-1/2}h^{d_1/2}\sqrt{\log(1/h^{d_1})}$, for $1/2 \le t' < 1$. For all large enough N, we have $t \ge t'c_1h^{d_1}/2$. Therefore, we have

$$\mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}}\left|\frac{1}{Nh^{d_1}}\sum_{i=1}^N K\left(\frac{V_i-\tilde{v}}{h}\right) - g_h(\tilde{v})\right| \ge t'c_1\right)$$

$$= \mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}} \left| \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{V_i-\tilde{v}}{h}\right) - \mathbb{E}K\left(\frac{V_i-\tilde{v}}{h}\right)\right| \geq t' c_1 h^{d_1} \right)$$
$$\leq c\exp\left\{-c_4 t' N h^{d_1}\right\},$$

where $c>0$ is universal constant and $c_4>0$ is a constant depending on C_2, C_3, A_1 and ν_1. □

Next, we derive the bound for $|\sum_{i=1}^{N}\gamma_i f(X_i)/N|$ uniformly for all $f\in\mathcal{F}_{h,v}$.

Lemma B.3. *Under Assumptions 2-7, there exists constants $c_6, c_7>0$ depending on b, C_2, A, C_1 and α such that with probability at least $1-c\exp(-c_6 t)$, $\forall f\in\mathcal{F}_{h,v}$, we have*

$$\frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right| \leq t\left\{ N^{-\frac{1}{2}}\|u\|_2^{\frac{2-\alpha}{2p}} h^{d_1\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)} + N^{-\frac{2}{2+\alpha}} h^{\frac{d_1\alpha}{2+\alpha}}\right\},$$

for any $t\geq c_7$ and $p\geq 1$.

Proof. Because $\mathbb{E}(\tilde{\gamma}_i \mid X_i)=0$ and $\gamma_i \mid X_i$, $i=1,\ldots,n$, are bounded sub-Gaussian random variables, there exists a constant $\sigma_\gamma>0$ depending on C_1, such that $\mathbb{E}\{\exp(\lambda\gamma)|X=x\}\leq\exp(\lambda^2\sigma_\gamma^2/2)$ for all x.

Define $\mathcal{F}_{h,v}(\delta):=\{f\in\mathcal{F}_{h,v}:\|f\|_2\leq\delta\}$ for $\delta>0$. We begin by deriving an upper bound for $\mathbb{E}[\sup_{f\in\mathcal{F}_{h,v}(\delta)}\sum_{i=1}^{N}\gamma_i f(X_i)/N]$. Conditioned on $X_i, i=1,\ldots,N$, $\sum_{i=1}^{N}\gamma_i f(X_i)/\sqrt{N}$ is a sub-Gaussian process with respect to the metric space $(\mathcal{F}_{h,v},\mathrm{dist})$, where $\mathrm{dist}^2(f_1,f_2)=\frac{\sigma_\gamma^2}{N}\sum_{i=1}^{N}(f_1(X_i)-f_2(X_i))^2$ for $f_1,f_2\in\mathcal{F}_{h,v}$. Therefore, by Dudley's entropy bound, and Lemma B.1, for any $\delta>0$, we have

$$\mathbb{E}\left\{\sup_{f\in\mathcal{F}_{h,v}(\delta)}\frac{1}{\sqrt{N}}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right| \mid X_i, i=1,\ldots,N\right\}$$
$$\leq c\int_0^{2\sigma_\gamma\delta_N}\sqrt{H(\tau,\mathcal{F}_{h,v},\|\cdot\|_N)}d\tau,$$

where $\delta_N^2=\sup_{f\in\mathcal{F}_{h,v}(\delta)}\left|\frac{1}{N}\sum_{i=1}^{N}f^2(X_i)\right|$.

Taking expectations on both sides and using Lemma B.1, there exists a constant $c_8>0$ depending on A, σ_γ, α and c_3 such that

$$\mathbb{E}\sup_{f\in\mathcal{F}_{h,v}(\delta)}\frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right| \leq \frac{c}{\sqrt{N}}\mathbb{E}\int_0^{2\sigma_\gamma\delta_N}\sqrt{H(\tau,\mathcal{F},\|\cdot\|_N)}d\tau$$
$$\leq \frac{c}{\sqrt{N}}\mathbb{E}\int_0^{2\sigma_\gamma\delta_N} A^{1/2}\sigma_{\mathcal{K}_h,N}^{\alpha/2}\tau^{-\alpha/2}d\tau$$
$$\leq \frac{cA^{1/2}}{\sqrt{N}}\frac{1}{1-\alpha/2}\mathbb{E}\sigma_{\mathcal{K}_h,N}^{\alpha/2}(2\sigma_\gamma\delta_N)^{1-\alpha/2}$$

$$
\begin{aligned}
&= \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} \mathbb{E}\sigma_{\mathcal{K}_h,N}^{\alpha/2} \delta_N^{1-\alpha/2} \qquad \text{(by Hölder's Inequality)} \\
&\le \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} (\mathbb{E}\delta_N)^{1-\alpha/2} (\mathbb{E}\sigma_{\mathcal{K}_h,N})^{\alpha/2} \qquad \text{(by Jensen's Inequality)} \\
&\le \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} (\mathbb{E}\delta_N^2)^{\frac{1-\alpha/2}{2}} (\mathbb{E}\sigma_{\mathcal{K}_h,N}^2)^{\alpha/4} \qquad \text{(by (E.5) in Lemma B.2)} \\
&\le \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} (\mathbb{E}\delta_N^2)^{\frac{1-\alpha/2}{2}} (c_3 h^{d_1})^{\alpha/4} \\
&\le c_8 N^{-1/2} h^{d_1\alpha/4} (\mathbb{E}\delta_N^2)^{\frac{1-\alpha/2}{2}}
\end{aligned}
$$

Next, we derive an upper bound for $\mathbb{E}\delta_N^2$. By symmetrization and contraction inequalities, we have

$$
\begin{aligned}
\mathbb{E}\delta_N^2 &\le \delta^2 + 2\mathbb{E} \sup_{f\in\mathcal{F}_{h,v}(\delta)} \left| \frac{1}{N}\sum_{i=1}^N f^2(X_i) - \mathbb{E}f^2(X_i) \right| \\
&\le \delta^2 + 2\mathbb{E} \sup_{f\in\mathcal{F}_{h,v}(\delta)} \left| \frac{1}{N}\sum_{i=1}^N r_i f^2(X_i) \right| \\
&\le \delta^2 + 8bC_2\mathbb{E} \sup_{f\in\mathcal{F}_{h,v}(\delta)} \left| \frac{1}{N}\sum_{i=1}^N r_i f(X_i) \right|,
\end{aligned}
$$

where r_i, $i = 1, \ldots, n$, are independent Rademacher random variables. Applying the entropy bound from Lemma B.1 and with Theorem 3.12 in [25], we have

$$
\mathbb{E} \sup_{f\in\mathcal{F}_{h,v}(\delta)} \left| \frac{1}{N}\sum_{i=1}^N r_i f(X_i) \right| \le c_9 \max\left\{ \frac{h^{d_1\alpha/4}}{\sqrt{N}} \delta^{1-\alpha/2}, \frac{h^{d_1\alpha/(2+\alpha)}}{N^{2/(2+\alpha)}} \right\}
$$

for some constant $c_9 > 0$ depending on A, b, C_2, α.

We now combine the above results. Also, as Assumption 7 indicates, for some constants $c_{10} > 0$ depending on $\alpha, C_2, b, c_\gamma, A$, we have

$$
\begin{aligned}
&\mathbb{E} \sup_{f\in\mathcal{F}_{h,v}(\delta)} \frac{1}{N} \left| \sum_{i=1}^N \gamma_i f(X_i) \right| \\
&\quad \le c_{10} \max\left\{ N^{-1/2} h^{d_1\alpha/4} \delta^{1-\alpha/2}, N^{-2/(2+\alpha)} h^{d_1\alpha/(2+\alpha)} \right\}.
\end{aligned}
$$

When $\delta \ge N^{\frac{-1}{2+\alpha}} h^{\frac{d_1\alpha}{2(2+\alpha)}}$, $\mathbb{E}\sup_{f\in\mathcal{F}(\delta)} \frac{1}{N}\sum_{i=1}^N \gamma_i f(X_i) \le c_{10} N^{-1/2} h^{d_1\alpha/4} \delta^{1-\alpha/2}$; By Talagrand concentration inequality, for $t \ge 1$, there exists a constant $c_{11} > 0$ depending on C_2, b, α, C_1, A, such that

$$
\mathbb{P}\left(\sup_{f\in\mathcal{F}_{h,v}(\delta)} \frac{1}{N} \left| \sum_{i=1}^N \gamma_i f(X_i) \right| > 2c_{10} t \frac{h^{d_1\alpha/4}}{N^{1/2}} \delta^{1-\alpha/2} \right) \le c \exp\left\{ -c_{11} t h^{d_1\alpha/2} \delta^{-\alpha} \right\}.
$$

When $\delta < N^{\frac{-1}{2+\alpha}} h^{\frac{d_1\alpha}{2(2+\alpha)}}$, $\mathbb{E}\sup_{f\in\mathcal{F}_{h,v}(\delta)} |\frac{1}{N}\sum_{i=1}^N \gamma_i f(X_i)| \le c_{10} N^{-2/(2+\alpha)} h^{d_1\alpha/(2+\alpha)}$. Then there exists a constant $c_{12} > 0$ depending on C_2, b, α, C_1, A, such that for $t \ge 1$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}_{h,v}(\delta)} \frac{1}{N} \left| \sum_{i=1}^N \gamma_i f(X_i) \right| > 2c_{10} t N^{\frac{-2}{2+\alpha}} h^{\frac{d_1\alpha}{2+\alpha}} \right) \le c \exp\left\{ -c_{12} t N^{\frac{\alpha}{2+\alpha}} h^{\frac{d_1\alpha}{2+\alpha}} \right\}.$$

Take $\xi_{N,h} = N^{\frac{-1}{2+\alpha}} h^{\frac{d_1\alpha}{2(2+\alpha)}}$. It is easy to see that $\|f\|_2^2 \le b^2 C_2^2 h^{d_1}$ for every $f \in \mathcal{F}_{h,v}$. We now apply the peeling technique. Take $t' = 2^{2-\alpha/2} c_{10} t$. When $\|f\|_2 > \xi_{N,h}$, there exists a constant $c_{13} > 0$ depending on C_2, b, α, C_1, A, such that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}_{h,v}:\xi_{N,h}\le\|f\|_2\le C_2 b h^{d_1/2}} \frac{\frac{1}{N}\left|\sum_{i=1}^N \gamma_i f(X_i)\right|}{\|f\|_2^{1-\alpha/2}} \ge t' N^{-1/2} h^{d_1\alpha/4} \right) \\
&\le \sum_{s=1}^{\left\lceil \log \frac{\xi_{N,h} h^{d_1/2}}{C_2 b} \right\rceil} \mathbb{P}\left(\sup_{f\in\mathcal{F}_{h,v}:2^{-s} C_2 b h^{d_1/2}\le\|f\|_2\le 2^{-s+1} C_2 b h^{d_1/2}} \frac{1}{N} \left| \sum_{i=1}^N \gamma_i f(X_i) \right| \right. \\
&\qquad\qquad \left. \ge t' N^{-1/2} h^{d_1\alpha/4} (2^{-s} C_2 b h^{d_1/2})^{1-\alpha/2} \right) \\
&= \sum_{s=1}^{\left\lceil \log \frac{\xi_{N,h}\sqrt{h}}{C_2 b} \right\rceil} \mathbb{P}\left(\sup_{f\in\mathcal{F}_{h,v}:2^{-s} C_2 b h^{1/2}\le\|f\|_2\le 2^{-s+1} C_2 b h^{1/2}} \frac{1}{N} \left| \sum_{i=1}^N \gamma_i f(X_i) \right| \right. \\
&\qquad\qquad \left. \ge 2t c_{10} N^{-1/2} h^{d_1\alpha/4} (2^{-s+1} C_2 b h^{d_1/2})^{1-\alpha/2} \right) \\
&\le \sum_{s=1}^{\infty} c \exp\{ -c_{11} t h^{d_1\alpha/2} (2^{-s+1} C_2 b h^{d_1/2})^{-\alpha} \} \\
&= \sum_{s=1}^{\infty} c \exp\{ -c_{11} t (2^{-s+1} C_2 b)^{-\alpha} \} \le c \exp\left(-c_{13} t'\right).
\end{aligned}$$

Therefore, with probability at least $1 - c\exp(-c_{13}t')$, we have $\forall f \in \mathcal{F}_{h,v}$,

$$\frac{1}{N} \left| \sum_{i=1}^N \gamma_i f(X_i) \right| \le t' \left(N^{-1/2} h^{d_1\alpha/4} \|f\|_2^{1-\alpha/2} + N^{-\frac{2}{2+\alpha}} h^{\frac{d_1\alpha}{2+\alpha}} \right), \tag{E.12}$$

for any $t' \ge 2^{2-\alpha/2} c_{10}$.

By Hölder's inequality,

$$\|f\|_2^2 = \|f^2\|_1 \le \|u^2(\cdot)\|_p \left\| K^2\left(\frac{V - \cdot}{h} \right) \right\|_q \le (b^{2p-2})^{\frac{1}{p}} \|u\|_2^{\frac{2}{p}} h^{\frac{d_1}{q}},$$

where $p, q \ge 1$ such that $1/p + 1/q = 1$. Plugging this result into (E.12) and

taking $t = t' \max\{b^2, 1\}$, we finally get $\forall f \in \mathcal{F}_{h,v}$

$$\frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right| \le t\left\{N^{-\frac{1}{2}}\|u\|_2^{\frac{2-\alpha}{2p}} h^{\frac{d_1\alpha}{4}+\frac{(2-\alpha)d_1}{4q}} + N^{-2/(2+\alpha)}h^{d_1\alpha/(2+\alpha)}\right\},$$

with probability at least $1-\exp(-c_6 t)$ for $t \ge c_7$, where $c_6, c_7 > 0$ are some constants depending on b, C_2, A, C_1 and α. □

We then relates $\|u\|_2$ to $\|u\|_N$ in the next lemma.

Lemma B.4. *There exist constants $c_{14}, c_{15} > 0$ depending on b and α, such that for $t \ge c_{14}$, we have with probability at least $1-\exp(-c_{15}tN^{\alpha/(2+\alpha)})$,*

$$\forall u \in \mathcal{H}(1) \qquad \|u\|_2^2 \le t(c_{15}N^{-\frac{2}{2+\alpha}} + \|u\|_N^2).$$

Proof. Take r_i, $i = 1, \ldots, n$, as independent rademacher random variables. From the proof of Lemma B.1, we know $\mathcal{N}(\epsilon, \mathcal{H}(1), \|\cdot\|_\infty) \le A\epsilon^{-\alpha}$ for some constant $A > 0$. Therefore, by Theorem 3.12 in [25], we have

$$\mathbb{E}\sup_{u\in\mathcal{H}(1),\|u\|\le\delta}\left|\frac{1}{N}\sum_{i=1}^{N} r_i u(X_i)\right| \le c_{16}\left(N^{\frac{-1}{2}}\delta^{1-\frac{\alpha}{2}} + N^{\frac{-1}{1+\alpha/2}}\right),$$

where $c_{16} > 0$ is a constant depending on b and α.

Next, we will adopt Theorem 3.3 in [6]. Note that

$$\mathrm{Var}\left\{u^2(X_i)\right\} \le \mathbb{E}\left\{u^4(X_i)\right\} \le b^2\|u\|_2^2.$$

Take $\psi(z) := 4c_{16}b^3\left(N^{-1/2}z^{\frac{2-\alpha}{4}}b^{(\alpha-2)/2} + N^{-1/(1+\alpha/2)}\right)$, $T(u) = b^2\|u\|_2^2$ and $B = b^2$ in Theorem 3.3 of [6]. It is easy to verify that $\psi(z)$ is non-decreasing and $\psi(z)/\sqrt{z}$ is non-increasing. In addition, we can also verify the condition that for every z,

$$b^2\mathbb{E}\sup_{u\in\mathcal{H}(1),T(u)\le z}\left|\frac{1}{N}\sum_{i=1}^{N} r_i u^2(X_i)\right| \le 4b^3\mathbb{E}\sup_{u\in\mathcal{H}(1),T(u)\le z}\left|\frac{1}{N}\sum_{i=1}^{N} r_i u(X_i)\right| \le \psi(z).$$

Then we will find the fixed points z^* of $\psi(z)$ (i.e., the solution of $\psi(z) = z$). It can be shown that $z^* = c_{15}N^{-2/(2+\alpha)}$ for some constant c_{15} depending on α and b. Therefore, Theorem 3.3 in [6] shows that with probability at least $1-\exp(-tNz^*)$,

$$\forall u \in \mathcal{H}(1) \qquad \|u\|_2^2 \le t(z^* + \|u\|_N^2),$$

with $t > c_{14}$ and a constant $c_{14} > 0$ depending on b and α. □

From Lemmas B.3 and B.4, we can see that for any $t_1, t_2 \ge \max\{c_7, c_{14}, 1\}$, with probability at least $1-\left\{c\exp(-c_6 t_1) + \exp(-c_{14}t_2N^{\alpha/(2+\alpha)})\right\}$, we have $\forall f \in \mathcal{F}_{h,v}$,

$$\frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right| \le t_1 t_2\left\{N^{\frac{-1}{2}}(\|u\|_N)^{\frac{2-\alpha}{2p}} h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1} + N^{\frac{-2}{2+\alpha}}h^{\frac{d_1\alpha}{2+\alpha}}\right.$$

$$+N^{\frac{-1}{2}-\frac{2-\alpha}{2p(2+\alpha)}}h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}\Big\}. \tag{E.13}$$

Let $s \geq 1$. Note that $\{u/\|u\|_{\mathcal{H}} : \|u\|_N \leq 1\} \subseteq \mathcal{H}(1)$. Using (E.13), we have, with probability at least $1-\left\{c\exp(-c_6t_1)+\exp(-c_{14}t_2N^{\alpha/(2+\alpha)})\right\}$, uniformly for all $u \in \mathcal{H}$ with $\|u\|_N \leq 1$,

$$\begin{aligned}\frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i\frac{u(X_i)}{\|u\|_{\mathcal{H}}}K\left(\frac{V_i-v}{h}\right)\right| &\leq t_1t_2\left\{N^{\frac{-1}{2}}\left\|\frac{u}{\|u\|_{\mathcal{H}}}\right\|_N^{\frac{2-\alpha}{2p}}h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}+N^{\frac{-2}{2+\alpha}}h^{\frac{d_1\alpha}{2+\alpha}}\right.\\ &\qquad\left.+N^{\frac{-1}{2}-\frac{2-\alpha}{2p(2+\alpha)}}h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}\right\}\\ \frac{1}{N}\left|\sum_{i=1}^{N}\gamma_iu(X_i)K\left(\frac{V_i-v}{h}\right)\right| &\leq t_1t_2\left\{N^{\frac{-1}{2}}\|u\|_{\mathcal{H}}^{1-\frac{2-\alpha}{2p}}h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}+\nu_{N,h}\|u\|_{\mathcal{H}}\right\},\end{aligned} \tag{E.14}$$

where $\nu_{N,h} := N^{-2/(2+\alpha)}h^{d_1\alpha/(2+\alpha)}+N^{\frac{-1}{2}-\frac{2-\alpha}{2p(2+\alpha)}}h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}$, $p \geq 1$. Next, we define

$$L(N,h,p,u) := N^{-\frac{1}{2}}\|u\|_{\mathcal{H}}^{1-\frac{2-\alpha}{2p}}h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}+\nu_{N,h}\|u\|_{\mathcal{H}}, \tag{E.15}$$

for any $N > 1$, $h > 0$, $p \geq 1$ and $u \in \mathcal{H}$.

Now we are able to bound $S_{N,h}(w^*,u)$ by the following lemma.

Lemma B.5. *Under Assumptions 2-7, we have*

$$\sup_{u\in\mathcal{H}_N}\frac{S_{N,h}(w^*,u)}{h^{-2d_1}\{L^2(N,h,p,u)\}} = \mathcal{O}_{\mathrm{p}}(1),$$

where L is defined in (E.15), $p \geq 1$, $h > 0$ can depend on N.

Proof. First, take

$$Q(v) := \sup_{u\in\mathcal{H}_N}\left|\frac{\frac{1}{N}\sum_{i=1}^{N}\gamma_iu(X_i)K\left(\frac{V_i-v}{h}\right)}{L(N,h,p,u)}\right|.$$

Due to (E.14), we can show that for any $t \geq \max\{c_7,c_{14},1\}$,

$$Q(v) \leq t^2,$$

with probability at least $1-2c\exp(-c_6t)$ for large enough N.

Take $\tilde{c}(k) = (\max\{c_7,c_{14},1\})^{4k}$. From the above upper bound for $Q(v)$, we have for any $v \in [0,1]^{d_1}$ and any integer $k \geq 1$,

$$\begin{aligned}\mathbb{E}\left\{Q^2(v)\right\}^k &= \int_0^\infty \mathbb{P}\left\{Q(v)^{2k} > t\right\}dt = \int_0^\infty \mathbb{P}\left\{Q(v) > t^{\frac{1}{2k}}\right\}dt\\ &\leq \tilde{c}(k)+\int_{\tilde{c}(k)}^\infty 2c\exp(-c_6t^{\frac{1}{4k}})dt\end{aligned}$$

$$= \tilde{c}(k) + 4k \int_{\max\{c_7, c_{14}, 1\}}^{\infty} 2c \exp(-c_6 t')(t')^{4k-1} dt'$$
$$\leq \tilde{c}(k) + c_{17} k\Gamma(4k),$$

where $c_{17} > 0$ is a constant depending on c_6. Note that for any fixed positive k, $\tilde{c}(k)$ and $k\Gamma(k)$ are bounded.

From (E.3), we have for $t > 0$ and positive integer k,

$$\mathbb{P}\left(\sup_{u \in \mathcal{H}_N} \frac{c_1^2 h^{2d_1} \tilde{S}_{N,h}(w^*, u)}{L^2(N, h, p, u)} \geq t \right) \leq \mathbb{P}\left(\left\{ \int_{[0,1]^{d_1}} Q^2(v) dv \right\} \geq t \right)$$
$$\leq \frac{\mathbb{E}\left[\int_{[0,1]^{d_1}} Q^2(v) dv \right]^k}{t^k} \leq \frac{\mathbb{E}\left[\int_{[0,1]^{d_1}} Q^{2k}(v) dv \right]}{t^k} \qquad \text{(by Jensen's inequality)}$$
$$\leq \frac{\int_[0, 1]^{d_1} \mathbb{E} Q^{2k}(v) dv}{t^k} \leq \frac{2^k (\tilde{c}(k) + c_{17} k\Gamma(4k))}{t^k} \leq \frac{c_{18}(k)}{t^k},$$

where $c_{18}(k) > 0$ is a constant depending on k. Then we have

$$\sup_{u \in \mathcal{H}_N} \frac{h^{2d_1} \tilde{S}_{N,h}(w^*, u)}{L^2(N, h, p, u)} = \mathcal{O}_{\mathrm{p}}(1).$$

From (E.7) in Lemma B.2, we can see that with probability at least $1 - c\exp\{-c_4 t' N h\}$, where $\frac{1}{2} \leq t' \leq 1$,

$$\forall \tilde{v} \in [0,1]^{d_1}, \qquad \left| \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h} \right) - g_h(\tilde{v}) \right| \leq t' c_1 \leq t' g_h(\tilde{v})$$
$$\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h} \right) - g_h(\tilde{v}) \geq -t' g_h(\tilde{v})$$
$$\frac{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h} \right)}{g_h(\tilde{v})} \geq 1 - t' \tag{E.16}$$
$$\frac{g_h(\tilde{v})}{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h} \right)} \leq \frac{1}{1 - t'}$$

Therefore,

$$\sup_{u \in \mathcal{H}_N} \frac{S_{N,h}(w^*, u)}{h^{-2d_1} L^2(N, h, p, u)}$$
$$\leq \sup_{u \in \mathcal{H}_N} \frac{\tilde{S}_{N,h}(w^*, u)}{h^{-2d_1} L^2(N, h, p, u)} \sup_{\tilde{v} \in [0,1]^{d_1}} \left\{ \frac{g_h(\tilde{v})}{\frac{1}{Nh^{d_1}} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h} \right)} \right\}^2 = \mathcal{O}_{\mathrm{p}}(1) \quad \square$$

Next, we control the penalty term $R_{N,h}(w^*)$ through the following lemma.

Lemma B.6. *Under Assumptions 2-7, we have*

$$R_{N,h}(w^*) = \mathcal{O}_{\mathrm{p}}(h^{-d_1}).$$

Proof. Take

$$\tilde{R}_{N,h}(w^*) := \int_{[0,1]^{d_1}} \frac{1}{g_h(v)^2} \left\{ \frac{1}{Nh^{2d_1}} \sum_{i=1}^{N} T_i w_i^{*2} K^2 \left(\frac{V_i - v}{h} \right) \right\} dv.$$

Notice that $T_i w_i^{*2}$ is upper bounded by C_1^2. By (E.6) in Lemma B.2,

$$\begin{aligned} &\sup_{\tilde{v}\in[0,1]^{d_1}} \left| \frac{1}{N} \sum_{i=1}^{N} T_i w_i^{*2} K^2 \left(\frac{V_i - \tilde{v}}{h} \right) \right| \le C_1^2 \sup_{\tilde{v}\in[0,1]^{d_1}} \left| \frac{1}{N} \sum_{i=1}^{N} K^2 \left(\frac{V_i - \tilde{v}}{h} \right) \right| \\ &= C_1^2 \sigma^2_{\mathcal{K}_h, N} \le 2C_1^2 c_3 t h^{d_1}, \end{aligned}$$

with probability at least $1 - c\exp(-c_2 t N h^{d_1})$ for $t \ge 1$. Therefore, we have

$$\tilde{R}_{N,h}(w^*) \le \int_{[0,1]^{d_1}} \frac{1}{g_h^2(v)} dv \left\{ \frac{1}{h^{2d_1}} 2C_1^2 c_3 t h^{d_1} \right\} \le \frac{2C_1^2 c_3^2}{c_1^2} t h^{-d_1}, \tag{E.17}$$

with probability at least $1 - c\exp(-c_2 t N h^{d_1}/c)$. Combining with the results from (E.16), we have

$$R_{N,h}(w^*) \le \tilde{R}_{N,h}(w^*) \left\{ \sup_{\tilde{v}\in[0,1]^{d_1}} \frac{g_h(\tilde{v})}{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h} \right)} \right\}^2 = \mathcal{O}_{\mathrm{p}}(h^{-d_1}) \qquad \square$$

Now, we are ready to prove Theorem 5.1.

Proof of Theorem 5.1. Take $u^* = \operatorname{argmax}_{u\in\mathcal{H}_N} \left\{ S_{N,h}(w^*, u) - \lambda_1 \|u\|^2_{\mathcal{H}} \right\}$. Its existence is shown in Section B.

Due to (3.6), we have the following basic inequality:

$$\begin{aligned} &S_{N,h}(\hat{w}, m_1) + \lambda_1 \|u^*\|^2_{\mathcal{H}} \|m_1\|^2_N + \lambda_2 R_{N,h}(\hat{w}) \|m_1\|^2_N \\ &\le S_{N,h}(w^*, u^*) \|m_1\|^2_N + \lambda_1 \|m_1\|^2_{\mathcal{H}} + \lambda_2 R_{N,h}(w^*) \|m_1\|^2_N. \end{aligned} \tag{E.18}$$

From Lemmas B.5 and B.6, we have $R_{N,h}(w^*) = \mathcal{O}_{\mathrm{p}}(h^{-d_1})$ and

$$S_{N,h}(w^*, u^*) = \mathcal{O}_{\mathrm{p}} \left\{ N^{-1} \|u^*\|_{\mathcal{H}}^{2-\frac{2-\alpha}{p}} h^{\left(-1-\frac{2-\alpha}{2p}\right)d_1} + \nu_{N,h}^2 h^{-2d_1} \|u^*\|^2_{\mathcal{H}} \right\}$$

for all $p \ge 1$.

We now compare different scenarios of (E.18).

Case 1: Suppose that $S_{N,h}(w^*, u^*)\|m\|^2_N$ is the largest in the right-hand side of (E.18).

If $\|m\|_N \ne 0$, we have

$$\lambda_1 \|u^*\|^2_{\mathcal{H}} \le \mathcal{O}_{\mathrm{p}} \left\{ N^{-1} \|u^*\|_{\mathcal{H}}^{2-(2-\alpha)/p} h^{-1-\frac{2-\alpha}{2p}} \right\} + \mathcal{O}_{\mathrm{p}} \left(\nu_{N,h}^2 h^{-2} \|u^*\|^2_{\mathcal{H}} \right).$$

By Assumptions 3 and 7, we can see that

$$\begin{aligned}\nu_{N,h}^2 h^{-2} &= N^{-\frac{4}{2+\alpha}} h^{\left(\frac{2\alpha}{2+\alpha}-2\right)d_1} + N^{-1-\frac{2-\alpha}{p(2+\alpha)}} h^{\left(1-\frac{2-\alpha}{2p}-2\right)d_1} \\ &= (N^{-1}h^{-d_1})^{\frac{4}{2+\alpha}} + (N^{-1}h^{-d_1})(h^{-\frac{d_1}{2}} N^{-\frac{1}{2+\alpha}})^{\frac{2-\alpha}{p}} \\ &= \mathcal{O}(N^{-1}h^{-d_1}) = o(\lambda_1).\end{aligned}$$

Thus it suffices to consider $\lambda_1\|u^*\|_{\mathcal{H}}^2 \le \mathcal{O}_{\mathrm{p}}\{N^{-1}\|u^*\|_{\mathcal{H}}^{2-(2-\alpha)/p} h^{\left(-1+\frac{\alpha-2}{2p}\right)d_1}\}$. Then we have

$$\|u^*\|_{\mathcal{H}} \le \lambda_1^{-\frac{p}{(2-\alpha)}} \mathcal{O}_{\mathrm{p}}\left\{N^{-\frac{p}{(2-\alpha)}} h^{\left(-\frac{p}{(2-\alpha)}-\frac{1}{2}\right)d_1}\right\},$$

and

$$S_{N,h}(\hat{w}, m) \le \lambda_1^{\frac{-2p+(2-\alpha)}{(2-\alpha)}} \mathcal{O}_{\mathrm{p}}\left(N^{\frac{-2p}{(2-\alpha)}} h^{\left(\frac{-2p}{(2-\alpha)}-1\right)d_1}\right) \|m\|_N^2.$$

If $\|m\|_N = 0$ we have $S_{N,h}(\hat{w}, m) = 0 \le \lambda_1^{\frac{-2p+(2-\alpha)}{(2-\alpha)}} \mathcal{O}_{\mathrm{p}}\left(N^{\frac{-2p}{(2-\alpha)}} h^{\left(\frac{-2p}{(2-\alpha)}-1\right)d_1}\right) \|m\|_N^2$.

Case 2: Suppose that $\lambda_1\|m\|_{\mathcal{H}}^2$ is the largest in right-hand side of (E.18). Then we have $S_{N,h}(\hat{w}, m) \le 3\lambda_1\|m\|_{\mathcal{H}}^2 = \mathcal{O}_{\mathrm{p}}(\lambda_1)\|m\|_{\mathcal{H}}^2$.

Case 3: Suppose that $\lambda_2 R_{N,h}(w^*)$ is the largest in right-hand side of (E.18). Then we have $S_{N,h}(\hat{w}, m) \le 3\lambda_2\, \mathcal{O}_{\mathrm{p}}(h^{-d_1})\|m\|_N^2 = \mathcal{O}_{\mathrm{p}}(\lambda_2 h^{-d_1})\|m\|_N^2$.

Combining these cases, we have

$$\begin{aligned}S_{N,h}(\hat{w}, m_1) = \max\Bigg\{ &\min\left\{\lambda_1^{\frac{-2p+(2-\alpha)}{(2-\alpha)}} \mathcal{O}_{\mathrm{p}}\left(N^{\frac{-2p}{(2-\alpha)}} h^{\left(\frac{-2p}{(2-\alpha)}-1\right)d_1}\right)\|m_1\|_N^2 : p \ge 1\right\}, \\ &\mathcal{O}_{\mathrm{p}}(\lambda_1)\|m_1\|_{\mathcal{H}}^2, \mathcal{O}_{\mathrm{p}}(\lambda_2 h^{-d_1})\|m_1\|_N^2\Bigg\}. \end{aligned} \tag{E.19}$$

Next, we compare the first two components of (E.19). We can see that as long as

$$\frac{2p}{2-\alpha}\log(\lambda_1^{-1}N^{-1}h^{-d_1}) \le \log h^{d_1},$$

the second component is dominant. Note that $\log h^{d_1} < 0$ as $h \to 0$. Because of the condition that $\lambda_1^{-1} = o(Nh^{d_1})$, the inequality is valid as long as $p \ge \frac{2-\alpha}{2}\frac{\log h^{d_1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}$. So we can pick any $p \ge \max\{1, \frac{2-\alpha}{2}\frac{\log h^{d_1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}\}$ to have the best order $\mathcal{O}_{\mathrm{p}}(\lambda_1)(\|m\|_{\mathcal{H}}^2 + \|m\|_N^2)$.

Then, we compare the first and the third components of (E.19). Similar to the previous analysis, as long as

$$\frac{2p}{2-\alpha}\log(\lambda_1^{-1}N^{-1}h^{-d_1}) \le \log(\lambda_2\lambda_1^{-1}),$$

the third component is dominant. Due to the condition that $\lambda_1^{-1} = o(Nh^{d_1})$, the inequality is valid if $p \ge \frac{2-\alpha}{2}\frac{\log \lambda_2\lambda_1^{-1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}$. So we can pick any $p \ge \max\{1, \frac{2-\alpha}{2}\frac{\log \lambda_2\lambda_1^{-1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}\}$ to have the best order $\mathcal{O}_{\mathrm{p}}(\lambda_2 h^{-d_1})\|m\|_N^2$.

Finally, we conclude that

$$S_{N,h}(\hat{w}, m_1) = \mathcal{O}_{\mathrm{p}}(\lambda_1 \|m_1\|_N^2 + \lambda_1 \|m_1\|_{\mathcal{H}}^2 + \lambda_2 h^{-d_1} \|m_1\|_N^2).$$

Moreover, further suppose that $\lambda_2^{-1} = \mathcal{O}(\lambda_1^{-1} h^{-d_1})$. From (E.18), by replacing m with a constant function and applying the similar analysis as above, we can conclude that $R_{N,h}(\hat{w}) = \mathcal{O}_{\mathrm{p}}(h^{-d_1})$. □

E.2. Proof of Theorem 5.2

Proof. First, we have

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i Y_i K_h\left(V_i, \cdot\right) - \mu_1 \right\|_2 \\
\leq &\left\| \frac{1}{N} \sum_{i=1}^{N} (T_i \hat{w}_i - 1) K_h\left(V_i, \cdot\right) m(X_i) \right\|_2 + \left\| \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i K_h(V_i, \cdot) \varepsilon_i \right\|_2 \qquad \text{(E.20)} \\
+ &\left\| \frac{1}{N} \sum_{i=1}^{N} m(X_i) K_h(V_i, \cdot) - \mu_1 \right\|_2 . \qquad \text{(E.21)}
\end{aligned}$$

Since $\|m_1\|_2 \leq b\|m_1\|_{\mathcal{H}} < \infty$ and $\|m_1\|_N = \|m_1\|_2 + o_{\mathrm{p}}(1)$, we have

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{i=1}^{N} (T_i \hat{w}_i - 1) K_h\left(V_i, \cdot\right) m_1(X_i) \right\|_2 = \{S_{N,h}(\hat{w}, m_1)\}^{1/2} \\
= &\mathcal{O}_{\mathrm{p}}(\lambda_1^{1/2} \|m_1\|_N + \lambda_1^{1/2} \|m_1\|_{\mathcal{H}} + \lambda_2^{1/2} h^{-d_1/2} \|m_1\|_N) \\
= &\mathcal{O}_{\mathrm{p}}(\lambda_1^{1/2} \|m_1\|_{\mathcal{H}} + \lambda_2^{1/2} h^{-d_1/2} \|m_1\|_2) + o_{\mathrm{p}}(\lambda_1^{1/2} + \lambda_2^{1/2} h^{-d_1/2})
\end{aligned}$$

due to Theorem 5.1 and the conditions of λ_1 and λ_2.

For the second term in (E.20), we have $\mathbb{E}(\varepsilon_i \mid T_i, \hat{w}_i, X_i, i = 1, \ldots, N) = 0$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left\{ \left\| \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i K_h(V_i, \cdot) \varepsilon_i \right\|_2^2 \mid T_i, \hat{w}_i, X_i, i = 1, \ldots, N \right\} \\
= &\int_0^1 \mathbb{E}\left\{ \left[\frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i K_h(V_i, v) \varepsilon_i \right]^2 \mid T_i, \hat{w}_i, X_i, i = 1, \ldots, N \right\} dv \\
= &\int_0^1 \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}\left\{ T_i \hat{w}_i^2 K_h^2(V_i, v) \epsilon_i^2 \mid T_i, \hat{w}_i, X_i, i = 1, \ldots, N \right\} dv \\
\leq &\frac{\sigma_0^2}{N} \int_0^1 \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i^2 K_h^2(V_i, v) dv = \frac{\sigma_0^2}{N} R_{N,h}(\hat{w}).
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left\{\frac{\left\|\frac{1}{N}\sum_{i=1}^{N} T_i\hat{w}_i K_h(V_i,\cdot)\varepsilon_i\right\|_2^2}{R_{N,h}(\hat{w})} \mid T_i, X_i, i=1,\ldots,N\right\} \le \frac{\sigma_0^2}{N},$$

$$\mathbb{E}\left\{\frac{\left\|\frac{1}{N}\sum_{i=1}^{N} T_i\hat{w}_i K_h(V_i,\cdot)\varepsilon_i\right\|_2^2}{R_{N,h}(\hat{w})}\right\} \le \frac{\sigma_0^2}{N},$$

$$\frac{\left\|\frac{1}{N}\sum_{i=1}^{N} T_i\hat{w}_i K_h(V_i,\cdot)\varepsilon_i\right\|_2^2}{R_{N,h}(\hat{w})} = \sigma_0^2\,\mathcal{O}_{\mathrm{p}}(\frac{1}{N}).$$

From the condition of λ_2, and the result from Theorem 5.1 that $R_{N,h}(\hat{w}) = \mathcal{O}_{\mathrm{p}}(h^{-d_1})$, we have

$$\left\|\frac{1}{N}\sum_{i=1}^{N} T_i\hat{w}_i K_h(V_i,\cdot)\varepsilon_i\right\|_2 = \mathcal{O}_{\mathrm{p}}(N^{-1/2}h^{-d_1/2}).$$

As for (E.21), it has a form of a typical Nadaraya–Watson estimator. By Theorem 5.44 in [44], we have

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N} m(X_i)K_h(V_i-\cdot) - \mu_1\right\|_2^2 = \mathcal{O}(N^{-1}h^{-d_1}).$$

Therefore, we have

$$\left\|\frac{1}{N}\sum_{i=1}^{N} m(X_i)K_h(V_i-\cdot) - \mu_1\right\|_2^2 = \mathcal{O}_{\mathrm{p}}(N^{-1}h^{-d_1}).$$

Overall, conclusion follows. □

E.3. Proof outline of Theorem 5.3

To obtain the rate, the entropy bound in Lemma B.1 needs to be modified to the bigger function class $\mathcal{F}_h := \{f : f(\check{x}) = u(\check{x})K(\frac{\check{v}-v}{h}), u \in \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \le 1\}, v \in [0,1]^{d_1}\}$. This can be done by combining the entropy bound for $\{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \le 1\}$ and Assumption 5(b). One can show that

$$H(\delta, \mathcal{F}_h, \|\cdot\|_N) \begin{cases} = 0 & \text{if } \delta > 2b\sigma_{\mathcal{K}_h,N} \\ \le A\sigma_{\mathcal{K}_h,N}^{\alpha}\delta^{-\alpha} + \log(A_1\epsilon^{-\nu_1}) & \text{otherwise} \end{cases}.$$

Then by adopting this entropy bound, the results in Lemma B.3 will be modified to that $\forall f \in \mathcal{F}_h$,

$$\frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right| \le t\left\{N^{-\frac{1}{2}}\|u\|_2^{\frac{2-\alpha}{2p}} h^{d_1\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)}\left(\log\frac{1}{h}\right)^{1/2} + N^{-\frac{2}{2+\alpha}}h^{\frac{d_1\alpha}{2+\alpha}}\log\frac{1}{h}\right\},$$

for any $t \geq c_1$, and $p \geq 1$ with probability at least $1 - c\exp(-c_6 t)$. Then the remaining argument is similar to those in the proof of Theorems 5.1 and 5.2.

E.4. Proof of Theorem 5.4

Proof. Following the same proof structure of Theorem 5.1, by replacing m with a constant function z of value 1, we have

$$S_{N,h}(\hat{w}, z) + \lambda_1 \|u^*\|_{\mathcal{H}}^2 + \lambda_2 R_{N,h}(\hat{w}) \leq S_{N,h}(w^*, u^*) + \lambda_1 \|z\|_{\mathcal{H}}^2 + \lambda_2 R_{N,h}(w^*).$$

By the condition of λ_1 such that $\lambda_1^{-1} = \mathcal{O}_{\mathrm{p}}(N^{-1}h^{-d_1})$, we have $R_{N,h}(\hat{w}) = \mathcal{O}_{\mathrm{p}}(\lambda_2^{-1}\lambda_1 + h^{-d_1})$. Since $\lambda_2^{-1}\lambda_1 = \mathcal{O}(h^{-d_1})$,

$$R_{N,h}(\hat{w}) = \mathcal{O}_{\mathrm{p}}(h^{-d_1}).$$

Again, following the same proof structure of Theorem 5.1, note that $\hat{e} \in \mathcal{H}$, by replacing m with $\hat{e}$, we have

$$\begin{aligned} S_{N,h}(\hat{w}, \hat{e}) + \lambda_1 \|u^*\|_{\mathcal{H}}^2 \|\hat{e}\|_N^2 + \lambda_2 R_{N,h}(\hat{w}) \|\hat{e}\|_N^2 \leq \\ S_{N,h}(w^*, u^*) \|\hat{e}\|_N^2 + \lambda_1 \|\hat{e}\|_{\mathcal{H}}^2 + \lambda_2 R_{N,h}(w^*) \|\hat{e}\|_N^2. \end{aligned}$$

By the condition of λ_1 such that $\lambda_1^{-1} = \mathcal{O}_{\mathrm{p}}(N^{-1}h^{-d_1})$, we can obtain

$$S_{N,h}(\hat{w}, e) = \mathcal{O}_{\mathrm{p}}(\lambda_1 \|e\|_N^2 + \lambda_1 \|e\|_{\mathcal{H}}^2 + \lambda_2 h^{-d_1} \|e\|_N^2).$$

Therefore,

$$\begin{aligned} &\left\| \frac{1}{N} \sum_{i=1}^N \tilde{K}_h(V_i, \cdot)\hat{m}(X_i) + \frac{1}{N} \sum_{i=1}^N T_i \hat{w}_i \tilde{K}_h(V_i, \cdot)\{Y_i - \hat{m}(X_i)\} - \mu_1 \right\|_2 \\ \leq &\left\| \frac{1}{N} \sum_{i=1}^N (T_i \hat{w}_i - 1) K_h(V_i, \cdot)\, e(X_i) \right\|_2 + \left\| \frac{1}{N} \sum_{i=1}^N T_i \hat{w}_i K_h(V_i, \cdot)\varepsilon_i \right\|_2 \\ &+ \left\| \frac{1}{N} \sum_{i=1}^N m(X_i) K_h(V_i, \cdot) - \mu_1 \right\|_2 \\ \leq &\{S_{N,h}(\hat{w}, e)\}^{1/2} + \mathcal{O}_{\mathrm{p}}(N^{-1/2}) R_{N,h}^{1/2}(\hat{w}) + \mathcal{O}_{\mathrm{p}}(N^{-1/2}h^{-d_1/2}) \\ \leq &\mathcal{O}_{\mathrm{p}}(\lambda_1^{1/2}\|e\|_N + \lambda_1^{1/2}\|e\|_{\mathcal{H}} + \lambda_2^{1/2} h^{-d_1/2}\|e\|_N) + \mathcal{O}_{\mathrm{p}}(N^{-1/2}h^{-d_1/2}) \\ \leq &\mathcal{O}_{\mathrm{p}}(N^{-1/2}h^{-d_1/2} + \lambda_1^{1/2}\|e\|_N + \lambda_1^{1/2}\|e\|_{\mathcal{H}} + \lambda_2^{1/2} h^{-d_1/2}\|e\|_N). \end{aligned}$$

□

References

[1] Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics* **21** 489–519. MR2236852

[2] Abrevaya, J. and Dahl, C. M. (2008). The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics* **26** 379–397. MR2459341
[3] Abrevaya, J., Hsu, Y.-C. and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* **33** 485–505. MR3416596
[4] Almond, D., Chay, K. Y. and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics* **120** 1031–1083.
[5] Athey, S., Imbens, G. W. and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 597–623. MR3849336
[6] Bartlett, P. L., Bousquet, O., Mendelson, S. et al. (2005). Local rademacher complexities. *The Annals of Statistics* **33** 1497–1537. MR2166554
[7] Birman, M. S. and Solomyak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes W_pˆα. *Matematicheskii Sbornik* **115** 331–355. MR0217487
[8] Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *arXiv preprint arXiv:1906.00198*.
[9] Chan, K. C. G., Yam, S. C. P. and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **78** 673. MR3506798
[10] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21** C1-C68. MR3769544
[11] Chernozhukov, V., Newey, W., Robins, J. and Singh, R. (2018). Double/de-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*.
[12] Einmahl, U., Mason, D. M. et al. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics* **33** 1380–1403. MR2195639
[13] Fan, Q., Hsu, Y.-C., Lieli, R. P. and Zhang, Y. (2020). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics* **just-accepted** 1–39. MR4356575
[14] Fox, S. H., Koepsell, T. D. and Daling, J. R. (1994). Birth weight and smoking during pregnancy-effect modification by maternal age. *American Journal of Epidemiology* **139** 1008–1015.
[15] Gu, C. (2013). *Smoothing spline ANOVA models* **297**. Springer Science & Business Media. MR3025869
[16] Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis* 25–46.

[17] Harder, R. L. and Desmarais, R. N. (1972). Interpolation using surface splines. *Journal of aircraft* **9** 189–191.
[18] Härdle, W. K. et al. (1991). *Smoothing techniques: with implementation in S*. Springer Science & Business Media. MR1140190
[19] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20** 217–240. MR2816546
[20] Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 243–263. MR3153941
[21] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics* **86** 4–29.
[22] Kallus, N. (2020). Generalized Optimal Matching Methods for Causal Inference. *Journal of Machine Learning Research* **21** 1–54. MR4095341
[23] Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 523–539. MR2420458
[24] Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.
[25] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008* **2033**. Springer Science & Business Media. MR2829871
[26] Kramer, M. S. (1987). Intrauterine growth and gestational duration determinants. *Pediatrics* **80** 502–511.
[27] Lee, S., Okui, R. and Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics* **32** 1207–1225. MR3734484
[28] Lin, Y. et al. (2000). Tensor product space ANOVA models. *The Annals of Statistics* **28** 734–755. MR1792785
[29] Mack, Y.-p. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61** 405–415. MR0679685
[30] Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*. MR4259133
[31] Oprescu, M., Syrgkanis, V. and Wu, Z. S. (2019). Orthogonal random forest for causal inference. In *International Conference on Machine Learning* 4932–4941. PMLR.
[32] Pearce, N. D. and Wand, M. P. (2006). Penalized splines and reproducing kernel methods. *The American Statistician* **60** 233–240. MR2246756
[33] Qin, J. and Zhang, B. (2007). 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** 101–122. MR2301502
[34] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the

propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

[35] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90** 1257–1270. MR1379468

[36] Semenova, V. and Chernozhukov, V. (2017). Estimation and Inference about Conditional Average Treatment Effect and Other Structural Functions. *arXiv* arXiv–1702.

[37] Semenova, V. and Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*. MR4281225

[38] Van der Vaart, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge university press. MR1652247

[39] Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113** 1228–1242. MR3862353

[40] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM. MR1045442

[41] Walker, M., Tekin, E. and Wallace, S. (2007). Teen smoking and birth outcomes Technical Report, National Bureau of Economic Research.

[42] Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. Crc Press. MR1319818

[43] Wang, Y. and Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142

[44] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science & Business Media. MR2172729

[45] Wong, R. K. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105** 199–213. MR3768874

[46] Zhao, Q. et al. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics* **47** 965–993. MR3909957

[47] Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning* 459–474. Springer. MR2867139

[48] Zimmert, M. and Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*.

[49] Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association* **110** 910–922. MR3420672