Estimation of Partially Conditional Average Treatment Effect by Hybrid Kernel-covariate Balancing

Jiayi Wang[1], Raymond K. W. Wong[1], Shu Yang[2], and Kwun Chuen Gary Chan[3]

[1]*Department of Statistics, Texas A&M University*
[2]*Department of Statistics, North Carolina State University*
[3]*Department of Biostatistics, University of Washington*

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 hybrid 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 subset of interested variables. In addition, we present an augmented version of our estimator which can incorporate estimations 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 propensity score function or the need of 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: Augmented weighting estimator; Causal inference; Reproducing kernel Hilbert space; Treatment effect heterogeneity

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 $\gamma(x)$ (Nie and Wager, 2017; Wager and Athey, 2018; Kennedy, 2020). 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 but not mother's educational attainment, although this variable can be a confounder. 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.

When V contains discrete covariates, one can divide the whole sample into different groups by constricting the same values of discrete covariates of V in the same group. Then, as long as there are enough samples in such stratum, $\tau(v)$ can be obtained by estimating the PCATE over the remaining continuous covariates in V separately for every stratum. Therefore, for simplicity, we focus on the setups with continuous V (Abrevaya et al., 2015; Lee et al., 2017; Fan et al., 2020; Zimmert and Lechner, 2019; Semenova and Chernozhukov, 2020) while keeping in mind that the proposed method can be used to handle V that consists of continuous and discrete variables. The 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, or even machine learning models. This step serves to adjust for confounding biases. In the second step, existing methods typically adopt nonparametric regression over V using the adjusted responses obtained from the first step. However, these methods suffer from many 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, Abrevaya et al. (2015), Zimmert and Lechner (2019), Fan et al. (2020) and Semenova and Chernozhukov (2020) specify restrictive requirements for the convergence rate of the estimators of the nuisance parameters. Detailed discussions are provided in 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 the weighting step. In particular, we generalize the idea of covariate balancing weighting in the average treatment effect (ATE) estimation literature (Qin and Zhang, 2007; Hainmueller,

2012; Imai and Ratkovic, 2014; Zubizarreta, 2015; Wong and Chan, 2018). 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, Wahba, 1990), a fairly flexible class of functions of X. 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. The construction of the proposed weights specifically involves two kernels — the reproducing kernel of the RKHS and the kernel of the kernel smoothing — and the goal of these weights can be understood as to balance covariate functions generated by the hybrid of these two kernels. 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. Invoking the well-known representer theorem, a finite-dimensional representation form of optimization objective can be derived and it can be solved by a gradient-based algorithm. Asymptotic properties of the proposed estimator are derived under the complex dependency structure of weights and kernel smoothing. 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 paper is organized as follows. Section 2 provides the basic setup for the CATE estimation. Section 3 introduces our proposed CFB weighting estimator, together with the computation techniques. Section 4 introduces an augmented version of our proposed estimator. In Section 5, the asymptotic properties of the proposed estimators are developed. A simulation study and a real data application are presented in Sections 6 and 7, respectively.

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 (Rosenbaum and Rubin, 1983).

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 (Nie and Wager, 2017; Wager and Athey, 2018; Kennedy, 2020; Semenova and Chernozhukov, 2020) focus on estimating the CATE given the full set of X, i.e., $\gamma(x) := \mathbb{E}\{Y(1) - Y(0) \mid X = x\}$, $x \in \mathcal{X}$, which we refer to as the FCATE. 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 a subset of confounders (Abrevaya et al., 2015; Zimmert and Lechner, 2019; Fan et al., 2020). 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. It is worth noting that $V = X$ is also allowed, and therefore $\gamma(x)$ can be estimated by 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 strata.

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 kernel smoother to the data $\{(V_i, Z_i), i = 1, \ldots, N\}$. 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 $\tau(v)$:

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

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. Abrevaya et al. (2015) propose 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. 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 proper 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., Hainmueller, 2012; Imai and

Ratkovic, 2014; Zubizarreta, 2015; Chan et al., 2016; Wong and Chan, 2018; Zhao et al., 2019; Kallus, 2020; Wang and Zubizarreta, 2020).

3 Covariate function balancing weighting for PCATE estimation

3.1 Motivation

To motivate the proposed estimator, suppose we are given the covariate balancing weights $\{\hat{w}_i : i = 1, \ldots, N\}$. We express the adjusted response as

$$Z_i = \hat{w}_i T_i Y_i - \hat{w}_i (1 - T_i) Y_i, \quad i = 1, \ldots, N. \tag{2}$$

Combining (1) and (2), the estimator of $\tau(v)$ is

$$\hat{\tau}(v) = \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i \tilde{K}_h(V_i, v) Y_i - \frac{1}{N} \sum_{i=1}^{N} (1 - T_i) \hat{w}_i \tilde{K}_h(V_i, v) Y_i. \tag{3}$$

One can see that the estimator (3) 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 discussion applies to the second term and the estimation of weights in the control group.

We assume $Y_i(1) = m_1(X_i) + \varepsilon_i$ such that the ε_i's are independent random errors with $\mathbb{E}(\varepsilon_i \mid X_i) = 0$ and $\mathbb{E}(\epsilon_i^2 \mid X_i) \leq \sigma_0^2 < \infty$. Focusing on the first term of (3), we obtain the following decomposition

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i \tilde{K}_h(V_i, v) Y_i &= \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i \tilde{K}_h(V_i, v) m_1(X_i) + \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i \tilde{K}_h(V_i, v) \varepsilon_i \\
&= \frac{1}{N} \sum_{i=1}^{N} (T_i \hat{w}_i - 1) \tilde{K}_h(V_i, v) m_1(X_i) + \frac{1}{N} \sum_{i=1}^{N} T_i \hat{w}_i \tilde{K}_h(V_i, v) \varepsilon_i \\
&\quad + \left[\frac{1}{N} \sum_{i=1}^{N} \tilde{K}_h(V_i, v) m_1(X_i) - \mu_1(v) \right] + \mu_1(v).
\end{aligned} \tag{4}$$

In the last equality, only the first two terms depend on the weights. 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. The 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., Hainmueller, 2012; Imai and Ratkovic, 2014; Zubizarreta, 2015; Chan et al., 2016; Wong and Chan, 2018; Zhao et al., 2019; Kallus, 2020; Wang and Zubizarreta, 2020), 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 respect to the sample size N.

3.2 Balancing via 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 $u \in \mathcal{H}$. Intuitively, from the first term of (4), we aim to find weights $w = \{w_i : T_i = 1\}$ to ensure the following function balancing criteria:

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

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

$$\|\mathcal{M}_{N,h,w}(u, \cdot)\|^2, \tag{5}$$

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:

$$S_{N,h}(w, u) = \|\mathcal{M}_{N,h,w}(u, \cdot)\|_2^2. \tag{6}$$

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 Wong and Chan (2018). Also, to overcome overfitting, we add a penalty on u with respect to $\|\cdot\|_{\mathcal{H}}$ and focus on controlling the balancing error over smoother functions. Inspired by the discussion for (4), we also introduce another penalty term

$$R_{N,h}(w) = \frac{1}{N}\sum_{i=1}^N \|T_i w_i \tilde{K}_h(V_i, \cdot)\|_2^2, \tag{7}$$

to control the variability of the weights.

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

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left[\sup_{u\in\mathcal{H}_N} \left\{ S_{N,h}(w, u) - \lambda_1 \|u\|_{\mathcal{H}}^2 \right\} + \lambda_2 R_{N,h}(w) \right], \tag{8}$$

where λ_1 and λ_2 are tuning parameters ($\lambda_1 > 0$ and $\lambda_2 > 0$). Note that (8) 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 also Section S1.1 in the supplementary material. Therefore, by the definition of $\mathcal{M}_{N,h,w}$, the weights are determined by achieving balance of the covariate functions generated by the hybrid of the two reproducing kernel κ and the smoothing kernel K.

Remark 2. Wong and Chan (2018) adopt a similar optimization form as in (8) to obtain weights. The key difference between their estimator and ours is the choice of balancing error tailored to the target quantity. In Wong and Chan (2018), 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 Wong and Chan (2018) as the limit of $S_{N,h}$ as $h \to \infty$. For finite h, two fundamental difficulties emerge that do not exist in Wong and Chan (2018). 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 (6). 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 Wong and Chan (2018) 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

Applying the standard representer theory, (8) can be reformulated as

$$\hat{w} = \underset{w \geq 1}{\text{argmin}} \left[\sigma_1 \left\{ \frac{1}{N} P^\intercal \text{diag}(T \circ w - J) G_h \text{diag}(T \circ w - J) P - N\lambda_1 D^{-1} \right\} + \lambda_2 R_{N,h}(w) \right], \tag{9}$$

where $\circ$ is the element-wise product of two vectors, $J = (1, 1, \ldots, 1)^\intercal$, $\sigma_1(A)$ represents the maximum eigenvalue of a symmetric matrix A, $P \in \mathbb{R}^{N \times r}$ consists of the singular vectors of 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 $M = PDP^\intercal$, and

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

The detailed derivation can be found in Section S1.1 in the supplementary material.

As for the computation, Lemma 1, whose proof can be found in Section S1.2 in the supplementary material, indicates that the underlying optimization is convex.

Lemma 1. *The optimization* (8) *is convex.*

Therefore, generic convex optimization algorithms are applicable. We note that the corresponding gradient has a closed-form expression[1]. Thus, gradient based algorithms can be applied efficiently to solve this problem.

Next we discuss several practical strategies to speed up the optimization. When optimizing (9), we need to compute the dominant eigenpair of an $r \times r$ matrix (for computing the gradient). Since common choices of the reproducing kernel κ are smooth, the corresponding Gram matrix M can be approximated well by a low-rank matrix. When N is large, to facilitate computation, one can use an M with r much smaller than N, such that the eigen decomposition of gram matrix M approximately holds. This would significantly reduce the burden of computing the dominant eigenpair of the $r \times r$ matrix. Although the form of G_h may seem complicated, this does not change with w. Therefore, for each h, we can precompute G_h once at the beginning of an algorithm for the optimization (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 integrations 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 (Harder and Desmarais, 1972) can be implemented to approximate unevaluated integrals in G_h to ease the computation burden.

4 Augmented estimator

Inspired by the augmented inverse propensity weighting (AIPW) estimators in the ATE literature, we also propose an augmented estimator that directly adjusts for the outcome models $m_1(\cdot)$ and $m_0(\cdot)$.

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 can then perform augmentation and obtain the adjusted response Z_i in (2) as

$$Z_i = w_i T_i\{Y_i - \hat{m}_1(X_i)\} + \hat{m}_1(X_i) - \left[w_i(1 - T_i)\left\{Y_i - \hat{m}_0(X_i)\right\} + \hat{m}_0(X_i)\right]. \tag{10}$$

Correspondingly, the decomposition in (4) becomes

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

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

$$+\left\{\frac{1}{N}\sum_{i=1}^{N}\tilde{K}_h(V_i,v)m_1(X_i)-\mu_1(v)\right\}+\mu_1(v).$$

Now, our goal is to control the difference between $N^{-1}\sum_{i=1}^{N}T_i\hat{w}_i\tilde{K}_h(V_i,v)\{m_1(X_i)-\hat{m}_1(X_i)\}$ 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 $S_{N,h}(\hat{w},m_1-\hat{m}_1):=\|\sum_{i=1}^{N}(T_i\hat{w}_i-1)\tilde{K}_h(V_i,\cdot)\{m_1(X_i)-\hat{m}_1(X_i)/N\}\|_2^2$ 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 $\|N^{-1}\sum_{i=1}^{N}\tilde{K}_h(V_i,\cdot)m_1(X_i)-\mu_1(\cdot)\|_2^2$ 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 to the tuning parameter selection.

Unlike other AIPW-type estimators (Lee et al., 2017; Fan et al., 2020; Zimmert and Lechner, 2019; Semenova and Chernozhukov, 2020) 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 also Remark 7. We defer the theoretical comparison between our estimator and the existing AIPW-type estimators in Remark 6 in Section 5.

Last, we note that there are existing work using weights to balance the residuals (e.g. Athey et al., 2016; Wong and Chan, 2018), which is similar to what we consider here for the proposed augmented estimator. These estimators are designed for ATE estimation and the balancing weights cannot be directly adopted here 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 $\{u\in\mathcal{H}:\|u\|_{\mathcal{H}}\le 1\}$ 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 $\|u\|_{\mathcal{W}^\ell}=\sqrt{\sum_{|\beta|\le\ell}\|D^\beta u\|_2^2}$, 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* $m(x) \in \mathcal{H}$.

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* $\{m(X), V\}$ *and the conditional expectation* $\mathbb{E}\{m(X) \mid V = v\}$ *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., Mack and Silverman, 1982; Einmahl et al., 2005; Wasserman, 2006) 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 (Van der Vaart, 2000).

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 $A_N \leq M B_N$ 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 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 1 specifies the control of the balancing error and the weight variability. They can be used to derive the convergence rate of the proposed estimator in the following theorem.

Theorem 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* $h^2 = o((N^{-1} h^{-d_1})^{1/2})$,

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

The proof can be found in Section S2.1 and S2.2 in the supplementary material. Since we require $\lambda_1^{-1} = \mathcal{O}(Nh^{d_1})$, the best convergence rate that we can achieve in Theorem 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 4 below, we show that this gap can be closed by using the proposed augmented estimator.

Remark 3. Abrevaya et al. (2015) adopt 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 (6). 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 L_2 convergence rate (Theorem 2), we can obtain the uniform convergence rate of this estimator in the following theorem.

Theorem 3. *Suppose Assumptions 1–9 hold, Let $\tilde{w}$ be the solution to (8) 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 $h^2 = o\{(N^{-1}h^{-d_1}\log(1/h))^{1/2}\}$,*

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

We provide the proof outline in Section S2.3 in the supplementary material.

Different from Theorem 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 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

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

Remark 4. In Theorem 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 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 (Kennedy, 2020; Nie and Wager, 2017). 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 (Lee et al., 2017; Semenova and Chernozhukov, 2017; Zimmert and Lechner, 2019; Fan et al., 2020) that adopts augmentation, estimations of both propensity score and outcome mean functions, referred to as nuisance parameters in below, are required. Lee et al. (2017) adopt 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. Semenova and Chernozhukov (2017); Zimmert and Lechner (2019); Fan et al. (2020) adopt nonparametric nusiance 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. Fan et al. (2020) require both nuisance parameters to be estimated consistently with respect to L_∞ norm. While Semenova and Chernozhukov (2017) and Zimmert and Lechner (2019) 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 $N^{-1/2}$. Unlike these existing estimators, our estimators does not rely on restrictive parametric modeling nor consistency of both nuisance parameter estimation.

Remark 7. Moreover, most existing work (discussed in Remark 6) require data-splitting or cross-fitting to remove the dependence between nuisance parameter estimations and the

Table 1: Models for simulation with two specifications for each of 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	$1/(1 + \exp Z_1 + Z_2 + Z_3)$	$10 + (2t - 1)(Z_1^2 + 2Z_1 \sin(2Z_1)) + Z_2^2 + \sin(2Z_3)Z_4^2$	$2v^2 + 4v\sin(2v)$

smoothing step for estimating $\tau(\cdot)$, which is crucial in their theoretical analyses. Zheng and van der Laan (2011) first propose cross-fitting in the context of Target Maximum Likelihood Estimator and Chernozhukov et al. (2017) subsequently apply to estimating equations. This technique can be used to relax the Donsker conditions required for the class of nuisance functions. Kennedy (2020) 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

We evaluate the finite-sample properties of various estimators with sample size $N = 100$. The covariate $X \in \mathbb{R}^4$ is generated by $X_1 = Z_1$, $X_2 = Z_1^2 + Z_2$, $X_3 = \exp(Z_3/2) + Z_2$ and $X_4 = \sin(2Z_1) + Z_4$ with $Z_j \sim \text{Uniform}[-2, 2]$ 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 $\tau(\cdot)$:

a) PROPOSED: the proposed estimator using the tensor product of second-order Sobolev kernel as the reproducing kernel κ.

b) ATE_{RKHS}: the weighted estimator described in Remark 2, whose weights are estimated based on the covariate balancing criterion in Wong and Chan (2018).

c) IPW: the inverse propensity weighting estimator from Abrevaya et al. (2015) with a logistic regression model for the propensity score. In Settings 1 and 3, the propensity score model is correctly specified.

d) Augmented estimators by augmenting the estimators in a)–c) by the outcome models. We consider two outcome models: linear regression (LM) and kernel ridge regression (KRR).

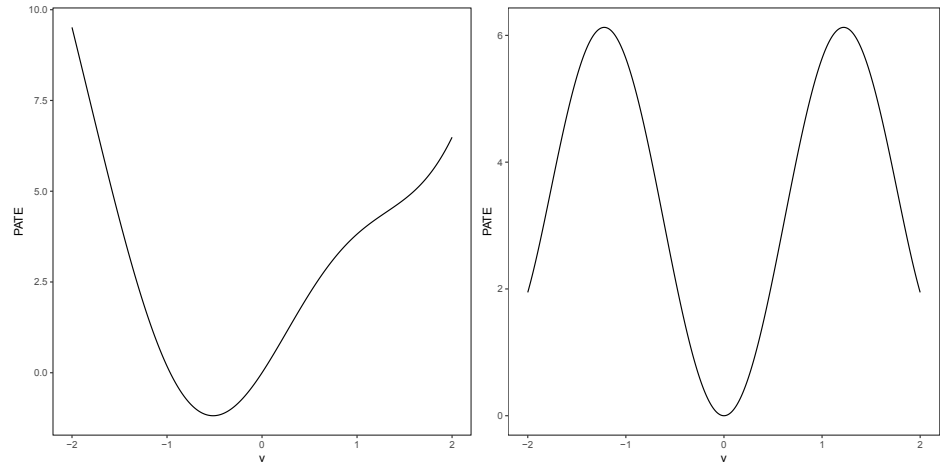

Figure 1: The target PCATEs in the simulation study: the left panel plots the PCATE in Settings 1 and 2; the right panel plots the PCATE in Settings 3 and 4.

e) REG: the estimator that uses outcome regressions. It directly smooths $\{(X_i, \hat{m}_1(X_i) - \hat{m}_0(X_i)) : i = 1, \ldots, N\}$ to estimate the PCATE, where $\hat{m}_1(X_i)$ and $\hat{m}_0(X_i)$ are estimated with outcome models considered in d).

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 (Ruppert, Sheather, and Wand, 1995; Wand and Jones, 1994; Calonico, Cattaneo, and Farrell, 2019). Throughout our analysis, $\hat{h}$ is computed via the R package "nprobust". The same bandwidth formula $\tilde{h}$ is also considered by Lee et al. (2017) and Fan et al. (2020) to estimate the CATE. For the proposed estimator, a bandwidth should be given prior to estimate the weights. We first compute the adjusted response by using weights from Wong and Chan (2018), and then obtain the bandwidth $\tilde{h}$ as the input to our proposed estimator.

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	80.212 (16.19)	25.706	40.898 (6.75)	18.838	105.509 (31.66)	31.146	49.04 (9.34)	20.989
	ATE_{RKHS}	16.136 (0.77)	9.633	9.653 (0.46)	6.139	18.458 (0.93)	10.264	11.367 (0.59)	7.257
	Proposed	4.223 (0.22)	2.725	2.232 (0.06)	1.997	4.769 (0.26)	3.229	3.214 (0.08)	3.006
LM	IPW	1.167 (0.04)	0.958	1.066 (0.03)	0.893	5.431 (1.37)	2.405	3.74 (0.78)	2.001
	ATE_{RKHS}	1.156 (0.03)	1.011	1.112 (0.03)	1.003	3.471 (0.19)	2.237	2.276 (0.06)	1.924
	Proposed	1.095 (0.03)	0.947	0.977 (0.02)	0.868	2.966 (0.17)	2.014	1.856 (0.05)	1.596
	REG	0.843 (0.03)	0.716	0.767 (0.02)	0.67	5.431 (0.18)	4.368	4.254 (0.05)	4.107
KRR	IPW	1.25 (0.04)	1.048	1.039 (0.03)	0.905	3.203 (0.19)	2.096	2.313 (0.14)	1.645
	ATE_{RKHS}	1.289 (0.04)	1.092	1.152 (0.03)	0.993	3.07 (0.13)	2.213	2.125 (0.06)	1.827
	Proposed	1.203 (0.04)	1.023	1.012 (0.03)	0.856	2.658 (0.12)	1.911	1.843 (0.05)	1.53
	REG	1.137 (0.03)	0.953	0.905 (0.02)	0.797	3.778 (0.12)	3.213	2.796 (0.06)	2.634

Table 2 shows the average integrated squared error (AISE) and median integrated squared error (MeISE) of above estimators over 500 simulated datasets. Without augmentation, Pro-

POSED 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 improvement 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 Lee et al. (2017), 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 Abrevaya et al. (2015), we include IPW: the IPW estimator with no augmentation. Following Lee et al. (2017), we include IPW(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 IPW(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 estimations in Proposed, we consider a tensor product RKHS, with the second order Sobolev space kernel for continuous covariates and the identity kernel for binary covariates.

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 (Kramer, 1987; Almond et al., 2005; Abrevaya, 2006; Abrevaya and Dahl, 2008). 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 (Fox et al., 1994; Walker et al., 2007). The REG(LM) estimator shows a linearly decreasing pattern, while the REG(KRR) estimator stops decreasing after age 30. For three weighted estimators, the effects are stable

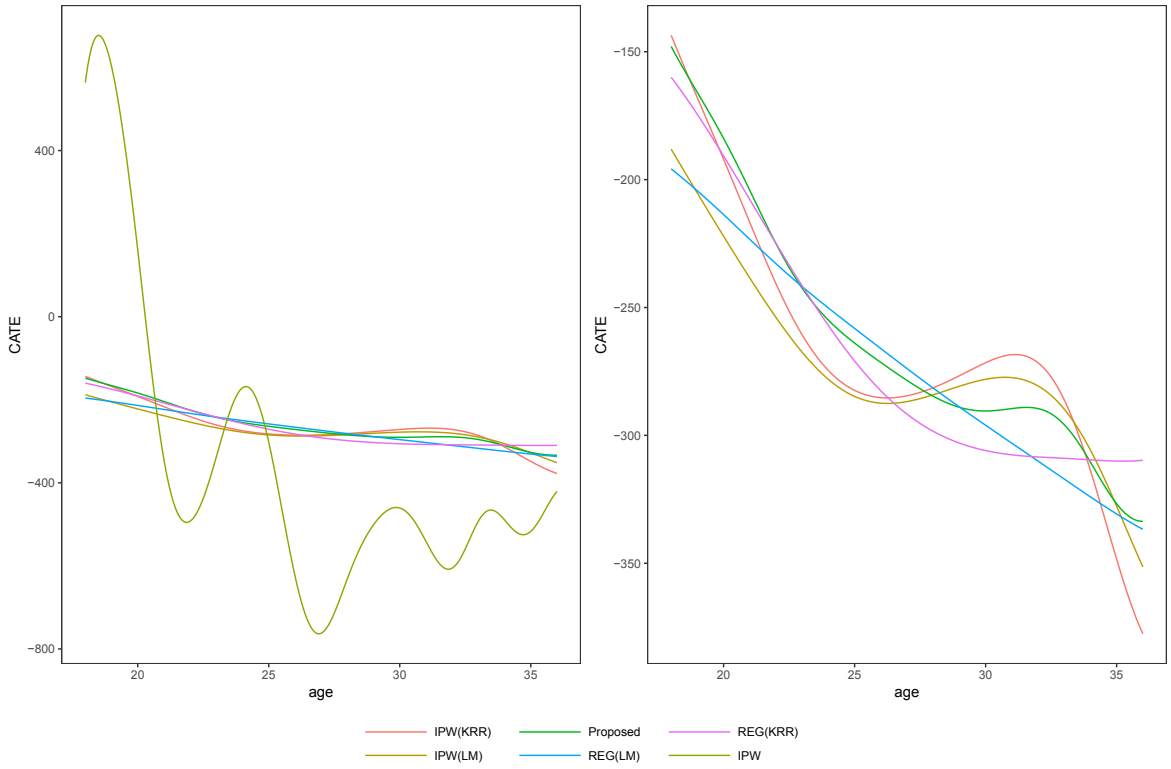

Figure 2: The estimated PCATEs of maternal smoking on birth weight as a function of mother's age: the left panel includes all estimators and the right panel excludes the IPW estimator.

around age 27 to 32, but tend to decrease quickly after age 32. Compared to IPW(LM) and IPW(KRR), Proposed does not show the increasing tendency before age 30 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 hybrid 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 proposed estimator is worth studying so that inference procedures can be developed.

Acknowledgement

The work of Raymond K. W. Wong is partially supported by the National Science Foundation (DMS-1711952 and CCF-1934904). The work of Shu Yang is partially supported by the National Institute on Aging (1R01AG066883) and the National Science Foundation (DMS-1811245). The work of Kwun Chuen Gary Chan is partially supported by the National Heart, Lung, and Blood Institute (R01HL122212) and the National Science Foundation (DMS-1711952).

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics 21*(4), 489–519.

Abrevaya, J. and C. M. Dahl (2008). The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics 26*(4), 379–397.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics 33*(4), 485–505.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *The Quarterly Journal of Economics 120*(3), 1031–1083.

Athey, S., G. W. Imbens, and S. Wager (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *arXiv preprint arXiv:1906.00198*.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (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*(3), 673.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins, et al. (2017). Double/debiased machine learning for treatment and causal parameters. Technical report.

Einmahl, U., D. M. Mason, et al. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics 33*(3), 1380–1403.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2020). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics* (just-accepted), 1–39.

Fox, S. H., T. D. Koepsell, and J. R. Daling (1994). Birth weight and smoking during pregnancy-effect modification by maternal age. *American Journal of Epidemiology 139*(10), 1008–1015.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 25–46.

Harder, R. L. and R. N. Desmarais (1972). Interpolation using surface splines. *Journal of aircraft 9*(2), 189–191.

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

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research 21*(62), 1–54.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kramer, M. S. (1987). Intrauterine growth and gestational duration determinants. *Pediatrics 80*(4), 502–511.

Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*(7), 1207–1225.

Mack, Y.-p. and B. W. Silverman (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 61*(3), 405–415.

Nie, X. and S. Wager (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

Qin, J. and B. Zhang (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*(1), 101–122.

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

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*(432), 1257–1270.

Semenova, V. and V. Chernozhukov (2017). Estimation and inference about conditional average treatment effect and other structural functions. *arXiv*, arXiv–1702.

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

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3. Cambridge university press.

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

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

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

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

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

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

Wong, R. K. and K. C. G. Chan (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika 105*(1), 199–213.

Zhao, Q. et al. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics 47*(2), 965–993.

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

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

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

Supplementary material for "Estimation of Partially Conditional Average Treatment Effect by Hybrid Kernel-covariate Balancing"

Jiayi Wang*, Raymond K. W. Wong†, Shu Yang‡ and Kwun Chuen Gary Chan§

S1 Computation

S1.1 Reparametrization

To solve (9), we focus on the inner optimization of (8): $\sup_{u\in\mathcal{H}_N}\{S_{N,h}(w,u)-\lambda_1\|u\|_{\mathcal{H}}^2\}$, which is equivalent to

$$\sup_{u\in\mathcal{H}}\left\{\frac{S_{N,h}(w,u)}{\|u\|_N}-\lambda_1\frac{\|u\|_{\mathcal{H}}^2}{\|u\|_N}\right\}. \tag{S1}$$

By the representer theorem (Wahba, 1990), the solution of this infinite dimensional optimization (S1) can be shown to lie in a finite dimensional subspace of $\mathcal{H}$: $\text{span}\{\kappa(X_i,\cdot):$

*Jiayi Wang is Ph.D. candidate, Department of Statistics, Texas A&M University, College Station, TX 77843, USA (Email: jiayiwang@stat.tamu.edu).

†Raymond K. W. Wong is Associate Professor, Department of Statistics, Texas A&M University, College Station, TX 77843, USA (Email: raywong@tamu.edu). His research is partially supported by the National Science Foundation (DMS-1711952 and CCF-1934904).

‡Shu Yang is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA (Email: syang24@ncsu.edu). Her research is partially supported by the National Institute on Aging (1R01AG066883) and the National Science Foundation (DMS-1811245).

§Kwun Chuen Gary Chan is Professor, Department of Biostatistics, University of Washington, Seattle, WA 98195, USA (Email: kcgchan@u.washington.edu). His research is partially supported by the National Heart, Lung, and Blood Institute (R01HL122212) and the National Science Foundation (DMS-1711952).

$i = 1, \ldots, N\}$. Take $M = [\kappa(X_i, X_j)]_{i,j=1}^N \in \mathbb{R}^{N\times N}$, we obtain

$$\sup_{u\in\mathcal{H}_N} \left\{S_{N,h}(w,u) - \lambda_1\|u\|_{\mathcal{H}}^2\right\} = \sup_{\alpha\in\mathbb{R}^N} \left[\frac{S_{N,h}\{w, \sum_{j=1}^N \alpha_j\kappa(X_j,\cdot)\}}{\alpha^\intercal M^2\alpha/N} - \lambda_1\frac{\alpha^\intercal M\alpha}{\alpha^\intercal M^2\alpha/N}\right]. \quad \text{(S2)}$$

By the definition of $S_{N,h}(w,u)$ in (6), we have

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

where $\circ$ represents the element-wise product of two vectors, $J = [1, 1, \ldots, 1] \in \mathbb{R}^N$, and

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

Note that M is positive semi-definite. We consider the eigen-decomposition of M as

$$M = PDP^\intercal \quad \text{(S3)}$$

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. Take $\beta = N^{-1/2}DP^\intercal\alpha$. Then (S2) is equivalent to

$$\sup_{\beta\in\mathbb{R}^r:\|\beta\|_2\le 1} \beta^\intercal\left\{\frac{1}{N}P^\intercal\mathrm{diag}(T\circ w - J)G_h\mathrm{diag}(T\circ w - J)P - N\lambda_1 D^{-1}\right\}\beta.$$

Therefore,

$$\hat{w} = \operatorname*{argmin}_{w\ge 1}\left[\sigma_1\left\{\frac{1}{N}P^\intercal\mathrm{diag}(T\circ w - J)G_h\mathrm{diag}(T\circ w - J)P - N\lambda_1 D^{-1}\right\} + \lambda_2 R_{N,h}(w)\right]. \quad \text{(S4)}$$

S1.2 Proof of Lemma 1

Proof of Lemma 1. By the definition (7), $R_{N,h}(w)$ 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 $\sigma_1\{\mathrm{diag}(y)G_h\mathrm{diag}(y) + B\}$ 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 $a \in \mathbb{R}^N$,

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

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 $\beta \in \mathbb{R}^r$,

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

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

$$\begin{aligned}
&\sigma_1 \left(\mathrm{diag}\{ty_1 + (1-t)y_2\} G_h \mathrm{diag}\{ty_1 + (1-t)y_2\} + B\right) \\
&\qquad \leq t\sigma_1 \left(\mathrm{diag}(y_1) G_h \mathrm{diag}(y_1) + B\right) + (1-t)\sigma_1 \left(\mathrm{diag}(y_2) G_h \mathrm{diag}(y_2) + B\right),
\end{aligned}$$

which leads to the conclusion. □

S2 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.

S2.1 Proof of Theorem 1

Firstly, we introduce some notations. Take $\gamma_i := T_i w_i^* - 1$, where $w_i^* = 1/\pi(X_i)$ for $i = 1, \ldots, N$. And define $\mathcal{H}(1) := \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \leq 1\}$. Due to Lemma 2.1 of Lin et al. (2000), 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

$$\tilde{S}_{N,h}(w^*,u) := \left\| \frac{1}{g_h(\cdot)} \left\{ \frac{1}{Nh^{d_1}} \sum_{i=1}^{N} \gamma_i u(X_i) K\left(\frac{V_i - \cdot}{h}\right) \right\} \right\|_2^2. \tag{S5}$$

Next, we show that g_h is lower bounded. Based on Assumption 7, without loss of generality, we take $h \leq 1$. Under Assumption 6, there exists a constant c_1 such that

$$\begin{aligned} g_h(v) &= \mathbb{E}\frac{1}{h^{d_1}} K\left(\frac{V_i - v}{h}\right) = \frac{1}{h^{d_1}} \int_I K\left(\frac{V-v}{h}\right) g(V) dV = \int_{(zh+v)\in[0,1]^{d_1}} K(z) g(zh+v) dz \\ &\geq C_3 \int_{(zh+v)\in[0,1]^{d_1}} K(z) dz \geq C_3 \int_{(z+v)\in[0,1]^{d_1}} K(z) dz \\ &\geq C_3 \min\left\{ \int_{[0,1/2]^{d_1}} K(z) dz, \int_{[-1/2,0]^{d_1}} K(z) dz \right\} \geq c_1. \end{aligned} \tag{S6}$$

Then,

$$\begin{aligned} \tilde{S}_{N,h}(w^*,u) &\leq \frac{1}{\inf_{v\in[0,1]^{d_1}} g_h^2(v)} \frac{1}{h^{2d_1}} \left\| \frac{1}{N} \sum_{i=1}^{N} \gamma_i u(X_i) K\left(\frac{V_i - \cdot}{h}\right) \right\|_2^2 \\ &\leq \frac{1}{c_1^2 h^{2d_1}} \left\| \frac{1}{N} \sum_{i=1}^{N} \gamma_i u(X_i) K\left(\frac{V_i - \cdot}{h}\right) \right\|_2^2. \end{aligned} \tag{S7}$$

Below, we will establish the bound of $\left| \frac{1}{N}\sum_{i=1}^{N} \gamma_i u(X_i) K\left((V_i - v)/h\right) \right|$ uniformly for every $u \in \mathcal{H}_N$ by conditioning on v. To start with, we define $\|f\|_N := \sqrt{\frac{1}{N}\sum_{i=1}^{N} f^2(X_i)}$ for some function f, $\mathcal{K}_h := \left\{ K\left((\cdot - v)/h\right) : v \in [0,1]^{d_1} \right\}$, $\sigma_{\mathcal{K}_h,N} := \sup_{\tilde{v}\in[0,1]^{d_1}} \sqrt{\frac{1}{N}\sum_{i=1}^{N} K^2((V_i - \tilde{v})/h)}$. And we take $\mathcal{F}_{h,v} := \left\{ f : f(\check{x}) = u(\check{x}) K(\frac{\check{v}-v}{h}); u \in \mathcal{H}(1) \right\}$.

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

Lemma S1. *For every 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. Notice that for every f_1, $f_2 \in \mathcal{F}_{h,v}$, $\|f_1 - f_2\|_N \leq \|f_1\|_N + \|f_2\|_N \leq 2b\sigma_{\mathcal{K}_h,N}$. Therefore, $H(\delta, \mathcal{F}_{h,v}, L^2(\mathbb{P}_N)) = 0$, when $\delta > 2b\sigma_{\mathcal{K}_h,N}$.

By Birman and Solomyak (1967), we have $H(\epsilon, \mathcal{H}(1), \|\cdot\|_\infty) \leq A\epsilon^{-\alpha}$ for some constant $A > 0$. Therefore, the covering number $\mathcal{N}(\epsilon, \mathcal{H}(1), \|\cdot\|_\infty) \leq \exp(A\epsilon^{-\alpha})$.

Take $\mathcal{N} \subset \mathcal{H}(1)$ as the ϵ-net of $\mathcal{H}(1)$ with respect to $\|\cdot\|_\infty$. By definition, for every $u \in \mathcal{H}(1)$, there exists a $u_0 \in \mathcal{N}$, such that

$$\sup_{x\in[0,1]^d} |u(x) - u_0(x)| \leq \epsilon. \tag{S8}$$

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

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

The last inequality is due to (S8) and $\frac{1}{N}\sum_{i=1}^N K^2\left((V_i - v)/h\right) \leq \sigma^2_{\mathcal{K}_h,N}$. Therefore, we have

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

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 S2. *Under Assumptions 5, 6 and 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:*

$$\mathbb{E}\sigma^2_{\mathcal{K}_h,N} \leq c_3 h^{d_1}, \tag{S9}$$

$$\mathbb{P}(\sigma^2_{\mathcal{K}_h,N} \geq 2tc_3h^{d_1}) < c\exp\left\{-c_2tNh^{d_1}\right\}, \quad t \geq 1, \tag{S10}$$

$$\mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}} \left|\frac{1}{Nh^{d_1}}\sum_{i=1}^N K\left(\frac{V_i - \tilde{v}}{h}\right) - g_h(\tilde{v})\right| \geq tc_1\right) \leq c\exp\left\{-c_4tNh^{d_1}\right\}, \quad \frac{1}{2} \leq t < 1. \tag{S11}$$

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

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

The second last inequality is due to the symmetrization inequality from Theorem 2.1 in Koltchinskii (2011), while the last inequality is due to the contraction inequality from Theorem 2.3 in Koltchinskii (2011). 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 $\mathcal{N}(\varepsilon, \mathcal{K}_h, \|\cdot\|_N) \leq A_1\varepsilon^{-\nu_1}$. 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 Koltchinskii (2011), we have

$$\mathbb{E}\|R_N\|_{\mathcal{K}_h} \leq c\left[\sqrt{\frac{\nu_1}{N}}\sigma_{\mathcal{K}_h}\sqrt{\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}} + \frac{\nu_1C_2}{N}\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}\right], \tag{S12}$$

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

$$\sigma^2_{\mathcal{K}_h} = \sup_{\tilde{v}\in[0,1]^{d_1}} \int_0^1 K^2\left(\frac{v-\tilde{v}}{h}\right) g(v)dv = h^{d_1} \sup_{\tilde{v}\in[0,1]^{d_1}} \int_{(zh+\tilde{v})\in[0,1]^{d_1}} K^2(z)g(zh+\tilde{v})dz, \tag{S13}$$

$$C_3h^{d_1} \sup_{\tilde{v}\in[0,1]^{d_1}} \int_{(zh+\tilde{v})\in[0,1]^{d_1}} K^2(z)dz \leq \sigma^2_{\mathcal{K}_h} \leq C_2^2h^{d_1}, \tag{S14}$$

$$C_3h^{d_1} \int_{[0,1]^{d_1}} K^2(z)dz \leq \sigma^2_{\mathcal{K}_h} \leq C_2^2h^{d_1}, \tag{S15}$$

where (S14) is due to $g(\cdot) \geq C_3$ and $K(\cdot) \leq C_2$; (S15) is valid for $h \leq 1$. 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

$$\begin{aligned}
\mathbb{E}\sigma^2_{\mathcal{K}_h,N} &\leq \sigma^2_{\mathcal{K}_h} + 8C_2\mathbb{E}\|R_N\|_{\mathcal{K}_h} \\
&\leq \sigma^2_{\mathcal{K}_h} + 8C_2c\left[\sqrt{\frac{\nu_1}{N}}\sigma_{\mathcal{K}_h}\sqrt{\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}} + \frac{\nu_1C_2}{N}\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}\right] \\
&\leq c_3h^{d_1}
\end{aligned}$$

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 Koltchinskii (2011)), and

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

we have for any $t \geq 1$,

$$\mathbb{P}(\sigma^2_{\mathcal{K}_h,N} \geq 2tc_3h^{d_1}) \leq c\exp\left\{-c_2tNh^{d_1}\right\},$$

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

$$\begin{aligned}
&\mathbb{E}\sup_{\tilde{v}\in[0,1]^{d_1}}\left|\frac{1}{N}\sum_{i=1}^N K\left(\frac{V_i-\tilde{v}}{h}\right) - \mathbb{E}K\left(\frac{V_i-\tilde{v}}{h}\right)\right| \leq 2\mathbb{E}\|R_N\|_{\mathcal{K}_h} \\
&\leq 2c\left[\sqrt{\frac{\nu_1}{N}}\sigma_{\mathcal{K}_h}\sqrt{\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}} + \frac{\nu_1C_2}{N}\log\frac{A_1C_2}{\sigma_{\mathcal{K}_h}}\right] \qquad \text{(S16)} \\
&\leq c_5N^{-1/2}h^{d_1/2}\sqrt{\log 1/h^{d_1}},
\end{aligned}$$

where the last inequality is due to Assumption 7, and the first term of (S16) is dominant for large enough N.

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}}\left|\frac{1}{N}\sum_{i=1}^N K\left(\frac{V_i-\tilde{v}}{h}\right) - \mathbb{E}K\left(\frac{V_i-\tilde{v}}{h}\right)\right| \geq c_5N^{-1/2}h^{d_1/2}\sqrt{\log 1/h^{d_1}} + t\right) \\
&\leq c\exp\left\{-\frac{1}{c}\frac{N^2t^2}{\tilde{V}+ntC_2}\right\},
\end{aligned}$$

where $\tilde{V} := NC_2^2 h^{d_1} + 16C_2c_5N^{1/2}h^{d_1/2}\sqrt{\log 1/h^{d_1}} \le 2NC_2^2h^{d_1}$, 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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}}\left|\frac{1}{Nh^{d_1}}\sum_{i=1}^{N}K\left(\frac{V_i-\tilde{v}}{h}\right)-g_h(\tilde{v})\right|\ge t'c_1\right)\\
&=\mathbb{P}\left(\sup_{\tilde{v}\in[0,1]^{d_1}}\left|\frac{1}{N}\sum_{i=1}^{N}K\left(\frac{V_i-\tilde{v}}{h}\right)-\mathbb{E}K\left(\frac{V_i-\tilde{v}}{h}\right)\right|\ge t'c_1h^{d_1}\right)\\
&\le c\exp\left\{-c_4t'Nh^{d_1}\right\},
\end{aligned}$$

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 every $f \in \mathcal{F}_{h,v}$.

Lemma S3. *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 $1 - c\exp(-c_6t)$,*

$$\forall f\in\mathcal{F}_{h,v},\qquad \frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right|\le t\left\{N^{-\frac{1}{2}}\|u\|_2^{\frac{2-\alpha}{2p}}h^{d_1\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)}+N^{-\frac{2}{2+\alpha}}h^{\frac{d_1\alpha}{2+\alpha}}\right\},$$

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

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

Define $\mathcal{F}_{h,v}(\delta) := \{f \in \mathcal{F}_{h,v} : \|f\|_2 \le \delta\}$ 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 $(\mathcal{F}_{h,v}, \text{dist})$, where $\text{dist}^2(f_1, f_2) = \frac{\sigma_\gamma^2}{N}\sum_{i=1}^{N}(f_1(X_i) - f_2(X_i))^2$ for $f_1, f_2 \in \mathcal{F}_{h,v}$. Therefore, by Dudley's entropy bound, and Lemma S1, for any $\delta > 0$, we have

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

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 S1, there exists a constant $c_8 > 0$ depending on A, σ_γ, α and c_3 such that

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}_{h,v}(\delta)} \frac{1}{N} \left| \sum_{i=1}^{N} \gamma_i f(X_i) \right| &\leq \frac{c}{\sqrt{N}} \mathbb{E} \int_0^{2\sigma_\gamma \delta_N} \sqrt{H(\tau, \mathcal{F}, \|\cdot\|_N)} d\tau \\
&\leq \frac{c}{\sqrt{N}} \mathbb{E} \int_0^{2\sigma_\gamma \delta_N} A^{1/2} \sigma_{\mathcal{K}_h,N}^{\alpha/2} \tau^{-\alpha/2} d\tau \\
&\leq \frac{cA^{1/2}}{\sqrt{N}} \frac{1}{1-\alpha/2} \mathbb{E} \sigma_{\mathcal{K}_h,N}^{\alpha/2} (2\sigma_\gamma \delta_N)^{1-\alpha/2} \\
&= \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} \mathbb{E} \sigma_{\mathcal{K}_h,N}^{\alpha/2} \delta_N^{1-\alpha/2} \qquad \text{(by Hölder's Inequality)} \\
&\leq \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} (\mathbb{E}\delta_N)^{1-\alpha/2} (\mathbb{E}\sigma_{\mathcal{K}_h,N})^{\alpha/2} \qquad \text{(by Jensen's Inequality)} \\
&\leq \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} (\mathbb{E}\delta_N^2)^{\frac{1-\alpha/2}{2}} (\mathbb{E}\sigma_{\mathcal{K}_h,N}^2)^{\alpha/4} \qquad \text{(by (S9) in Lemma S2)} \\
&\leq \frac{cA^{1/2}}{\sqrt{N}} \frac{(2\sigma_\gamma)^{1-\alpha/2}}{1-\alpha/2} (\mathbb{E}\delta_N^2)^{\frac{1-\alpha/2}{2}} (c_3 h^{d_1})^{\alpha/4} \\
&\leq c_8 N^{-1/2} h^{d_1 \alpha/4} (\mathbb{E}\delta_N^2)^{\frac{1-\alpha/2}{2}}
\end{aligned}$$

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

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

where r_i, $i = 1, \ldots, n$, are independent Rademacher random variables. Applying the entropy bound from Lemma S1 and with Theorem 3.12 in , we have

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

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

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

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

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

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 $t \ge 1$,

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

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

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

Therefore, with probability at least $1 - c\exp(-c_{13}t')$, we have

$$\forall f \in \mathcal{F}_{h,v} \qquad \frac{1}{N}\left|\sum_{i=1}^{N}\gamma_i f(X_i)\right| \le t'\left\{N^{-1/2}h^{d_1\alpha/4}\|f\|_2^{1-\alpha/2} + N^{-\frac{2}{2+\alpha}}h^{\frac{d_1\alpha}{2+\alpha}}\right\}, \tag{S17}$$

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 (S17) and taking $t = t'\max\{b^2, 1\}$, we finally get

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

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 S4. *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 S1, 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 Koltchinskii (2011), 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 Bartlett et al. (2005). 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 Bartlett et al. (2005). 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 Bartlett et al. (2005) 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 S3 and S4, we can see that for any $t_1, t_2 \ge \max\{c_7, c_{14}, 1\}$, with probability at least $1 - \left\{c\exp(-c_6 t_1) + \exp(-c_{14}t_2 N^{\alpha/(2+\alpha)})\right\}$, we have

$$\forall f \in \mathcal{F}_{h,v} \qquad \frac{1}{N}\left|\sum_{i=1}^N \gamma_i f(X_i)\right| \le t_1 t_2 \left\{N^{\frac{-1}{2}}(\|u\|_N)^{\frac{2-\alpha}{2p}} h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1} + N^{\frac{-2}{2+\alpha}} h^{\frac{d_1\alpha}{2+\alpha}} + N^{\frac{-1}{2}-\frac{2-\alpha}{2p(2+\alpha)}} h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}\right\}. \tag{S18}$$

Let $s \ge 1$. Note that $\{u/\|u\|_{\mathcal{H}} : \|u\|_N \le 1\} \subseteq \mathcal{H}(1)$. Using (S18), we have, with probability at least $1 - \left\{c\exp(-c_6 t_1) + \exp(-c_{14}t_2 N^{\alpha/(2+\alpha)})\right\}$, uniformly for all $u \in \mathcal{H}$ with $\|u\|_N \le 1$,

$$\frac{1}{N}\left|\sum_{i=1}^N \gamma_i \frac{u(X_i)}{\|u\|_{\mathcal{H}}} K\left(\frac{V_i - v}{h}\right)\right| \le t_1 t_2 \left\{N^{\frac{-1}{2}} \left\|\frac{u}{\|u\|_{\mathcal{H}}}\right\|_N^{\frac{2-\alpha}{2p}} h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1} + N^{\frac{-2}{2+\alpha}} h^{\frac{d_1\alpha}{2+\alpha}} + N^{\frac{-1}{2}-\frac{2-\alpha}{2p(2+\alpha)}} h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1}\right\}$$

$$\frac{1}{N}\left|\sum_{i=1}^N \gamma_i u(X_i) K\left(\frac{V_i - v}{h}\right)\right| \le t_1 t_2 \left\{N^{\frac{-1}{2}} \|u\|_{\mathcal{H}}^{1-\frac{2-\alpha}{2p}} h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1} + \nu_{N,h}\|u\|_{\mathcal{H}}\right\}, \tag{S19}$$

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

$$L(N, h, p, u) := N^{-\frac{1}{2}} \|u\|_{\mathcal{H}}^{1-\frac{2-\alpha}{2p}} h^{\left(\frac{1}{2}-\frac{2-\alpha}{4p}\right)d_1} + \nu_{N,h}\|u\|_{\mathcal{H}}, \tag{S20}$$

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

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

Lemma S5. *Under Assumption 2-7,*

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

where L is defined in (S20), $p \ge 1$, $h > 0$ can depend on N.

Proof. First, take

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

Due to (S19), we can show that for any $t \geq \max\{c_7, c_{14}, 1\}$,

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

with probability at least $1 - 2c\exp(-c_6 t)$ 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 $k \geq 1$,

$$\begin{aligned}
\mathbb{E}\left(Q^2(v)\right)^k &= \int_0^\infty \mathbb{P}\left(Q(v)^{2k} > t\right) dt = \int_0^\infty \mathbb{P}\left(Q(v) > t^{\frac{1}{2k}}\right) dt \\
&\leq \tilde{c}(k) + \int_{\tilde{c}(k)}^\infty 2c\exp(-c_6 t^{\frac{1}{4k}}) dt \\
&= \tilde{c}(k) + 4k \int_{\max\{c_7, c_{14}, 1\}}^\infty 2c\exp(-c_6 t')(t')^{4k-1} dt' \\
&\leq \tilde{c}(k) + c_{17} k\Gamma(4k),
\end{aligned}$$

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 (S7), we have for $t > 0$ and positive integer k,

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

where $c_{18}(k) > 0$ is a constant depending on k. And 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 (S11) in Lemma S2, we can see that with probability at least $1 - c\exp\{-c_4 t' N h\}$,

where $\frac{1}{2} \leq t' \leq 1$,

$$\begin{aligned}
\forall \tilde{v} \in [0,1]^{d_1}, \qquad & \left| \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h}\right) - g_h(\tilde{v}) \right| \leq t' c_1 \leq t' g_h(\tilde{v}) \\
& \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h}\right) - g_h(\tilde{v}) \geq -t' g_h(\tilde{v}) \\
& \frac{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h}\right)}{g_h(\tilde{v})} \geq 1 - t' \\
& \frac{g_h(\tilde{v})}{\frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{V_i - \tilde{v}}{h}\right)} \leq \frac{1}{1 - t'}
\end{aligned} \tag{S21}$$

Therefore,

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

□

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

Lemma S6. *Under Assumptions 2-7,*

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

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 (S10) in Lemma S2,

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

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

$$\tilde{R}_{N,h}(w^*) \leq \int_{[0,1]^{d_1}} \frac{1}{g_h^2(v)} dv \left\{ \frac{1}{h^{2d_1}} 2 C_1^2 c_3 t h^{d_1} \right\} \leq \frac{2 C_1^2 c_3^2}{c_1^2} t h^{-d_1}, \tag{S22}$$

with probability at least $1 - c\exp(-c_2 tNh^{d_1}/c)$. Combining with the results from (S21), we have

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

□

Now, we are ready to prove Theorem 1.

Proof of Theorem 1. Take $u^* = \operatorname{argmax}_{u\in\mathcal{H}_N}\left\{S_{N,h}(w^*,u) - \lambda_1\|u\|_{\mathcal{H}}^2\right\}$. Its existence is shown in Section S1.

Due to (8), we have the following basic inequality:

$$S_{N,h}(\hat{w},m_1) + \lambda_1\|u^*\|_{\mathcal{H}}^2\|m_1\|_N^2 + \lambda_2 R_{N,h}(\hat{w})\|m_1\|_N^2 \le S_{N,h}(w^*,u^*)\|m_1\|_N^2 + \lambda_1\|m_1\|_{\mathcal{H}}^2 + \lambda_2 R_{N,h}(w^*)\|m_1\|_N^2 \tag{S23}$$

From Lemmas S5 and S6, we have $R_{N,h}(w^*) = \mathcal{O}_{\mathrm{p}}(h^{-d_1})$ and

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

for all $p \ge 1$.

We now compare different scenarios of (S23).

Case 1: Suppose that $S_{N,h}(w^*,u^*)\|m\|_N^2$ is the largest in the right-hand side of (S23).

If $\|m\|_N \ne 0$, we have $\lambda_1\|u^*\|_{\mathcal{H}}^2 \le \mathcal{O}_{\mathrm{p}}\left(N^{-1}\|u^*\|_{\mathcal{H}}^{2-(2-\alpha)/p} h^{-1-\frac{2-\alpha}{2p}}\right) + \mathcal{O}_{\mathrm{p}}\left(\nu_{N,h}^2 h^{-2}\|u^*\|_{\mathcal{H}}^2\right)$. By Assumptions 3 and 7, we can see that

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

Therefore we only need to consider $\lambda_1\|u^*\|_{\mathcal{H}}^2 \le \mathcal{O}_{\mathrm{p}}\left(N^{-1}\|u^*\|_{\mathcal{H}}^{2-(2-\alpha)/p} h^{\left(-1+\frac{\alpha-2}{2p}\right)d_1}\right)$. 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 $S_{N,h}(\hat{w}, m) = 0 \leq \lambda_1^{\frac{-2p+(2-\alpha)}{(2-\alpha)}} \mathcal{O}_{\mathrm{p}}\left(N^{\frac{-2p}{(2-\alpha)}} h^{\left(\frac{-2p}{(2-\alpha)}-1\right)d_1}\right) \|m\|_N^2$.

Case 2: Suppose that $\lambda_1\|m\|_{\mathcal{H}}^2$ is the largest in right-hand side of (S23). Then we have $S_{N,h}(\hat{w}, m) \leq 3\lambda_1\|m\|_{\mathcal{H}}^2 = \mathcal{O}_{\mathrm{p}}(\lambda_1)\|m\|_{\mathcal{H}}^2$.

Case 3: Suppose that $\lambda_2 R_{N,h}(w^*)$ is the largest in right-hand side of (S23). Then we have $S_{N,h}(\hat{w}, m) \leq 3\lambda_2\, \mathcal{O}_{\mathrm{p}}(h^{-d_1})\|m\|_N^2 = \mathcal{O}_{\mathrm{p}}(\lambda_2 h^{-d_1})\|m\|_N^2$.

Combining these cases, we have

$$S_{N,h}(\hat{w}, m_1) = \max\left\{\min\left\{\lambda_1^{\frac{-2p+(2-\alpha)}{(2-\alpha)}} \mathcal{O}_{\mathrm{p}}\left(N^{\frac{-2p}{(2-\alpha)}} h^{\left(\frac{-2p}{(2-\alpha)}-1\right)d_1}\right) \|m_1\|_N^2 : p \geq 1\right\},\right.$$
$$\left.\mathcal{O}_{\mathrm{p}}(\lambda_1)\|m_1\|_{\mathcal{H}}^2, \mathcal{O}_{\mathrm{p}}(\lambda_2 h^{-d_1})\|m_1\|_N^2\right\}. \quad \text{(S24)}$$

Next, we compare the first two components of (S24). We can see that as long as

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

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 $p \geq \frac{2-\alpha}{2}\frac{\log h^{d_1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}$. So we can pick any $p \geq \max\{1, \frac{2-\alpha}{2}\frac{\log h^{d_1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}\}$ 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 (S24). Similar to the previous analysis, as long as

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

the third component is dominant. Due to the condition that $\lambda_1^{-1} = o(Nh^{d_1})$, the inequality is valid if $p \geq \frac{2-\alpha}{2}\frac{\log \lambda_2\lambda_1^{-1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}$. So we can pick any $p \geq \max\{1, \frac{2-\alpha}{2}\frac{\log \lambda_2\lambda_1^{-1}}{\log(\lambda_1^{-1}N^{-1}h^{-d_1})}\}$ 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 (S23), by replacing m with a constant function and applying the similar analysis as above, we can conclude that $R_{N,h}(\hat{w}) = \mathcal{O}_{\mathrm{p}}(h^{-d_1})$.

□

S2.2 Proof of Theorem 2

Proof. First,

$$\begin{aligned}
&\left\| \frac{1}{N}\sum_{i=1}^{N} T_i\hat{w}_i Y_i K_h\left(V_i,\cdot\right) - \mathbb{E}\left\{Y(1) \mid V=\cdot\right\} \right\|_2 \\
\leq &\left\| \frac{1}{N}\sum_{i=1}^{N} (T_i\hat{w}_i - 1) K_h\left(V_i,\cdot\right) m(X_i) \right\|_2 + \left\| \frac{1}{N}\sum_{i=1}^{N} T_i\hat{w}_i K_h(V_i,\cdot)\varepsilon_i \right\|_2 \qquad \text{(S25)} \\
+ &\left\| \frac{1}{N}\sum_{i=1}^{N} m(X_i) K_h(V_i,\cdot) - \mathbb{E}\left\{Y(1) \mid V=\cdot\right\} \right\|_2 . \qquad \text{(S26)}
\end{aligned}$$

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

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

due to Theorem 1 and the conditions of λ_1 and λ_2.

For the second term in (S25), we have $\mathbb{E}(\varepsilon_i \mid T_i, \hat{w}_i, X_i, i = 1, \ldots, N) = 0$. Then

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

Therefore,

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

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

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

From the condition of λ_2, and the result from Theorem 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 (S26), it has a form of a typical Nadaraya–Watson estimator. By Theorem 5.44 in Wasserman (2006) we have

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

Therefore,

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

Overall, conclusion follows.

□

S2.3 Proof outline of Theorem 3

To obtain the rate, the entropy bound in Lemma S1 needs to be modified to the bigger function class $\mathcal{F}_h := \{f : f(\check{x}) = u(\check{x})K(\frac{\check{v}-v}{h}), u \in \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \leq 1\}, v \in [0,1]^{d_1}\}$. This can be done by combining the entropy bound for $\{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \leq 1\}$ and Assumption 5(b). One can show that

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

Then by adopting this entropy bound, the results in Lemma S3 will be modified to

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

for any $t \ge c_1$, and $p \ge 1$ with probability at least $1 - c\exp(-c_6 t)$. Then the remaining argument is similar to those in the proof of Theorems 1 and 2.

S2.4 Proof of Theorem 4

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

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

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 1, by replacing m with $\hat{e}$, we have

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

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,

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

□

References

Bartlett, P. L., O. Bousquet, S. Mendelson, et al. (2005). Local rademacher complexities. *The Annals of Statistics 33*(4), 1497–1537.

Birman, M. S. and M. Z. Solomyak (1967). Piecewise-polynomial approximations of functions of the classes w_pˆα. *Matematicheskii Sbornik 115*(3), 331–355.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, Volume 2033. Springer Science & Business Media.

Lin, Y. et al. (2000). Tensor product space anova models. *The Annals of Statistics 28*(3), 734–755.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

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