

# Estimation of Conditional Average Treatment Effects with High-Dimensional Data

Qingliang Fan\*    Yu-Chin Hsu<sup>†</sup>    Robert P. Lieli<sup>‡</sup>    Yichong Zhang<sup>§</sup>

August 8, 2019

## Abstract

Given the unconfoundedness assumption, we propose new nonparametric estimators for the reduced dimensional conditional average treatment effect (CATE) function. In the first stage, the nuisance functions necessary for identifying CATE are estimated by machine learning methods, allowing the number of covariates to be comparable to or larger than the sample size. This is a key feature since identification is generally more credible if the full vector of conditioning variables, including possible transformations, is high-dimensional. The second stage consists of a low-dimensional kernel regression, reducing CATE to a function of the covariate(s) of interest. We consider two variants of the estimator depending on whether the nuisance functions are estimated over the full sample or over a hold-out sample. Building on Belloni et al. (2017) and Chernozhukov et al. (2018), we derive functional limit theory for the estimators and provide an easy-to-implement procedure for uniform inference based on the multiplier bootstrap. The empirical application revisits the effect of maternal smoking on a baby's birth weight as a function of the mother's age.

**Keywords:** heterogenous treatment effects, high-dimensional data, uniform confidence band

**JEL codes:** C14, C21, C55

---

\*Wang Yanan Institute for Studies in Economics (WISE), School of Economics, and Fujian Key Lab of Statistics, Xiamen University. E-mail: michaelqfan@gmail.com.

<sup>†</sup>Institute of Economics, Academia Sinica, Taiwan. Department of Finance, National Central University and Department of Economics, University Chengchi University. E-mail: ychsu@econ.sinica.edu.tw.

<sup>‡</sup>Department of Economics, Central European University, Budapest. E-mail: lielir@ceu.edu.

<sup>§</sup>School of Economics, Singapore Management University. E-mail: yczhang@smu.edu.sg. Zhang acknowledges financial support from the Singapore Ministry of Education Tier 2 grant under grantno. MOE2018-T2-2-169 and the Lee Kong Chian Fellowship.

# 1 Introduction

In settings with individual level treatment effect heterogeneity, the unconfoundedness assumption theoretically permits consistent estimation of the conditional average treatment effect (CATE) for all possible values of the set of covariates  $X$  used in adjusting for selection bias. One way to think about these covariates is that they are ex-ante predictors of an individual’s potential outcomes with and without treatment, and hence are highly correlated with the treatment participation decision as well. Unconfoundedness states that the econometrician observes all relevant predictors so that conditional on  $X$ , the treatment takeup decision is no longer statistically related to the potential outcomes.<sup>1</sup> Nevertheless, in many situations the individual deciding on treatment participation is likely to have access to private signals about their potential outcomes. Relying on the unconfoundedness assumption amounts to hoping that a set of publicly observed characteristics can still proxy for the information content of these signals. Therefore, the unconfoundedness assumption is more credible in applications in which  $X$  is a rich, detailed set of covariates, i.e., the dimension of  $X$  is high. This has several practical consequences.

First, while CATE as a function of  $X$  provides a detailed characterization of treatment effect heterogeneity across observable subpopulations, this information is very hard to analyze and convey if  $X$  is high dimensional. Of course, one could examine slices of this function along some component(s)  $X_1$  of  $X$  while holding the other components  $X_{-1}$  of  $X$  constant. Nevertheless, how CATE varies as a function of  $X_1$  will generally depend on the level at which  $X_{-1}$  is held constant, requiring the examination of (infinitely) many different slices. For this reason, instead of holding the variables in  $X_{-1}$  constant, [Abrevaya, Hsu, and Lieli \(2015\)](#) suggest integrating them out with respect to the conditional distribution of  $X_{-1}$  given  $X_1$  or, in practice, a smoothed estimate of this distribution. This gives rise to a reduced dimensional CATE function that is easier to present and interpret. If all covariates  $X$  are integrated out, one obtains an estimator of ATE as in [Hahn \(1998\)](#) or [Hirano, Imbens, and Ridder \(2003\)](#).

Second, while the difference in the mean outcome across participants and non-participants

---

<sup>1</sup>This condition was formalized by [Rosenbaum and Rubin \(1983\)](#); since then, unconfoundedness (or “selection on observables” or “conditional independence”) has become one of the standard paradigms for modeling selection effects. See, e.g., [Imbens and Wooldridge \(2009\)](#) for further discussion.

conditional on  $X$  identifies CATE as a function of  $X$ , flexible estimation of these regression functions can become cumbersome in practice. Consider, for example, a control function approach (see, e.g., [Wooldridge \(2010, Chapter 21\)](#)), where one first builds a “dictionary”  $b(X)$  of technical controls consisting of powers and interactions of  $X$  up to some order, and then runs an OLS regression of the outcome on the treatment dummy  $D$ , the full set of controls  $b(X)$  and the interaction between  $D$  and  $b(X)$ . With  $\dim(X) = 10$  raw covariates,  $\dim(b(X)) \leq 65$  if up to quadratic terms are included<sup>2</sup>, and  $\dim(b(X)) \leq 285$  if cubic terms are added as well. With  $\dim(X) = 30$ , the corresponding figures are 495 and 5455(!), respectively. Clearly, in many applications the number of regressors will be comparable to the sample size or will even exceed it. While other nonparametric regression methods (such as kernel based or local linear regression) do not use a dictionary, they too will suffer from statistical and implementation problems related to the curse of dimensionality already when  $\dim(X)$  is moderately large.

In this paper we propose two-step estimators of the reduced dimensional CATE function where in the first step the required high dimensional nuisance regressions are conducted by machine learning methods designed specifically to handle such problems, while the second integration step is implemented by a traditional kernel-based nonparametric regression. (This step assumes that  $X_1$  is a continuous variable, which is the technically challenging and interesting case.) We derive the statistical properties of two variants of the estimator. In the first case, the first step (nuisance function estimation) and the second step (kernel regression) are both implemented over the full sample of available observations. In the second case, the available sample is split into parts, and the first step is implemented in one subsample while the second step is done in the complement sample. The roles of the subsamples are then rotated and the results are averaged. This is the “cross-fitting” approach to machine-learning-aided causal inference advocated by [Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins \(2018\)](#). The first approach is used by [Belloni, Chernozhukov, Fernández-Val, and Hansen \(2017\)](#) in estimating unconditional treatment effects.

In proposing and studying these estimators, we contribute to two recent strands of the econometrics literature. First, we advance the currently available flexible methods for the estimation of reduced dimensional CATE functions due to [Abrevaya et al. \(2015\)](#) and [Lee,](#)

---

<sup>2</sup>We use the less than or equal sign to account for the fact that  $X$  might contain dummy variables, which makes powers redundant.

Okui, and Whang (2017) (henceforth LOW). Second, we make technical contributions to the recent literature that employs machine learning methods in tackling the prediction component of causal inference problems; see Belloni, Chernozhukov, and Hansen (2014a,b), Belloni et al. (2017), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017), Chernozhukov et al. (2018). We now describe in detail how our paper fits into and extends these lines of research.

Abrevaya et al. (2015) use an inverse probability weighted conditional moment of the data to identify CATE. They consider both kernel based and parametric estimation of the propensity score in the first step, and derive the asymptotic distribution of the estimated CATE function evaluated at a fixed point  $x_1$  in the support of  $X_1$ . LOW advance these results in two respects: their estimator is based on a Neyman-orthogonal moment condition and they also provide a method for uniform inference about the CATE function as a whole (rather than point-by-point). While LOW only use parametric models to estimate the nuisance functions involved in the moment condition, orthogonality lends their estimator a “double robustness” property: either the model for the propensity score or the models for the conditional means of the potential outcomes are allowed to be misspecified (but not both).

The CATE estimator proposed here is based on the same Neyman-orthogonal moment condition as in LOW, but the required nuisance functions are estimated by machine learning methods, which allows for data-driven flexible functional forms as well as a (very) high dimensional set of covariates. Neyman-orthogonality is crucial in ensuring that the proposed CATE estimators are robust to the regularization bias inherent in the first stage, making post-selection inference possible. As the asymptotic theory is derived from high level assumptions, there are a number of applicable first stage estimation methods in practice, such as a random forest or  $\ell_1$ -penalized lasso or post-lasso. In this paper we use lasso estimation as the leading example, and discuss in some detail the primitive sparseness conditions under which it satisfies the stated high level assumptions.

In light of the discussion of the unconfoundedness assumption above, replacing the parametric estimators in LOW with machine learning methods greatly enhances the applicability and empirical relevance of flexible CATE estimation. At the same time, the asymptotic theory remains tractable: we provide methods for pointwise as well as uniform inference about the CATE function under both the full sample and sample splitting implementation schemes. The

uniform methods utilize the multiplier bootstrap while pointwise inference can be based either on the bootstrap or the analytic results.

Turning to the literature on machine learning in treatment effect estimation, we build primarily on [Chernozhukov et al. \(2017, 2018\)](#) for the split sample method and [Belloni et al. \(2017\)](#) for the full sample method, whilst providing the necessary extension of the theory to account for the use of kernel regression in second step. In these papers the parameter of interest is identified by the restriction that the unconditional expectation of a “score function” evaluated at the true parameter value (and the true nuisance functions) is zero. By contrast, the identifying restriction in our case is that the *conditional* expectation of the same score function is zero. Hence, our estimation procedure does not simply consist of substituting in the estimated nuisance functions and setting the sample average score to zero; instead, the score function is first multiplied by a kernel  $\mathcal{K}((X_{1i} - x_1)/h)/h$  and then averaged over the sample data, where  $h$  denotes a smoothing parameter (bandwidth).

The multiplication by the kernel function has consequences for the asymptotic theory. The key high level assumptions we employ in deriving our asymptotic results involve bounding the  $L_\infty$  norm of the difference between the true and estimated nuisance functions, and the  $L_2$  norm of the same difference multiplied by the kernel. The rates at which these error bounds are required to converge to zero are closely linked to the rate at which the bandwidth sequence converges to zero. (For example, the bandwidth must obey a standard “undersmoothing” condition in order to eliminate the asymptotic bias emanating from the second stage kernel regression.) From a purely technical standpoint, incorporating the bandwidth conditions into the high level norm bounds in the full-sample as well as the cross-fitting case is a central contribution of the paper. Similarly to [Abrevaya et al. \(2015\)](#) and [LOW](#), the resulting convergence rate of the CATE estimators is  $\sqrt{Nh^d}$ , where  $N$  is the sample size and  $d = \dim(X_1)$ .

In addition to the errors bounds, the full-sample estimator also requires controlling the complexity (entropy) of the function space in which the nuisance functions take values. In case of lasso estimation, this can be accomplished by restricting how fast the number of covariates and the sparsity indices associated with the nuisance functions are allowed to increase with the sample size. These conditions are more stringent than in case of estimating ATE.

The only other paper we are aware of that contributes to both the literature on CATE estimation and machine-learning-aided causal inference is [Chernozhukov and Semenova \(2019\)](#).

There are, however, substantial technical differences between their paper and ours. First, the traditional nonparametric estimator used by *ibid.* in the second stage is series regression (a global method) rather than kernel regression (a local method). Second, *ibid.* only consider the cross-fitting approach and do not address the problem of estimating both the nuisance functions and the target function on the full sample. Finally, we also provide a reasonably detailed discussion of the primitive conditions under which lasso estimation fulfills the high level conditions posited in the paper.

We further study and illustrate our methods through Monte Carlo simulations and taking a new look at the application in [Abrevaya et al. \(2015\)](#) and LOW. The Monte Carlo exercises study the small sample performance of the estimators and compare the full sample and cross-fitting approaches. The proposed estimators perform well in terms of bias, MSE, and coverage rates. In general, we find that the cross-fitting estimator has somewhat better finite sample properties than the full sample estimator, and thus we suggest using the former in real data applications.

Our application uses vital statistics data from North Carolina to estimate the effect of a (first-time) mother’s smoking during pregnancy on the baby’s birth weight as a function of the mother’s age. Despite the previous analyses, the application is well worth revisiting with the help of machine learning methods, as there is a large number of covariates describing the mother’s characteristics and events during pregnancy, and the specification of the propensity score is known to have a substantial impact on the results (see [Abrevaya et al. \(2015, Section 4.2\)](#)). Our results provide some corroborating evidence that the negative effect of smoking on birth weight becomes more detrimental with age. This pattern is less prevalent than some of the results reported in [Abrevaya et al. \(2015\)](#) but stronger than that found by LOW.

The rest of the paper proceeds as follows. In Section 2 we describe the formal setup and the estimators. Section 3 states and discusses the assumptions underlying the first-order asymptotic theory and provides the main results. Section 4 describes how to conduct uniform inference using the multiplier bootstrap. The Monte Carlo exercises and the empirical application are presented in Section 5 and 6, respectively. Section 7 concludes.

## 2 The formal framework, identification and the estimators

Population units are characterized by a random vector  $(D, Y(1), Y(0), X)$ , where  $D \in \{0, 1\}$  indicates the receipt of a binary treatment,  $Y(1)$  and  $Y(0)$  are the potential outcomes with and without the treatment, respectively, and  $X$  is a vector of pre-treatment covariates. The actually observed variables are given by the vector  $W = (D, Y, X)$ , where  $Y = DY(1) + (1 - D)Y(0)$ . The distribution of  $(D, Y(1), Y(0), X)$ , and hence  $W$ , is induced by a fixed underlying probability measure  $\mathbb{P}$ ; parameter values computed under  $\mathbb{P}$  will be denoted by the subscript '0' and represent the true values of these parameters. The expectation operator corresponding to  $\mathbb{P}$  is denoted by  $\mathbb{E}$ , but we also use the linear functional notation  $\mathbb{P}f := \int f(w)d\mathbb{P} = \mathbb{E}[f(W)]$ .

Given a  $d$ -dimensional subvector  $X_1 \subset X$  comprised of continuous variables, the reduced dimensional CATE function is defined as<sup>3</sup>

$$\tau_0(x_1) = CATE(x_1) = \mathbb{E}[Y(1) - Y(0)|X_1 = x_1].$$

The identification of  $\tau_0(x_1)$  from the joint distribution of  $W$  is facilitated by the unconfoundedness assumption along with some technical conditions:

**Assumption 2.1.** *The distribution  $\mathbb{P}$  satisfies:*

- (i) *(Unconfoundedness)*  $(Y(1), Y(0)) \perp D|X$
- (ii) *(Moments)*  $\mathbb{E}[|Y(j)|^q] < \infty$ ,  $j = 0, 1$  and  $q \geq 4$ .
- (iii) *(Propensity score)* Let  $\pi_0(x) = \mathbb{P}(D = 1|X = x)$ . There exists some constant  $\underline{C} > 0$  so that  $\mathbb{P}(\underline{C} \leq \pi_0(X) \leq 1 - \underline{C}) = 1$ .

Let  $\mu_0(j, x) = \mathbb{E}[Y|X = x, D = j]$ ,  $j = 0, 1$ . It follows immediately from Assumption 2.1 that  $E[Y(j)|X_1 = x_1] = E[\mu_0(j, X)|X_1 = x_1]$ , and hence  $\tau_0(x_1)$  is identified as

$$\tau_0(x_1) = \mathbb{E}[\mu_0(1, X) - \mu_0(0, X)|X_1 = x_1].$$

We now state a less obvious but more robust result based on a Neyman-orthogonal moment condition. Given any probability measure satisfying Assumption 2.1, let  $\tau(\cdot)$ ,  $\mu(1, \cdot)$ ,  $\mu(0, \cdot)$ ,

---

<sup>3</sup>The most relevant case in practice is  $d = 1$  or perhaps  $d = 2$ , for otherwise the motivating properties of the reduced dimensional CATE function (interpretability and presentability) are lost. As we will see below, under a fourth moment condition on  $Y$ , the general theory requires  $d \leq 3$ . There are no restrictions on  $d$  for bounded outcomes.

$\pi(\cdot)$  denote the functions corresponding to  $\tau_0(\cdot)$ ,  $\mu_0(1, \cdot)$ ,  $\mu_0(0, \cdot)$ ,  $\pi_0(\cdot)$ , respectively. Let  $\eta = (\pi(\cdot), \mu(1, \cdot), \mu(0, \cdot))$  represent the infinite dimensional nuisance parameters needed to identify CATE, and define

$$\psi(W; \eta) = \frac{D(Y - \mu(1, X))}{\pi(X)} + \mu(1, X) - \frac{(1 - D)(Y - \mu(0, X))}{1 - \pi(X)} - \mu(0, X).$$

The following theorem gives a moment condition that is (at least approximately) satisfied at  $(\tau_0, \eta)$  even when  $\eta$  deviates from  $\eta_0$ .

**Theorem 2.1.** (i) Under Assumption 2.1,

$$\begin{aligned} \mathbb{E} \left[ \frac{D(Y - \mu_0(1, X))}{\pi_0(X)} + \mu_0(1, X) \middle| X_1 = x_1 \right] &= \mathbb{E}(Y(1) \mid X_1 = x_1) \\ \mathbb{E} \left[ \frac{(1 - D)(Y - \mu_0(0, X))}{1 - \pi_0(X)} + \mu_0(0, X) \middle| X_1 = x_1 \right] &= \mathbb{E}(Y(0) \mid X_1 = x_1) \end{aligned}$$

for all  $x_1$  in the support of  $X_1$ .

(ii)  $\mathbb{E}[\psi(W; \eta_0) - \tau_0(X_1) \mid X_1 = x_1] = 0$  by part (i), and this moment equation satisfies the Neyman-orthogonality condition

$$\partial_r \mathbb{E}[\psi(W; \eta_0 + r(\eta - \eta_0)) - \tau_0(X_1) \mid X_1 = x_1] \Big|_{r=0} = 0. \quad (1)$$

**Remarks:**

1. Assumption 2.1(iii) is not necessary for Theorem 2.1; a weaker moment condition such as  $\mathbb{E}[1/\pi_0^2(X)] < \infty$  would suffice. Nevertheless, the overlap condition stated under Assumption 2.1(iii) is indispensable for subsequent results concerned with the asymptotic distribution of our CATE estimators. Similarly, for identification only, the fourth moment condition in Assumption 2.1(ii) could be replaced by a second moment condition.
2. If  $\eta = (\pi_0, \mu(0, \cdot), \mu(1, \cdot))$  or  $\eta = (\pi, \mu_0(0, \cdot), \mu_0(1, \cdot))$ , i.e.,  $\eta$  deviates from  $\eta_0$  along select coordinates at a time, then  $\mathbb{E}[\psi(W; \eta_0 + r(\eta - \eta_0)) - \tau_0(X_1) \mid X_1 = x_1] = 0$  for any value of  $r$ , which of course implies (1). This is the “double robustness property” emphasized by LOW; it implies that if  $\pi(\cdot)$  and  $(\mu(0, \cdot), \mu(1, \cdot))$  are parametric models for  $\pi_0(\cdot)$  and  $(\mu_0(0, \cdot), \mu_0(1, \cdot))$ , respectively, and one of these models is misspecified, then one can still consistently estimate  $\tau_0(x_1)$  based on the moment condition  $\mathbb{E}[\psi(W; \eta) - \tau_0(X_1) \mid X_1 = x_1] = 0$ .

We now present the proposed CATE estimators. While the high level assumptions stated in Section 3 can accommodate multiple machine learning procedures for estimating  $\eta_0$ , here we describe the first stage using lasso as a concrete example. Specifically, let  $b(X) = (b_1(X), \dots, b_p(X))$  be a dictionary of control terms based on  $X$ , where  $p$  is potentially larger than the sample size  $N$  and can grow with  $N$ . Typically,  $b(X)$  consists of powers and interactions of the components of  $X$ , but other bases could also be used. The lasso approximates the nuisance functions  $\eta_0$  with linear combinations of the components  $b_i(X)$ ; in particular, for  $p$ -vectors  $\beta$ ,  $\alpha$  and  $\theta$ , set

$$r_\alpha(x) := \mu_0(0, x) - b(X)' \alpha, \quad r_\beta(x) := \mu_0(1, x) - b(X)' \beta, \quad r_\theta(x) := \pi_0(x) - \Lambda(b(X)' \theta) \quad (2)$$

where  $\Lambda(\cdot)$  is the logistic c.d.f. The primitive condition that justifies using the lasso is approximate sparsity. Intuitively, it means that it is possible to make the approximation errors  $r_\alpha$ ,  $r_\beta$ ,  $r_\theta$  small with just a small number of approximating terms, i.e., with  $\alpha$ ,  $\beta$  and  $\theta$  having only a handful of non-zero components (we will discuss this assumption formally in Section 3).<sup>4</sup> The coefficients  $\alpha$ ,  $\beta$  and  $\theta$  are estimated by penalized least squares or maximum likelihood, where a penalty is imposed for any non-zero component. For suitable specifications of the penalty and other implementation details see Belloni et al. (2014b) or Belloni et al. (2017).

We propose two versions of the CATE estimator, depending on whether the first stage approximation to  $\eta_0$  and the second stage kernel regression targeting  $\tau_0$  take place over the same sample or not. To facilitate the formal definitions, the following assumption describes the properties and use of the sample data:

**Assumption 2.2.** (i) *The observed data consists of  $N$  independent and identically distributed (i.i.d.) random vectors  $\{W_i\}_{i=1}^N = \{(D_i, Y_i, X_i)\}_{i=1}^N$  with the same distribution as the population distribution of  $W$ .*

(ii) *Let  $K$  be a (small) positive integer, and (for simplicity) suppose that  $n = N/K$  is also an integer. Let  $I_1, \dots, I_K$  be a random partition of the index set  $I = \{1, \dots, N\}$  so that  $\#I_k = n$  for  $k = 1, \dots, K$ .*

**The full sample estimator** Let  $\hat{\mu}_0(0, x; I) = b(x)' \hat{\alpha}(I)$ ,  $\hat{\mu}_0(1, x; I) = b(x)' \hat{\beta}(I)$  and  $\hat{\pi}_0(x) = \Lambda(b(x)' \hat{\theta}(I))$ , where  $\hat{\alpha}(I)$ ,  $\hat{\beta}(I)$  and  $\hat{\theta}(I)$  are lasso regression coefficients estimated over the full

---

<sup>4</sup>The linear index structure and approximate sparsity are specific to the lasso; other machine learning methods provide different types of approximations which do not necessarily rely on sparsity.

sample  $I$ . Furthermore, let

$$\hat{f}(x_1; I) = \frac{1}{Nh^d} \sum_{i \in I} \mathcal{K}_h(X_{1i} - x_1)$$

denote a kernel density estimator of the p.d.f. of  $X_1$  over  $I$ , where  $\mathcal{K}$  is a  $d$ -dimensional product kernel,  $h$  is a smoothing parameter (bandwidth), and  $\mathcal{K}_h(u) = \mathcal{K}\left(\frac{u}{h}\right)$ . Set  $\hat{\eta}(I) = (\hat{\mu}_0(0, \cdot; I), \hat{\mu}_0(1, \cdot; I), \hat{\pi}(\cdot; I))$ . The second stage of the full sample estimator consists of the nonparametric regression

$$\hat{\tau}(x_1) = \frac{1}{Nh^d \hat{f}(x_1; I)} \sum_{i \in I} \psi(W_i, \hat{\eta}(I)) \mathcal{K}_h(X_{1i} - x_1). \quad (3)$$

**The  $K$ -fold cross fitting estimator** For each  $k = 1, \dots, K$ , let  $\hat{\mu}_0(0, x; I_k^c) = b(x)' \hat{\alpha}(I_k^c)$ ,  $\hat{\mu}_0(1, x; I_k^c) = b(x)' \hat{\beta}(I_k^c)$  and  $\hat{\pi}_0(x; I_k^c) = \Lambda(b(x)' \hat{\theta}(I_k^c))$ , where  $\hat{\alpha}(I_k^c)$ ,  $\hat{\beta}(I_k^c)$  and  $\hat{\theta}(I_k^c)$  are lasso regression coefficients estimated over the subsample  $I_k^c = I \setminus I_k$ . Furthermore, let

$$\hat{f}(x_1; I_k) = \frac{1}{nh^d} \sum_{i \in I_k} \mathcal{K}_h(X_{1i} - x_1)$$

denote a kernel density estimator of the p.d.f. of  $X_1$  over the subsample  $I_k$ . Set  $\hat{\eta}(I_k^c) = (\hat{\mu}_0(0, \cdot; I_k^c), \hat{\mu}_0(1, \cdot; I_k^c), \hat{\pi}(\cdot; I_k^c))$ . The second stage of the  $K$ -fold cross-fitting estimator consists of  $K$  nonparametric regressions over the samples  $I_1, \dots, I_K$ :

$$\tilde{\tau}_k(x_1) = \frac{1}{nh^d \hat{f}(x; I_k)} \sum_{i \in I_k} \psi(W_i, \hat{\eta}(I_k^c)) \mathcal{K}_h(X_{1i} - x_1). \quad (4)$$

Finally, in the third stage we take the average of the  $K$  preliminary estimates to obtain an efficient estimator:

$$\tilde{\tau}(x_1) = \frac{1}{K} \sum_{k=1}^K \tilde{\tau}_k(x_1).$$

### 3 Asymptotic properties

In this section we provide the fundamental asymptotic results for our CATE estimators which form the basis of the uniform inference procedures to be given in Section 4. To this end, we state and discuss several assumptions.

Let  $\mathcal{X}_1 \subset \mathbb{R}^d$  denote the support of  $X_1$  and let  $\bar{\mathcal{X}}_1$  be the subset of  $\mathcal{X}_1$  over which  $\tau_0(x_1)$  is to be estimated. In addition, let  $f(x_1)$  denote the p.d.f. of  $X_1$ .

**Assumption 3.1.** *Assume that*

- (i) *The set  $\bar{\mathcal{X}}_1$  is contained in the interior of  $\mathcal{X}_1$  and is the Cartesian product of closed intervals, i.e.,  $\bar{\mathcal{X}}_1 = \prod_{j=1}^d [x_{1\ell}^{(j)}, x_{1u}^{(j)}]$  with  $x_{1\ell}^{(j)} < x_{1u}^{(j)}$ . Furthermore, there exist positive constants  $\underline{C}$  and  $\bar{C}$  such that:*

$$\underline{C} \leq \inf_{x_1 \in \bar{\mathcal{X}}_1} f(x_1) \leq \sup_{x_1 \in \bar{\mathcal{X}}_1} f(x_1) \leq \bar{C} \quad \text{and} \quad \sup_{x_1 \in \bar{\mathcal{X}}_1} (|\mu_0(0, x_1)| + |\mu_0(1, x_1)|) \leq \bar{C}.$$

- (ii) *The functions  $f(x_1)$ ,  $\mu_0(0, x_1)$ , and  $\mu_0(1, x_1)$  are twice differentiable with bounded derivatives over  $\bar{\mathcal{X}}_1$ :*

$$\sup_{x_1 \in \bar{\mathcal{X}}_1, 1 \leq j, s \leq d} \left( |\partial_j f(x_1)| + |\partial_{j,s} f(x_1)| + |\partial_j \mu_0(0, x_1)| + |\partial_{j,s} \mu_0(0, x_1)| \right. \\ \left. + |\partial_j \mu_0(1, x_1)| + |\partial_{j,s} \mu_0(1, x_1)| \right) \leq \bar{C},$$

where  $\partial_{j,s} f(x_1)$  is the derivative of  $f(x_1)$  w.r.t.  $x_{1j}$  and  $x_{1s}$ .

- (iii) *For  $u \in \mathbb{R}^d$ ,  $\mathcal{K}(u) = \kappa(u_1) \times \dots \times \kappa(u_d)$ , where  $\kappa$  is a bounded, symmetric p.d.f. with  $\int t\kappa(t)dt = 0$  and  $\int t^2\kappa(t)dt = \nu < \infty$ . Furthermore, there exists a positive constant  $\bar{C}_\kappa$  such that  $|t|\kappa(t) \leq \bar{C}_\kappa$  for all  $t \in \mathbb{R}$ .*
- (iv) *The bandwidth  $h = h_N$  satisfies  $h = CN^{-H}$  for some  $H > 1/(4+d)$  and  $H < (1-2/q)/d$ , where  $C > 0$  and  $q$  satisfies Assumption 2.1(ii).*

For the most part, Assumption 3.1 is a collection of standard regularity conditions used in the nonparametric treatment effect estimation literature. The functions  $f(x_1)$ ,  $\mu_0(0, x_1)$ , and  $\mu_0(1, x_1)$  are required to be sufficiently smooth over  $\bar{\mathcal{X}}_1$ , the density of  $X_1$  must be bounded away from zero over the same set, and the kernel function  $\mathcal{K}$  must obey some mild restrictions, satisfied by usual choices of  $\kappa$  such as Guassian kernel or Epanechnikov kernel. In simulations and empirical studies, we use the Guassian kernel. Of course, Assumption 3.1(ii) also implies that we restrict attention to the technically more interesting case in which the distribution of  $X_1$  is continuous so that one cannot simply use sample splitting to estimate CATE at various points in the support of  $X_1$ .

The conditions imposed on the bandwidth in Assumption 3.1(iv) are motivated as follows. The restriction  $H > 1/(4+d)$  means that  $h$  converges to zero faster than the MSE-optimal

bandwidth choice; this undersmoothing condition is needed to ensure that the bias from the second stage kernel regression is asymptotically negligible. In addition, we require  $H < (1 - 2/q)/d$  to be able to use a Gaussian approximation as in [Chernozhukov, Chetverikov, and Kato \(2014, Proposition 3.2\)](#). If the outcome variable is bounded, one can set  $q = \infty$  in Assumption 2.1(ii) so that  $H < 1/d$  as in [Chernozhukov et al. \(2014, Proposition 3.1\)](#). If one only assumes  $q = 4$ , then the convergence rate must satisfy  $H \in (1/(4 + d), 1/2d)$ . For this interval to be nonempty,  $d$  can be at most 3, which is consistent with Assumption 1 in LOW.

We now state high level conditions that specify the convergence rates required of the first stage nuisance function estimators. The stated rates are linked to the bandwidth sequence  $h$  used in the second stage regression(s). More specifically, we make the following assumption about the full sample first stage estimator  $\hat{\eta}(I)$ .

**Assumption 3.2** (Full sample first stage). *Let  $\delta_{1N}, \delta_{2N}, \delta_{3N}$  and  $A_N$  be sequences of positive numbers, and  $\mathcal{G}_N^{(j)}, j \in \{0, 1, \pi\}$  be classes of real-valued functions defined on the support of  $X$  with corresponding envelope functions  $G_N^{(j)}, j \in \{0, 1, \pi\}$ . For  $\epsilon > 0$ , let  $\mathcal{N}(\mathcal{G}_N^{(j)}, \|\cdot\|, \epsilon)$  be the covering number associated with  $\mathcal{G}_N^{(j)}$  under some norm  $\|\cdot\|$  defined on  $\mathcal{G}_N^{(j)}$ .<sup>5</sup> The following conditions are satisfied.*

(i) *The estimator  $\hat{\eta}(I)$  obeys the error bounds*

$$\begin{aligned} & \sup_{x_1 \in \bar{\mathcal{X}}_1} \sum_{j=0,1} \left\| (\hat{\mu}(j, X; I) - \mu_0(j, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \\ & + \sup_{x_1 \in \bar{\mathcal{X}}_1} \left\| (\hat{\pi}(X; I) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} = O_p(\delta_{1N}) \end{aligned} \quad (5)$$

$$\sum_{j=0,1} \|\hat{\mu}(j, \cdot; I) - \mu_0(j, \cdot)\|_{\mathbb{P}, \infty} + \|\hat{\pi}(X; I) - \pi_0(X)\|_{\mathbb{P}, \infty} = O(\delta_{2N}). \quad (6)$$

(ii) *With probability approaching one,*

$$\hat{\mu}(j, \cdot; I) \in \mathcal{G}_N^{(j)}, \quad j = 0, 1, \quad \text{and} \quad \hat{\pi}(\cdot; I) \in \mathcal{G}_N^{(\pi)}$$

*where the classes of functions  $\mathcal{G}_N^{(j)}, j \in \{0, 1, \pi\}$  are such that*

$$\sup_Q \log \mathcal{N}(\mathcal{G}_N^{(j)}, \|\cdot\|_{Q, 2}, \epsilon \|G_N^{(j)}\|_{Q, 2}) \leq \delta_{3N} (\log(A_N) + \log(1/\epsilon) \vee 0), \quad j = 0, 1, \pi \quad (7)$$

---

<sup>5</sup>The covering number is the minimal number of balls with radius  $\epsilon$  needed to cover  $\mathcal{G}_N^{(j)}$ . A ball with radius  $\epsilon$  centered on  $g$  is the collection of functions  $g' \in \mathcal{G}_N^{(j)}$  with  $\|g' - g\| < \epsilon$ .

with the supremum taken over all finitely supported discrete probability measures  $Q$ .

(iii) The sequences  $\delta_{1N}$ ,  $\delta_{2N}$ ,  $\delta_{3N}$  and  $A_N$  satisfy:

$$\min(\delta_{1N}/h^{d/2}, \delta_{2N}) = o((\log(N)Nh^d)^{-1/4}), \quad \delta_{2N} = o(1), \quad (8)$$

$$\delta_{3N} \log(A_N \vee N) \log(N) \min(h^{-d}\delta_{1N}^2, \delta_{2N}^2) = o(1), \quad (9)$$

$$\text{and } \delta_{2N}\delta_{3N} \log^{1/2}(N)N^{1/q} \log(A_N \vee N) = o((Nh^d)^{1/2}). \quad (10)$$

**Remarks:**

1. Part (i) of Assumption 3.2 controls the difference between  $\eta_0$  and  $\hat{\eta}(I)$  (i.e., the estimation error) in various norms.
2. Part (ii) controls the complexity of the nuisance functions and the estimators through restrictions on the entropy of the classes  $\mathcal{G}_N^{(j)}$ .
3. It is of course part (iii) that fills parts (i) and (ii) with content through specifying the behavior of the sequences  $\delta_{1N}$ ,  $\delta_{2N}$ ,  $\delta_{3N}$  and  $A_N$ . In particular, condition (8) extends the fairly standard requirement in semiparametric settings that the first stage nuisance function estimators converge faster than  $N^{-1/4}$ ; see [Ai and Chen \(2003\)](#) and [Belloni et al. \(2017\)](#). However, in estimating  $\tau_0(x_1)$ , the second stage kernel regression relies only on observations local to  $x_1$ , and hence the relevant effective sample size is  $Nh^d$  rather than  $N$ . The extra  $\log(N)$  factor that appears in the convergence rate is the price to pay for uniform results in  $x_1$ .
4. If the first stage estimators are based on (correctly specified) parametric models, then, under standard regularity conditions,  $\hat{\eta}(I)$  converges to  $\eta_0$  at the rate of  $N^{1/2}$  both in  $L_2$  and  $L_\infty$  norm. Thus, in this case (5) and (6) both hold with  $\delta_{1N} = \delta_{2N} = N^{-1/2}$  (recall that  $\mathcal{K}_h$  is bounded). In addition, conditions (8), (9) and (10) are also easily satisfied with  $\delta_{3N} = O(1)$ , and  $A_N = O(1)$ . This is essentially the setting in LOW (with allowance for partial misspecification).

The corresponding assumption about the cross-fitting (split sample) estimator is as follows.

**Assumption 3.3** (Split sample first stage). *The split-sample first stage estimators  $\hat{\eta}(I_k^c)$ ,  $k = 1, \dots, K$  are assumed to satisfy:*

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} \left\{ \left\| (\hat{\pi}(X; I_k^c) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \right. \\ \left. \times \left\| (\hat{\mu}(j, X; I_k^c) - \mu_0(j, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \right\} = O(\delta_{1n}^2), \quad (11)$$

$$\|\pi_0(X) - \hat{\pi}_0(X; I_k^c)\|_{\mathbb{P}, \infty} + \sum_{j=0,1} \|\mu_0(j, X) - \hat{\mu}(j, X; I_k^c)\|_{\mathbb{P}, \infty} = O(\delta_{2n}), \quad (12)$$

where  $h^{-d} \delta_{1n}^2 = o((\log(n)nh^d)^{-1/2})$ , and  $\delta_{2n} = o((\log(n))^{-1})$ .

**Remarks:**

1. Because  $K$  is fixed and  $n = N/K$ ,  $\log(N)Nh^d$  and  $\log(n)nh^d$  have the same order of magnitude.
2. For the split sample estimation, there is no requirement on the entropy of the space where the estimated nuisance functions take values. This weakening of the theoretical conditions is due to the fact that, because of the cross-fitting technique, we can treat the estimators of the nuisance parameters as fixed by conditioning on the subsample of the data used for the estimation.
3. Assumption 3.3 is weaker than Assumption 3.2 in another respect. For the full sample estimation, we need the  $L_2$  bound for all three nuisance functions separately, while for the split sample estimation, we only require the bound for the product of  $L_2$  norms of two nuisance parameters. The product structure of condition (11) allows for tradeoffs between how fast  $\hat{\pi}(\cdot; I_k^c)$  versus  $\hat{\mu}(j, \cdot; I_k^c)$  converges; we will discuss this point further below. In essence, Assumption 3.3 is the local analog of Assumption 5.1(f) used by Chernozhukov et al. (2018) to estimate the *unconditional* average treatment effect via the cross-fitting (split sample) approach.

As a leading example, we will now discuss some primitive conditions under which lasso estimation of  $\eta_0$  will satisfy Assumptions 3.2 and 3.3. As mentioned in Section 2, lasso estimation is theoretically justified if the nuisance functions are approximately sparse; intuitively, this means that they can be represented (up to a small approximation error) as a linear combination of a small subset of terms chosen from a large dictionary  $b(X)$ . What distinguishes

this assumption from a parametric model specification (as in LOW) is that the small relevant subset is chosen in a data-driven way and does not need to be pre-specified by the researcher based on a combination of intuition, subject matter theory, and ad-hoc testing. Furthermore, the size  $s$  of the relevant subset (known as the “sparsity index”) can grow with the sample size  $N$ , though it is required to be substantially less than  $N$ .

To formalize the idea that the dimension  $p$  of  $b(X)$  is comparable with or larger than the sample size, we let  $p = p_N$  be a function of  $N$  and allow  $p_N$  to grow to infinity as  $N$  increases, possibly (much) faster than  $N$ . For example, one could set  $p_N = O(N^\lambda)$  for any  $\lambda > 0$ , but even  $\log(p_N) = O(N^\lambda)$  is allowed if  $\lambda$  is not too large. The linear approximation errors to the components of  $\eta_0$ , defined in display (2), will be controlled by the sparsity index  $s = s_N$ , a nondecreasing sequence of positive numbers potentially converging to infinity with  $N$ . Also needing control is the upper bound on the components of  $b(X)$ ; to this end, let  $\zeta = \zeta_N = \max_{1 \leq j \leq p_N} \|b_j(X)\|_{\mathbb{P}, \infty}$ , and note that  $\zeta$  (weakly) increases as  $p_N \rightarrow \infty$ . The following assumption formalizes the notion of approximate sparsity.

**Assumption 3.4.** *There exist sequences of coefficients  $\alpha = \alpha_N$ ,  $\beta = \beta_N$  and  $\theta = \theta_N$  so that the linear approximations defined in (2) satisfy the following conditions.*

- (i) *The number of nonzero coefficients is bounded by  $s$ , i.e.,  $\max\{\|\alpha\|_0, \|\beta\|_0, \|\theta\|_0\} \leq s$ .*
- (ii) *The approximation errors are asymptotically small in the sense that*

$$\begin{aligned} \|r_\alpha(X)\|_{\mathbb{P}, 2} + \|r_\beta(X)\|_{\mathbb{P}, 2} + \|r_\theta(X)\|_{\mathbb{P}, 2} &= O\left(\sqrt{s \log(p)/N}\right) \\ \|r_\alpha(X)\|_{\mathbb{P}, \infty} + \|r_\beta(X)\|_{\mathbb{P}, \infty} + \|r_\theta(X)\|_{\mathbb{P}, \infty} &= O\left(\sqrt{s^2 \zeta^2 \log(p)/N}\right), \end{aligned}$$

where  $s^2 \zeta^2 \log(p)/N \rightarrow 0$  (and therefore  $s \log(p)/N \rightarrow 0$ ).

Part (i) of Assumption 3.4 states that the number of nonzero coefficients in the  $b(X)$ -based linear approximations to  $\eta_0$  is at most  $s$ . Part (ii) requires that the approximation errors associated with these linear combinations asymptotically vanish both in  $L_2$  and  $L_\infty$  norm. This generally requires  $s \rightarrow \infty$ , but  $s$  needs to stay small relative to  $N$  in the sense that  $s^2 \zeta^2 \log(p)/N \rightarrow 0$ .

Given Assumption 3.4 and additional regularity conditions, results by Belloni et al. (2017) imply that conditions (5) and (6) hold with

$$\delta_{1N} = \sqrt{s \log(p \vee N)/N} \text{ and } \delta_{2N} = \sqrt{s^2 \zeta^2 \log(p \vee N)/N}. \quad (13)$$

Furthermore, [Belloni et al. \(2017\)](#) also establish (7) with  $\delta_{3N} = s$  and  $A_N = p$  for the following function classes:

$$\mathcal{G}_N^{(j)} = \{b(X)' \beta : \|\beta\|_0 \leq \ell_N s, \sup_{x \in \bar{X}} |b(x)' \beta - \mu_0(j, x)| \leq M \delta_{2N}\}, \quad j = 0, 1$$

and

$$\mathcal{G}_N^{(\pi)} = \{\Lambda(b(X)' \theta) : \|\theta\|_0 \leq \ell_N s, \sup_{x \in \bar{X}} |\Lambda(b(x)' \theta) - \pi_0(x)| \leq M \delta_{2N}\},$$

where  $\ell_N$  is some slowly diverging sequence, e.g.  $\ell_N = \log(\log(N))$  and  $M > 0$ . (As  $\pi_0(\cdot)$ ,  $\mu_0(1, \cdot)$ , and  $\mu_0(0, \cdot)$  are uniformly bounded,  $\mathcal{G}_N^{(0)}$ ,  $\mathcal{G}_N^{(1)}$ ,  $\mathcal{G}_N^{(\pi)}$  have bounded envelope functions.)

Given these results, Assumption 3.2 with first-stage lasso estimation boils down to the following conditions:

$$\begin{aligned} \min \left( \frac{s \log(p \vee N) \log^{1/2}(N)}{(Nh^d)^{1/2}}, \frac{\zeta^2 s^2 \log(p \vee N) \log^{1/2}(N) h^{d/2}}{N^{1/2}} \right) &= o(1), \\ \zeta^2 s^2 \log(p \vee N) &= o(N), \\ \min \left( \frac{s^2 \log^2(p \vee N) \log(N)}{Nh^d}, \frac{\zeta^2 s^3 \log^2(p \vee N) \log(N)}{N} \right) &= o(1), \\ \text{and } \frac{\zeta^2 s^4 \log^3(p \vee N) \log(N)}{N^{2-2/q} h^d} &= o(1). \end{aligned} \tag{14}$$

These conditions all hold if  $\frac{s^2 \log^2(p \vee N) \log(N)}{Nh^d} = o(1)$  and  $\zeta^2 s^2 \log(p \vee N) \log(N) = o(N^{1-2/q})$ . For example, if  $q = 4$ ,  $p = O(N^\lambda)$ ,  $\lambda > 0$ , and  $\zeta = O(N^{1/4})$ , then a sparsity index of order  $s = o(\sqrt{Nh^d})$  is essentially sufficient for Assumption 3.2, ignoring logarithmic factors of  $N$ .

By contrast, Assumption 3.3 holds under substantially weaker sparsity conditions. Let  $s_\pi$  and  $s_\mu$  denote the individual sparsity index sequences associated with  $\pi_0(\cdot)$  and  $\mu_0(j, \cdot)$ , respectively. Given the rates in (13), the l.h.s. of (11) is at most of order  $O(\sqrt{s_\pi} \sqrt{s_\mu} \log(p)/N)$ , as  $\mathcal{K}_h$  is a bounded function. Hence, Assumption 3.3 essentially reduces to  $\sqrt{s_\pi} \sqrt{s_\mu} \log(p)/(Nh^d) = o((\log(N)Nh^d)^{-1/2})$ . Again, setting  $p = O(N^\lambda)$ ,  $\lambda > 0$ , and ignoring the logged factors of  $N$  gives  $s_\pi s_\mu = o(Nh^d)$ . This condition is of course satisfied if  $s_\pi = s_\mu = o(\sqrt{Nh^d})$ , but there can be tradeoffs between the two sparsity indexes. For example, if  $s_\pi = O(1)$ , i.e., the propensity score essentially obeys a finite dimensional model linear in parameters, then  $s_\mu = o(Nh^d)$  is possible, i.e.,  $\mu_0(j, \cdot)$  can be a function that is substantially harder to approximate.

We are now in position to present the following fundamental representation result.

**Theorem 3.1.** (a) If Assumptions 2.1, 2.2, 3.1 and 3.2 are satisfied, then

$$\hat{\tau}(x_1) - \tau_0(x_1) = (\mathbb{P}_N - \mathbb{P}) \left[ \frac{1}{h^d f(x_1)} (\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right] + R_N(x_1)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_N(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ .

(b) If Assumptions 2.1, 2.2, 3.1 and 3.3 are satisfied, then the representation established in part (a) also holds for the  $K$ -fold cross fitting estimator  $\check{\tau}(x_1)$ , i.e.,

$$\check{\tau}(x_1) - \tau_0(x_1) = (\mathbb{P}_N - \mathbb{P}) \left[ \frac{1}{h^d f(x_1)} (\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right] + \check{R}_N(x_1)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{R}_N(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ .

Theorem 3.1 provides the linear (Bahadur) representation of the nonparametric estimators  $\hat{\tau}(x_1)$  and  $\check{\tau}(x_1)$  with uniform control of the remainder terms. It serves as a building block both for pointwise and uniform inference about  $\tau_0(x_1)$ . Starting with the former, we define

$$\sigma_N^2(x_1) = h^d \text{Var} \left( \frac{1}{h^d f(x_1)} (\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right),$$

and suppose that  $\sigma_N^2(x_1)$  satisfies:

**Assumption 3.5.** There exists some  $\underline{C} > 0$  such that  $\min_{x_1 \in \bar{\mathcal{X}}_1} \sigma_N^2(x_1) \geq \underline{C}$  for all  $N$ .

Then Theorem 3.1, together with Lyapunov's CLT, implies

$$\frac{\sqrt{Nh^d}(\hat{\tau}(x_1) - \tau_0(x_1))}{\sigma_N(x_1)} \xrightarrow{d} \mathcal{N}(0, 1) \tag{15}$$

for any fixed  $x_1 \in \bar{\mathcal{X}}_1$ . One can estimate the variance  $\sigma_N^2(x_1)$  as

$$\hat{\sigma}_N^2(x_1) = \frac{1}{Nh^d \hat{f}^2(x_1; I)} \sum_{i=1}^n (\psi(W_i, \hat{\eta}(I)) - \hat{\tau}(x_1))^2 \mathcal{K}_h^2(X_{1i} - x_1),$$

and we will show that

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\hat{\sigma}_N(x_1) - \sigma_N(x_1)| = o_p(1) \text{ and } \sup_{x_1 \in \bar{\mathcal{X}}_1} |\hat{\sigma}_N^{-1}(x_1) - \sigma_N^{-1}(x_1)| = o_p(1).$$

Of course, this means that inference in practice can proceed based on (15) with  $\hat{\sigma}_N(x_1)$  replacing  $\sigma_N(x_1)$ . Furthermore, result (15) remains valid if one uses the estimator  $\check{\tau}(x_1)$  in place of  $\hat{\tau}(x_1)$ ;

in this case  $\sigma_N^2(x_1)$  can be estimated as

$$\check{\sigma}_k^2(x_1) = \frac{1}{nh^d} \sum_{i \in I_k} \frac{1}{\hat{f}^2(x_1; I_k)} (\psi(W_i, \hat{\eta}(I_k^c)) - \check{\tau}_k(x_1))^2 \mathcal{K}_h^2(X_{1i} - x_1),$$

and

$$\check{\sigma}_N^2(x_1) = \frac{1}{K} \sum_{k=1}^K \check{\sigma}_k^2(x_1).$$

**Theorem 3.2.** *If Assumptions in Theorems 3.1 hold, then*

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\hat{\sigma}_N(x_1) - \sigma_N(x_1)| = o_p(1), \quad \sup_{x_1 \in \bar{\mathcal{X}}_1} |\hat{\sigma}_N^{-1}(x_1) - \sigma_N^{-1}(x_1)| = o_p(1),$$

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{\sigma}_N(x_1) - \sigma_N(x_1)| = o_p(1), \quad \text{and} \quad \sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{\sigma}_N^{-1}(x_1) - \sigma_N^{-1}(x_1)| = o_p(1).$$

Turning to uniform inference, one option is to construct uniform confidence bands analytically similarly to LOW. In this paper we provide an alternative method based on the multiplier bootstrap; this is the subject of the following section.

## 4 Uniform inference based on the multiplier bootstrap

We propose a computationally efficient multiplier bootstrap procedure which takes the nuisance function estimators from the first stage as given and only recomputes the nonparametric regression estimator(s) from the second stage. This step simply involves a random rescaling of the terms in the sums (3) and (4). As lasso estimation is usually time consuming, our procedure is less costly to implement than, say, a standard nonparametric bootstrap requiring new samples from the original data and recomputing the whole estimator.

To describe the procedure formally, we make the following assumption.

**Assumption 4.1.** *The random variable  $\xi$  is independent of  $W$  with  $\mathbb{E}(\xi) = \text{var}(\xi) = 1$  and its distribution has sub-exponential tails.<sup>6</sup>*

Assumption 4.1 is standard for multiplier bootstrap inference. For example, a normal random variable with unit mean and standard deviation satisfies this assumption. The bootstrap is implemented as follows:

1. Compute the first stage nuisance function estimates  $\hat{\mu}(0, x; I)$ ,  $\hat{\mu}(1, x; I)$ ,  $\hat{\pi}(x; I)$  OR  $\hat{\mu}(0, x; I_k^c)$ ,  $\hat{\mu}(1, x; I_k^c)$ ,  $\hat{\pi}(x; I_k^c)$ ,  $k = 1, \dots, K$ .
2. Draw an i.i.d. sequence  $\{\xi_i\}_{i=1}^N$  from the distribution of  $\xi$ .

---

<sup>6</sup>A random variable  $\xi$  has sub-exponential tails if  $\mathbb{P}(|\xi| > x) \leq K \exp(-Cx)$  for every  $x$  and some constants  $K$  and  $C$ .

3. Compute

$$\begin{aligned}\hat{\tau}^b(x_1) &= \frac{1}{Nh^d \hat{f}^b(x_1; I)} \sum_{i \in I} \xi_i \psi(W_i, \hat{\eta}(I)) \mathcal{K}_h(X_{1i} - x_1) \quad \text{OR} \\ \tilde{\tau}_k^b(x_1) &= \frac{1}{nh^d \hat{f}^b(x_1; I_k)} \sum_{i \in I_k} \xi_i \psi(W_i, \hat{\eta}(I_k^c)) \mathcal{K}_h(X_{1i} - x_1), \quad k = 1, \dots, K \\ \check{\tau}^b(x_1) &= \frac{1}{K} \sum_{k=1}^K \tilde{\tau}_k^b(x_1),\end{aligned}$$

where

$$\hat{f}^b(x_1; I) = \frac{1}{Nh^d} \sum_{i \in I} \xi_i \mathcal{K}_h(X_{1i} - x_1) \quad \text{and} \quad \hat{f}^b(x_1; I_k) = \frac{1}{nh^d} \sum_{i \in I_k} \xi_i \mathcal{K}_h(X_{1i} - x_1).$$

The following theorem is the bootstrap version of Theorem 3.1 and it forms the basis of our inference procedure.

**Theorem 4.1.** (a) *If Assumptions 2.1, 2.2, 3.1, 3.2 and 4.1 are satisfied, then*

$$\hat{\tau}^b(x_1) - \hat{\tau}(x_1) = (\mathbb{P}_N - \mathbb{P}) \left[ \frac{\xi - 1}{h^d f(x_1)} (\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right] + R_N^b(x_1)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_N^b(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ .

(b) *If Assumptions 2.1, 2.2, 3.1, 3.3 and 4.1 are satisfied, the representation established in part (a) also holds for  $\check{\tau}^b(x_1) - \check{\tau}(x_1)$ , i.e.,*

$$\check{\tau}^b(x_1) - \check{\tau}(x_1) = (\mathbb{P}_N - \mathbb{P}) \left[ \frac{\xi - 1}{h^d f(x_1)} (\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right] + \check{R}_N^b(x_1)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{R}_N^b(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ .

Theorem 4.1 justifies the validity of the multiplier bootstrap in implying that  $\sqrt{Nh^d}(\hat{\tau}^b(x_1) - \hat{\tau}(x_1))$  converges in distribution to the limiting distribution of  $\sqrt{Nh^d}(\hat{\tau}(x_1) - \tau_0(x_1))$  conditional on the sample path (data) with probability 1. Therefore, if Assumption 3.5 also holds, then, conditional on data,

$$\frac{\sqrt{Nh^d}(\hat{\tau}^b(x_1) - \hat{\tau}(x_1))}{\hat{\sigma}_N(x_1)} \xrightarrow{d} \mathcal{N}(0, 1), \quad (16)$$

The same statements of course hold true if  $\check{\tau}^b(x_1)$  and  $\check{\tau}(x_1)$  replaces  $\hat{\tau}^b(x_1)$  and  $\hat{\tau}(x_1)$ , respectively. In addition to pointwise inference, the uniform control of the error term  $R_N^b(\cdot)$  in Theorem 4.1 makes it possible to employ the multiplier bootstrap for uniform inference. We propose the following algorithm.

## Uniform Confidence Band Implementation Procedure

1. Compute  $\hat{\tau}(x_1)$  and  $\hat{\sigma}_N(x_1)$  for a suitably fine grid over  $\bar{\mathcal{X}}_1$ .
2. Choose the number of bootstrap replications  $B$  (say,  $B = 300$ ). Compute  $\hat{\tau}^b(x_1)$  over the same grid for  $b = 1, \dots, B$  while generating a new set of i.i.d.  $\mathcal{N}(1, 1)$  random variables  $\{\eta_i^b\}_{i=1}^N$  in each step  $b$ .
3. For  $b = 1, \dots, B$ , compute

$$M_b^{1\text{-sided}} = \sup_{x_1 \in \bar{\mathcal{X}}_1} \frac{\sqrt{Nh^d}(\hat{\tau}^b(x_1) - \hat{\tau}(x_1))}{\hat{\sigma}_N(x_1)}, \quad M_b^{2\text{-sided}} = \sup_{x_1 \in \bar{\mathcal{X}}_1} \frac{\sqrt{Nh^d}|\hat{\tau}^b(x_1) - \hat{\tau}(x_1)|}{\hat{\sigma}_N(x_1)},$$

where the supremum is approximated by the maximum over the chosen gridpoints.

4. Given a confidence level  $1 - \alpha$ , find the empirical  $(1 - \alpha)$  quantile of the sets of numbers  $\{M_b^{1\text{-sided}} : b = 1, \dots, B\}$  and  $\{M_b^{2\text{-sided}} : b = 1, \dots, B\}$ . Denote these quantiles as  $\widehat{C}_\alpha^{1\text{-sided}}$  and  $\widehat{C}_\alpha^{2\text{-sided}}$ , respectively.
5. The uniform confidence bands are constructed as

$$\begin{aligned} I_L &= \left\{ \left( \hat{\tau}(x_1) - \widehat{C}_\alpha^{1\text{-sided}} \frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}}, \infty \right) : x_1 \in \bar{\mathcal{X}}_1 \right\}, \\ I_R &= \left\{ \left( -\infty, \hat{\tau}(x_1) + \widehat{C}_\alpha^{1\text{-sided}} \frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}} \right) : x_1 \in \bar{\mathcal{X}}_1 \right\}, \\ I_2 &= \left\{ \left( \hat{\tau}(x_1) - \widehat{C}_\alpha^{2\text{-sided}} \frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}}, \hat{\tau}(x_1) + \widehat{C}_\alpha^{2\text{-sided}} \frac{\widehat{\sigma}_N(x_1)}{\sqrt{Nh^d}} \right) : x_1 \in \bar{\mathcal{X}}_1 \right\}. \end{aligned}$$

The following theorem formally states the asymptotic validity of the confidence regions proposed above.

**Theorem 4.2.** *If Assumptions 2.1, 2.2, 3.1, 3.2, 3.5 and 4.1 are satisfied, then*

$$\lim_{N \rightarrow \infty} \mathbb{P}(\tau_0 \in I_L) = \lim_{N \rightarrow \infty} \mathbb{P}(\tau_0 \in I_R) = \lim_{N \rightarrow \infty} \mathbb{P}(\tau_0 \in I_2) = 1 - \alpha.$$

### Remarks:

1. Theorem 4.2 states that the random confidence bands  $I_R$ ,  $I_L$  and  $I_2$  contain the *entire* function  $\tau_0$  with the prescribed probability  $1 - \alpha$  in large samples.
2. If the grid in step 1. is chosen to be a single point  $x_1$ , then the algorithm provides pointwise confidence intervals  $I_L(x_1)$ ,  $I_R(x_1)$  and  $I_2(x_2)$ .

3. One can construct uniform confidence bands for  $\tau_0$  based on the cross-fitting estimator  $\check{\tau}$  following the exact same steps as above; of course, one needs to replace  $\hat{\tau}$ ,  $\hat{\tau}^b$  and  $\hat{\sigma}_N$  with  $\check{\tau}$ ,  $\check{\tau}^b$  and  $\check{\sigma}_N$ , respectively.

## 5 Monte Carlo simulations

In this section, we investigate the finite sample properties of the proposed high dimensional CATE estimators using Monte Carlo experiments. The goal is to evaluate estimation accuracy and the validity of the proposed inference procedures. In accordance with the theoretical setup of the paper, we do not assume that either  $\pi_0$  or  $\mu_0$  are known by the researcher, but rather we use variable selection techniques (specifically, post-lasso), to estimate these functions.<sup>7</sup> We generate an initial pool of covariates whose size  $p$  is comparable with, or even exceeds, the sample size  $N$ . Nevertheless, only a small subset of these variables forms an important component of the data generating process (DGP), i.e., a sparsity condition is satisfied. We consider two cases: in the first, the DGP is strictly sparse (i.e., the number of covariates that actually enter the DGP is fixed and small), and in the second we emulate an approximately sparse DGP (i.e., all  $p$  covariates are relevant to some extent but only a few are truly important). Throughout this section we repeat the simulations 5000 times. For the cross-fitting estimator, we set  $K = 4$  as suggested by Chernozhukov et al. (2018).

These simulation designs are meant to imitate a data-rich modeling environment, which is a common scenario in modern applications. For instance, in our empirical study of maternal smoking and birth we observe a large number of personal characteristics (e.g., mothers age, education, ethnicity, etc.), and other medical variables that are potential predictors of birth weight as well as the treatment status. It is a priori unknown how strongly and in what form  $\mu_0$  and  $\pi_0$  depend on these variables. Thus, preliminary estimation of these functions by an appropriate selection method that chooses the most important covariates and their functional form is a crucial component of estimating CATE. The asymptotic theory presented in this paper accommodates the use of such methods, and the simulations show that the subsequent kernel regression targeting CATE has good statistical properties even when the sample size is

---

<sup>7</sup>Post-lasso estimation starts with the ordinary lasso, i.e., a penalized least squares (or logit) regression. The regressors with non-zero estimated coefficients are retained, and the regression is re-estimated by OLS (or standard logit). The theory presented in Sections 3 and 4 covers post-lasso estimation as well.

relatively small.

## 5.1 Data-generating process

We consider the following two DGPs in the simulation studies.

### DGP 1 (strict sparsity):

This design is based on LOW. We specify the following linear model for the potential outcomes:

$$Y(1) = 10 + \sum_{k=1}^p \beta_k X_k + \epsilon, \quad (17)$$

and we set  $Y(0) = 0$ . The covariates  $\mathbf{X} = (X_1, X_2, \dots, X_p)'$  are drawn from the  $N(0, I_p)$  distribution, where  $I_p$  is the  $p$ -dimensional identity matrix. The error  $\epsilon$  follows standard normal distribution and is independent of  $\mathbf{X}$ . We consider  $p = 100, 200, 500$ , and sample sizes  $N = 500, 1000$ . We set  $\beta_i = 1$ , for  $i = 1, 2, \dots, p_1$ , and 0 for  $i > p_1$ . Therefore, only the first  $p_1$  covariates are actually present in the conditional mean function. In our simulations we set  $p_1 = 4$ .

The treatment status is determined by

$$D = \mathbf{1} \left\{ \Lambda \left( \sum_{k=1}^p \gamma_k X_k \right) > U \right\}, \quad (18)$$

where  $U \sim \text{unif}(0, 1)$ , independent of  $(\mathbf{X}, \epsilon)$ , and  $\Lambda(\cdot)$  is the logistic link function. The observed outcome is then  $Y = DY(1)$ . We set  $\gamma_i = 0.5$ , for  $i = 1, 2, \dots, p_1$ , and 0 for  $i > p_1$ . We use the same value of  $p_1$  as in the conditional mean function. The estimation target is  $CATE(x_1) = E[Y(1)|X_1 = x_1]$ , given by

$$CATE(x_1) = 10 + x_1. \quad (19)$$

### DGP 2 (approximate sparsity):

The strict sparsity assumption in DGP 1 may be too strong in applications. Given the functional forms specified above, approximate sparsity means that there are many small but nonzero coefficients  $\beta_k$  and  $\gamma_k$ . In DGP 2 we mimic this situation using a dwindling coefficients

setup as in [Belloni et al. \(2017\)](#). Similarly to DGP 1, we define

$$\begin{aligned}
Y(1) &= \sum_{k=1}^p \beta_k X_k + \epsilon, \\
D &= \mathbf{1} \left\{ \Lambda \left( \sum_{k=1}^p \gamma_k X_k \right) > U \right\},
\end{aligned} \tag{20}$$

where  $\epsilon \sim N(0, 1)$ ,  $U \sim \text{unif}(0, 1)$ , and  $\epsilon$  and  $U$  are independent. We also set  $Y(0) = 0$ . Nevertheless, the distribution of  $X$  and the coefficients are generated differently. Specifically, we draw  $X_1$  from the  $N(0, 1)$  distribution independently of the other covariates. The rest of the covariates are generated as  $X_i \sim N(0, \Sigma)$  with  $\Sigma_{kj} = (0.5)^{|j-k|}$ , for  $i = 2, 3, \dots, p$ . We consider  $p = 100, 200$  and  $500$  and sample sizes  $N = 500, 1000$ .

The model coefficients are specified in the following way. Let  $\gamma_k = c_d \theta_{0,k}$ ,  $\beta_k = c_y \theta_{0,k}$ , where  $\theta_0$  is a  $p \times 1$  vector such that  $\theta_{0,k} = (1/k)^2$  for  $k = 1, \dots, p$ ,  $c_d$  and  $c_y$  are scalars that control the strength of the relationship between the controls, the outcome, and the treatment variable. We use several different combinations of  $c_d$  and  $c_y$ , setting  $c_d = \sqrt{\frac{(\pi^2/3)R_d^2}{(1-R_d^2)\theta_0'\Sigma\theta_0}}$  and  $c_y = \sqrt{\frac{R_y^2}{(1-R_y^2)\theta_0'\Sigma\theta_0}}$  for  $R_d^2 = R_y^2 \in \{0.1, 0.5\}$ .<sup>8</sup> The estimation target is again  $CATE(x_1) = E[Y(1)|X_1 = x_1]$ , given by

$$CATE(x_1) = \beta_1 x_1. \tag{21}$$

The computation procedure is the following: we first estimate  $\mu_0$  and  $\pi_0$  using the post lasso method of [Belloni et al. \(2014b\)](#). After this step,  $\hat{\beta}$  and  $\hat{\gamma}$  are sparse as some (most) coefficients are estimated as exact zeros. Thus, this step also serves as the model selection procedure, which is the advantage of a lasso type approach. In the second step, we substitute the estimated values of  $\mu_0(j, X_i)$  and  $\pi_0(X_i)$  into the score function  $\psi$  and run kernel regressions of  $\psi(W_i, \hat{\eta})$  on  $X_{1i}$  at various evaluation points  $x_1$ . In particular, we use a grid of 201 equally spaced points over the interval  $\mathcal{X}_1 = [-1, 1]$ , and estimate CATE at these points. We use the Gaussian kernel throughout and the bandwidth is set as  $h_N = \hat{h} \times N^{1/5} \times N^{-2/7}$ , where  $\hat{h}$  is a commonly used optimal bandwidth in the literature (e.g., the plug-in method of [Ruppert, Sheather, and Wand \(1995\)](#)).<sup>9</sup>

<sup>8</sup>The formulas for  $c_d$  and  $c_y$  ensure that the regressions under (20) attain the given (pseudo)  $R^2$  values. A higher value of  $R^2$  means that the regressors are more informative about  $D$  and  $Y$ .

<sup>9</sup>E.g., in univariate case, let  $Y$  be the dependent variable and  $X$  the regressor. Letting  $a = \min_{1 \leq i \leq N} X_i$ ,

We examine the coverage probability of the (2-sided) uniform confidence band for  $\text{CATE}(x_1)$  over the grid described above. The nominal coverage probabilities  $(1 - \alpha)$  are the standard 99%, 95% and 90%. We compute the empirical coverage (EMP), the mean critical value (Mcri), and the standard deviation of critical value (Sdcri).<sup>10</sup> We also compute the coverage probabilities of the confidence band based on the critical values computed by the Gumbel approximation (using the full sample) as in LOW. In addition, we report various statistics describing the properties of the estimates for  $x_1 \in \{-1, -0.5, 0, 0.5, 1\}$ . These are the bias (BIAS), standard deviation (SD), the average estimated standard error of  $\widehat{\text{CATE}}(x_1)$  (ASE), and the root mean squared error (RMSE).

## 5.2 Simulation Results

Tables 1-2 show the simulation results for DGP 1 (strict sparsity). In particular, Table 1 shows that the proposed uniform confidence band has very good finite sample coverage properties in this setting. For example, for  $p = 200$  and  $N = 500$ , the uniform confidence band for the cross-fitting method covers  $\tau_0$  over the  $[-1, 1]$  interval with 99.1%, 93.7% and 86.6% probability given the nominal rates 99%, 95% and 90%, respectively. When the sample size increases to  $N = 1000$ , the corresponding empirical probabilities further improve to 99.3%, 95.5% and 90.3%, respectively. The coverage rates are similarly accurate in all other cases, including for the full-sample estimator. Thus, our approach is capable of providing sound inference for the unknown  $\tau_0$ .

The importance of conducting uniform inference properly is apparent from the average size of the critical values. For example, the pointwise critical value for  $1 - \alpha = 95\%$  is 1.96, while the critical value for the corresponding uniform confidence band is around 2.75. Hence, using pointwise confidence bands to draw inferences about the global properties of the function  $\tau_0$  can be misleading. The Gumbel approximation, on the other hand, is shown to be too conservative to provide any useful inference in practice.

Table 2 displays various statistics describing finite sample estimation accuracy. Both vari-

---

$b = \max_{1 \leq i \leq N} X_i$ , the bandwidth is computed as  $\hat{h} = (\frac{1}{2\sqrt{\pi}})^{1/5} [\frac{\hat{\sigma}^2(b-a)}{\hat{\theta}N}]^{1/5}$ , where  $\hat{\sigma}^2$  is the estimated variance of the local linear regression residuals, and  $\hat{\theta} = \frac{1}{N} \sum_{i=1}^N (\hat{m}^{(2)}(x_i))^2$ , where  $\hat{m}^{(2)}$  is the estimate of the second derivative of the regression function from a local cubic regression.

<sup>10</sup>“Critical value” refers to the statistic  $\widehat{C}_\alpha^{2\text{-sided}}$ , whose computation is discussed in Section 4.

Table 1: DGP 1, high dimensional CATE estimation

p	N	Conf. level	cross-fitting			full sample			
			EMP	Mcri	Sdcri	EMP	Mcri	Sdcri	Gumbel
p=100	N=500	99%	0.993	3.290	0.083	0.996	3.292	0.084	1.000
		95%	0.949	2.753	0.098	0.950	2.752	0.099	0.998
		90%	0.878	2.474	0.106	0.872	2.476	0.109	0.996
	N=1000	99%	0.985	3.294	0.084	0.982	3.296	0.085	1.000
		95%	0.947	2.758	0.099	0.946	2.757	0.101	0.998
		90%	0.908	2.479	0.108	0.910	2.481	0.111	0.992
p=200	N=500	99%	0.991	3.292	0.091	0.990	3.293	0.093	0.998
		95%	0.937	2.754	0.107	0.932	2.753	0.109	0.998
		90%	0.866	2.476	0.115	0.863	2.477	0.120	0.996
	N=1000	99%	0.993	3.287	0.079	0.996	3.290	0.080	1.000
		95%	0.955	2.748	0.091	0.958	2.749	0.094	0.998
		90%	0.903	2.474	0.102	0.904	2.473	0.104	0.994
p=500	N=500	99%	0.983	3.292	0.094	0.980	3.295	0.097	0.998
		95%	0.940	2.753	0.112	0.938	2.755	0.114	0.998
		90%	0.887	2.477	0.123	0.884	2.480	0.125	0.996
	N=1000	99%	0.991	3.290	0.079	0.992	3.293	0.080	1.000
		95%	0.947	2.751	0.093	0.946	2.753	0.095	1.000
		90%	0.896	2.475	0.102	0.894	2.478	0.104	0.998

*Notes:* The nominal coverage probabilities that we consider are 99%, 95% and 90%. We compute the empirical coverage (“EMP”), the mean critical value (“Mcri”), and the standard deviation of critical value (“Sdcri”). We also compute the coverage probabilities of the confidence band based on the critical values computed by the Gumbel approximation (“Gumbel”).

ants of the estimator exhibit small bias, and as the sample size increases, RMSE improves.<sup>11</sup> The estimated standard errors show a small downward bias. Comparing the two variants, the differences between the RMSE values are small, but it is always the cross fitting approach that comes out slightly better in each setting.

Tables 3-6 show the simulation results for DGP 2 (approximate sparsity), for  $R^2 = 0.1$  and 0.5. The overall performance of the proposed estimators is satisfactory both in terms of empirical coverage and estimation accuracy. For example, in Table 3 ( $R_d^2 = R_y^2 = 0.1$ ), for  $p = 200$ ,  $N = 500$ , the empirical coverage rate of the uniform confidence band for the cross-fitting estimator is 97.9%, 92.7% and 86.9% given the respective nominal coverage rates 99%, 95% and 90%. As sample size increases to  $n = 1000$ , the corresponding figures improve to 98.5%, 93.4% and 87.7%, respectively. In most cases, the difference between the nominal and

<sup>11</sup>Our results show that the post-lasso method has good performance in selecting the important variables in the  $\mu_0$  and  $\pi_0$  functions. Due to space limitations, the model selection results (i.e., information about the estimated  $\beta$  and  $\gamma$  coefficients) are not shown here; these results are available upon request to the authors.

Table 2: DGP1, high dimensional CATE estimation

		HD-DR								
		cross-fitting				full sample				
	At $x=$	BIAS	SD	ASE	RMSE	BIAS	SD	ASE	RMSE	
p=100	N=500	-1	0.006	0.269	0.257	0.268	0.005	0.276	0.262	0.276
		-0.5	0.005	0.220	0.209	0.223	0.004	0.229	0.213	0.229
		0	0.015	0.211	0.189	0.213	0.021	0.217	0.193	0.218
		0.5	0.008	0.217	0.190	0.216	0.009	0.222	0.197	0.222
		1	-0.007	0.244	0.226	0.246	-0.006	0.252	0.230	0.252
	N=1000	-1	0.008	0.192	0.180	0.191	0.009	0.198	0.186	0.198
		-0.5	0.008	0.160	0.146	0.162	0.009	0.167	0.151	0.168
		0	-0.005	0.152	0.133	0.151	-0.005	0.157	0.139	0.157
		0.5	0.002	0.156	0.139	0.155	0.001	0.162	0.142	0.162
		1	-0.013	0.170	0.159	0.171	-0.015	0.176	0.164	0.176
p=200	N=500	-1	0.017	0.268	0.264	0.297	0.020	0.275	0.269	0.306
		-0.5	-0.010	0.224	0.213	0.228	-0.009	0.231	0.219	0.231
		0	0.008	0.212	0.192	0.213	0.009	0.218	0.197	0.218
		0.5	0.010	0.217	0.194	0.218	0.011	0.222	0.199	0.221
		1	-0.007	0.233	0.224	0.235	-0.008	0.241	0.231	0.241
	N=1000	-1	0.006	0.196	0.181	0.199	0.005	0.206	0.187	0.206
		-0.5	0.007	0.158	0.144	0.159	0.008	0.163	0.150	0.163
		0	-0.002	0.134	0.132	0.135	-0.001	0.139	0.137	0.138
		0.5	0.002	0.136	0.134	0.137	0.002	0.140	0.140	0.141
		1	0.008	0.163	0.158	0.165	0.009	0.170	0.163	0.170
p=500	N=500	-1	-0.004	0.281	0.262	0.283	-0.004	0.294	0.270	0.293
		-0.5	0.005	0.242	0.215	0.245	0.006	0.252	0.223	0.252
		0	-0.008	0.205	0.199	0.207	-0.011	0.214	0.203	0.215
		0.5	-0.008	0.215	0.201	0.217	-0.010	0.224	0.206	0.224
		1	-0.005	0.247	0.231	0.249	-0.007	0.256	0.239	0.256
	N=1000	-1	0.003	0.197	0.181	0.199	0.002	0.205	0.189	0.205
		-0.5	-0.003	0.154	0.146	0.157	-0.002	0.162	0.153	0.162
		0	-0.004	0.147	0.136	0.149	-0.005	0.152	0.141	0.152
		0.5	0.006	0.140	0.137	0.142	0.007	0.149	0.143	0.149
		1	0.003	0.171	0.162	0.173	0.004	0.180	0.167	0.180

*Notes:* In both tables, we estimate  $\text{CATEF}(x_1)$  for  $x_1 \in \{-1, -0.5, 0, 0.5, 1\}$  and compute the mean bias (“BIAS”), standard deviation (“SD”), the average of standard error for  $\text{CATEF}(x_1)$  (“ASE”), and the root mean squared error (“RMSE”).

the empirical coverage rate is somewhat larger than under strict sparsity, but inference is still very precise.

Turning to Table 4 (estimation accuracy for  $R^2 = 0.1$ ), one observes some of the same patterns as in the case of DGP 1. Specifically, the RMSE decreases with the sample size, and is slightly but consistently smaller for the cross-fitting estimator than for the full sample estimator. The estimated standard error is also slightly downward biased. Compared with

Table 3: DGP 2,  $R^2 = 0.1$ , HD CATE estimation

p	n	Conf.level	cross-fitting			full sample			
			EMP	Mcri	Sdcri	EMP	Mcri	Sdcri	Gumbel
p=100	N=500	99%	0.990	3.291	0.090	0.988	3.292	0.091	1.000
		95%	0.935	2.753	0.105	0.932	2.752	0.107	1.000
		90%	0.874	2.475	0.116	0.870	2.476	0.118	0.996
	N=1000	99%	0.987	3.295	0.090	0.984	3.296	0.092	1.000
		95%	0.936	2.758	0.106	0.934	2.757	0.108	1.000
		90%	0.887	2.481	0.117	0.884	2.482	0.119	0.998
p=200	N=500	99%	0.979	3.297	0.085	0.978	3.298	0.087	1.000
		95%	0.927	2.758	0.101	0.924	2.759	0.103	1.000
		90%	0.869	2.485	0.112	0.866	2.484	0.113	0.992
	N=1000	99%	0.985	3.292	0.079	0.982	3.293	0.081	1.000
		95%	0.934	2.755	0.093	0.932	2.754	0.096	0.998
		90%	0.877	2.479	0.104	0.874	2.478	0.106	0.996
p=500	N=500	99%	0.987	3.301	0.087	0.986	3.300	0.089	1.000
		95%	0.935	2.762	0.102	0.932	2.761	0.105	1.000
		90%	0.874	2.486	0.114	0.871	2.487	0.116	0.992
	N=1000	99%	0.992	3.297	0.095	0.994	3.298	0.097	1.000
		95%	0.945	2.758	0.112	0.944	2.759	0.115	1.000
		90%	0.890	2.485	0.123	0.888	2.484	0.126	0.996

Notes: See notes to Table 1.

DGP 1, the estimators have somewhat larger bias.

Finally, Tables 5 and 6 show the results for DGP 2 with  $R^2 = 0.5$ . Generally speaking, less sparsity brings about a deterioration in finite sample performance both in terms of coverage rates and estimation accuracy. While inference is still reasonably accurate, the RMSE of both estimators increase substantially, in some cases by as much as 50 percent or more. This is likely due to the fact that variable selection, i.e., obtaining a parsimonious approximation to the DGP, is more difficult when the DGP is less sparse. In case of  $R^2 = 0.5$ , the model coefficients are larger and shrink to zero slower compared to the case of  $R^2 = 0.1$ . Thus, it is hard to single out a handful of variables as the most important ones, and there can be substantial small sample variation in the set of regressors selected by the lasso. (The  $R^2$  associated with DGP 1 is greater than 0.5, which shows that it is not the high  $R^2$  itself that is the root of the problem but the increased difficulty of model selection in finite samples.) This finding is consistent with the theory presented in Section 3.

Table 4: DGP 2,  $R^2 = 0.1$ , HD CATE estimation

		HD-DR								
		cross-fitting				full sample				
	At $x=$	BIAS	SD	ASE	RMSE	BIAS	SD	ASE	RMSE	
p=100	N=500	-1	-0.038	0.192	0.172	0.197	-0.040	0.200	0.179	0.204
		-0.5	-0.016	0.159	0.132	0.160	-0.018	0.166	0.138	0.167
		0	0.006	0.113	0.116	0.117	0.005	0.120	0.121	0.120
		0.5	0.020	0.110	0.111	0.114	0.023	0.117	0.117	0.122
		1	0.021	0.131	0.124	0.136	0.024	0.139	0.131	0.146
	N=1000	-1	-0.035	0.134	0.121	0.136	-0.037	0.140	0.126	0.145
		-0.5	-0.021	0.102	0.092	0.105	-0.023	0.108	0.097	0.110
		0	-0.005	0.093	0.079	0.094	-0.004	0.098	0.084	0.098
		0.5	0.017	0.084	0.078	0.087	0.020	0.089	0.084	0.091
		1	0.012	0.098	0.088	0.106	0.014	0.103	0.094	0.113
p=200	N=500	-1	-0.041	0.139	0.124	0.145	-0.043	0.146	0.128	0.152
		-0.5	-0.020	0.102	0.093	0.107	-0.023	0.109	0.098	0.112
		0	-0.007	0.080	0.081	0.082	-0.008	0.085	0.085	0.086
		0.5	0.021	0.081	0.079	0.083	0.023	0.085	0.084	0.088
		1	0.048	0.094	0.090	0.099	0.052	0.101	0.095	0.114
	N=1000	-1	-0.037	0.136	0.121	0.141	-0.039	0.142	0.126	0.147
		-0.5	-0.014	0.097	0.091	0.099	-0.015	0.102	0.096	0.103
		0	-0.006	0.082	0.079	0.084	-0.007	0.088	0.084	0.088
		0.5	0.014	0.078	0.077	0.080	0.015	0.083	0.083	0.085
		1	0.039	0.093	0.091	0.102	0.046	0.099	0.094	0.109
p=500	N=500	-1	-0.034	0.218	0.201	0.220	-0.036	0.223	0.204	0.226
		-0.5	-0.014	0.153	0.148	0.155	-0.015	0.159	0.153	0.159
		0	0.013	0.124	0.122	0.125	0.014	0.132	0.129	0.133
		0.5	0.022	0.134	0.120	0.136	0.025	0.141	0.127	0.143
		1	0.043	0.141	0.137	0.147	0.048	0.145	0.140	0.153
	N=1000	-1	-0.031	0.221	0.198	0.224	-0.033	0.228	0.206	0.230
		-0.5	-0.020	0.183	0.149	0.187	-0.022	0.191	0.154	0.192
		0	0.008	0.122	0.124	0.125	0.009	0.128	0.129	0.128
		0.5	0.019	0.121	0.118	0.124	0.022	0.129	0.126	0.131
		1	0.039	0.135	0.133	0.139	0.041	0.141	0.139	0.147

*Note.* See note to Table 2.

## 6 Empirical application

In this section, we employ the proposed high dimensional CATE estimators in analyzing the average effect of maternal smoking on birth weight while allowing for virtually unrestricted treatment effect heterogeneity conditional on the mother's age. Birth weight has been associated with health and human capital development throughout life (Black, Devereux, and Salvanes (2007), Almond and Currie (2011)), and maternal smoking is considered to be the most important preventable cause of low birth weight (Kramer, 1987). In recent studies, Abre-

Table 5: DGP 2,  $R^2 = 0.5$ , HD CATE estimation

p	N	Conf.level	cross-fitting			full sample			
			EMP	Mcri	Sdcri	EMP	Mcri	Sdcri	Gumbel
p=100	N=500	99%	0.981	3.288	0.089	0.978	3.289	0.090	1.000
		95%	0.925	2.749	0.104	0.920	2.748	0.106	0.998
		90%	0.861	2.471	0.113	0.856	2.472	0.116	0.998
	N=1000	99%	0.984	3.316	0.106	0.981	3.315	0.109	1.000
		95%	0.934	2.781	0.125	0.932	2.780	0.128	1.000
		90%	0.865	2.508	0.138	0.862	2.507	0.141	0.984
p=200	N=500	99%	0.964	3.295	0.093	0.959	3.294	0.096	0.998
		95%	0.920	2.754	0.110	0.918	2.755	0.113	0.994
		90%	0.863	2.480	0.122	0.858	2.479	0.124	0.990
	N=1000	99%	0.970	3.302	0.101	0.965	3.303	0.104	1.000
		95%	0.928	2.766	0.119	0.924	2.765	0.122	0.996
		90%	0.865	2.491	0.131	0.860	2.490	0.134	0.984
p=500	N=500	99%	0.963	3.296	0.098	0.959	3.295	0.102	0.998
		95%	0.920	2.758	0.115	0.915	2.757	0.119	0.996
		90%	0.856	2.478	0.127	0.851	2.479	0.131	0.992
	N=1000	99%	0.969	3.307	0.101	0.966	3.306	0.104	1.000
		95%	0.930	2.770	0.119	0.924	2.771	0.122	0.998
		90%	0.867	2.495	0.130	0.863	2.496	0.134	0.994

*Note.* See note to Table 1.

vaya et al. (2015) and LOW both explored this causal relationship using the CATE approach, and found different degrees of heterogeneity by age. Using observations from 3,754 white mothers in Pennsylvania, LOW found that the CATE of smoking is decreasing from 17 to around 29 years of age but they differ from Abrevaya et al. (2015) in that the contrast between young and 30-year-old mothers is still not large.<sup>12</sup>

Our study improves on these previous investigations by considering a much larger pool of covariates and explicitly incorporating a variable selection mechanism into the estimation. This initial pool consists of a vector  $X$  of raw covariates as well as technical regressors (powers and interactions) to account for the fact that the functional form of  $\pi_0$  and  $\mu_0$  is unknown. By contrast, Abrevaya et al. (2015) assume that a low dimensional parametric model (known up to its coefficients) is correctly specified for  $\pi_0$ , while LOW assume that either  $\pi_0$  or  $\mu_0$  obeys such a model. While we still assume that  $\pi_0$  and  $\mu_0$  are sparse functions, we let a data-driven procedure (lasso) select the most relevant regressors.

<sup>12</sup>As the smoking effect is negative, 'decreasing' means that the detrimental effects of smoking become stronger with age.

Table 6: DGP 2,  $R^2 = 0.5$ , HD CATE estimation

		HD-DR								
		cross-fitting				full sample				
	At $x=$	BIAS	SD	ASE	RMSE	BIAS	SD	ASE	RMSE	
p=100	N=500	-1	-0.044	0.325	0.256	0.352	-0.048	0.331	0.262	0.359
		-0.5	-0.025	0.181	0.163	0.185	-0.028	0.187	0.169	0.193
		0	0.008	0.122	0.120	0.123	0.009	0.128	0.124	0.128
		0.5	0.041	0.103	0.107	0.124	0.043	0.109	0.111	0.131
		1	0.055	0.108	0.112	0.173	0.060	0.114	0.117	0.181
	N=1000	-1	-0.033	0.229	0.202	0.257	-0.034	0.234	0.208	0.261
		-0.5	-0.022	0.138	0.121	0.148	-0.025	0.143	0.128	0.153
		0	-0.006	0.090	0.091	0.091	-0.005	0.094	0.094	0.094
		0.5	0.026	0.083	0.079	0.102	0.032	0.088	0.084	0.113
		1	0.028	0.089	0.083	0.160	0.031	0.093	0.089	0.169
p=200	N=500	-1	-0.051	0.346	0.268	0.366	-0.053	0.359	0.275	0.375
		-0.5	-0.035	0.197	0.168	0.201	-0.037	0.203	0.176	0.209
		0	0.011	0.128	0.124	0.130	0.012	0.136	0.129	0.136
		0.5	0.065	0.108	0.109	0.129	0.070	0.116	0.114	0.136
		1	0.068	0.118	0.115	0.174	0.073	0.124	0.120	0.182
	N=1000	-1	-0.030	0.221	0.192	0.260	-0.032	0.229	0.198	0.268
		-0.5	-0.026	0.142	0.118	0.154	-0.029	0.149	0.125	0.162
		0	0.007	0.091	0.089	0.092	0.008	0.095	0.091	0.095
		0.5	0.033	0.080	0.078	0.101	0.037	0.084	0.081	0.107
		1	0.038	0.091	0.082	0.157	0.042	0.094	0.087	0.162
p=500	N=500	-1	-0.058	0.343	0.277	0.374	-0.061	0.351	0.285	0.381
		-0.5	-0.040	0.201	0.175	0.213	-0.042	0.209	0.182	0.220
		0	0.012	0.141	0.125	0.143	0.014	0.147	0.132	0.147
		0.5	0.062	0.110	0.112	0.125	0.066	0.114	0.116	0.132
		1	0.096	0.128	0.116	0.188	0.101	0.133	0.121	0.194
	N=1000	-1	-0.039	0.235	0.202	0.261	-0.041	0.240	0.208	0.269
		-0.5	-0.036	0.149	0.125	0.162	-0.037	0.153	0.130	0.167
		0	0.014	0.096	0.090	0.097	0.015	0.101	0.094	0.101
		0.5	0.044	0.082	0.079	0.110	0.047	0.087	0.082	0.115
		1	0.035	0.092	0.085	0.163	0.039	0.096	0.089	0.169

*Note.* See note to Table 2.

## 6.1 Data Description

We start with the same data set as [Abrevaya et al. \(2015\)](#), composed of vital statistics collected by the North Carolina State Center Health Services, and extract the records of first-time mothers<sup>13</sup> between 1988 and 2002. The variables include whether the mother smokes (the treatment dummy), the baby’s birth weight (the main outcome variable, measured in grams),

<sup>13</sup>The motivation for focusing on first time mothers is discussed in [Abrevaya et al. \(2015\)](#). In effect, the restricted sample enables more credible identification of the causal effect, as there cannot be uncaptured feedback from the previous birth experience to the current one.

the parents’ socio-economic information, such as age, education, income, race, etc., as well as the mothers’ medical and health records. The dataset includes 45 raw covariates and 591,547 observations in total. Table 7 summarizes the most important pre-treatment covariates in the data set.<sup>14</sup>

Table 7: Variable definitions

	<b>Name</b>	<b>Type</b>	<b>Description</b>
<b>Outcome Variable</b>	bweight	real number	birth weight(g)
<b>Treatment</b>	smoke	dummy	Whether mother smokes or not?
	mage	real number*	mother’s age
	meduc	integer	mother’s years of schooling
<b>Parents</b>	fage	integer	father’s age
<b>Basic</b>	feduc	integer	father’s years of schooling
<b>Info</b>	fagemiss	integer	Whether or not father’s age is missing?
	married	dummy	Whether or not mother is married?
<b>Covariates</b>	popdens	real number	population density in mother’s zip code (units/km <sup>2</sup> )
	prenatal	integer	month of first prenatal visit (=10 if prenatal care is foregone)
<b>Mothers’</b>	pren.visits	integer	number of prenatal visits
<b>Medical</b>	terms	integer	previous (terminated) pregnancies
<b>Care</b>	amnio	dummy	Did mother take amniocentesis?
<b>&amp;</b>	anemia	dummy	Did mother suffer from anemia?
<b>Health</b>	diabetes	dummy	Did mother suffer from gestational diabetes?
<b>Status</b>	hyperpr	dummy	Did mother suffer from hypertension?
	ultra	dummy	Did mother take ultra sound exams?
	male	dummy	Whether or not baby is male?
<b>Others</b>	drink	dummy	mother’s alcohol use
	by88-02	dummy	13 birth year dummies (from 1988 to 2002)

\*Note: mother’s age is originally recorded as an integer but for the purposes of this exercise we add a uniform [-1,1] random number to this value to make it a continuous variable.

## 6.2 High Dimensional CATE Estimation

In this section we estimate the CATE of maternal smoking on the baby’s birth weight with mother’s age as the conditioning variable. Following [Abrevaya et al. \(2015\)](#) and LOW,

<sup>14</sup>We drop some covariates from the analysis for various reasons. For example, the mother’s weight gain during pregnancy is arguably not a pre-treatment variable. Or the Kessner index of prenatal care is basically a function of the number of prenatal visits and the timing of the first visit.

we estimate CATE separately for black mothers and white mothers. We only report the estimation results for white mothers in this section; the results for the black mothers are available upon request. The dependent variable  $Y$  is the baby’s birth weight measured in grams. The treatment dummy  $D$  takes on the value 1 if the mother smokes and 0 otherwise. We start from the set of variables displayed in Table 7, and construct an even larger dictionary  $b(X)$  by adding polynomial terms to account for the unknown form of the nuisance functions in a flexible way. Specifically, we include, up to degree 3, the powers and interaction terms of key dummy variables and continuous/integer covariates. We then end up with 792 covariates in total.

With such a large set of covariates, it is not clear which variables are important in estimating the CATE function. The true set of variables which belong to the estimating equations is assumed to be sparse, as discussed in the previous sections. We hence apply the lasso method in Belloni et al. (2017) to estimate propensity score ( $\pi_0$ ) and conditional mean function ( $\mu_0$ ). We then compute the robust score function  $\psi$  for each observation  $i$ , and run a kernel regression of  $\psi_i$  on mother’s age evaluated at numerous grid points in the interval [15, 36] (years of age). We use the cross-fitting variant of the estimator, i.e., the nuisance function estimation and the kernel regression (with the bandwidth choice the same as we discussed in the simulations) takes place in different subsamples, and then these roles are rotated. We refer to the resulting point estimates as HDCATE (HD stands for ‘high dimensional’).

The HDCATE estimates are displayed in Figures 1, 2 and 3, along with 90%, 95% and 99% confidence bands, respectively. For a given confidence level, we compute two types of intervals. “HDCATE CB” is the proposed uniform confidence band computed according to the algorithm given in Section 4. “PW CB” is a pointwise confidence band, given for purposes of comparison, where the critical value  $\widehat{C}_\alpha^{2\text{-sided}}$  is replaced by the corresponding value from the standard normal distribution (e.g., 1.96 for  $\alpha = 5\%$ ). The constant function labeled “ATE” represents the estimated average treatment effect across all ages.

Figures 1 - 3 show that maternal smoking has a negative effect on birth weight at all ages (the upper bounds of the confidence bands are negative), and the average effect is likely becoming *more negative* with age. For example, the point estimates show that for teenage mothers of age 18 or younger the negative effect of smoking is, on average, less than 180 grams in absolute value. For mothers around age 24, the same effect is -220 grams, and it approaches -250 above

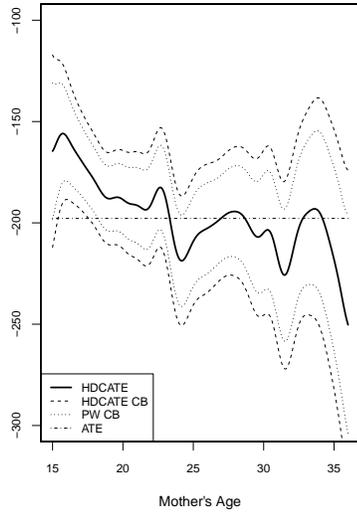


Figure 1: CATE for the effect of smoking on birth weights conditional on mother's age, 99% confidence bands.

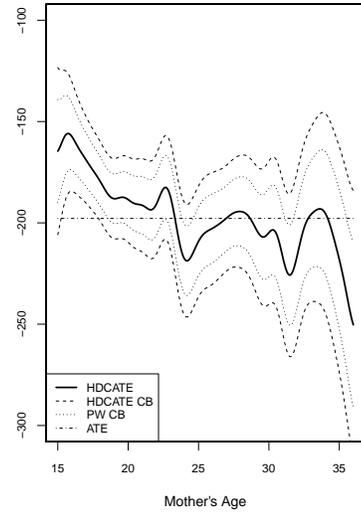


Figure 2: CATE for the effect of smoking on birth weights conditional on mother's age, 95% confidence bands.

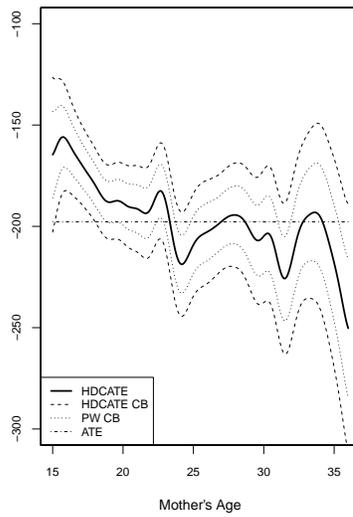


Figure 3: CATE for the effect of smoking on birth weights conditional on mother's age, 90% confidence bands.

35 years of age.<sup>15</sup> Thus, there is substantial variation in the estimated average treatment effect by age. A potential explanation is that older mothers are likely to have smoked for a longer period and the detrimental effects of smoking are cumulative (the smoking dummy does not control for duration or intensity of smoking).

Examining the confidence bands qualifies the analysis of the point estimate in important ways. In Figure 1, the lower bound of the 99% uniform confidence band (dashed line) attains its maximum at around 16 years of age, and the value of this maximum lies just below the minimum of the upper bound attained at around age 24. Thus, it is possible to fit a constant function (at about -185 grams) inside the uniform confidence bands. Nevertheless, if one is less conservative and uses the 95% or 90% uniform confidence bands displayed on Figures 2 and 3, respectively, then it is no longer possible to do so. Thus, there is fairly compelling (statistically significant) evidence that the smoking effect becomes more negative at least between the ages of 16 and 24. Based on the pointwise confidence band, there is some evidence of further decline in HDCATE at higher ages but it is possible to fit constant functions even within the 90% uniform confidence bands over the interval [25,35]. (Again note that these bands become rather wide at higher ages due to the relatively small number of observations.)

## 7 Conclusion

We advance the literature on the estimation of the reduced dimensional CATE function by proposing that the nuisance functions necessary for identification be estimated by flexible machine learning (selection) methods, followed by a traditional kernel regression. This is a significant innovation for several reasons. On the practical side, large data sets in which the number of variables is comparable to the sample size are increasingly common, but traditional parametric or nonparametric methods both have difficulties in dealing with such data. However, the unconfoundedness assumption used to identify CATE gains credibility precisely in data rich environments. It is therefore practically important to establish an estimation method that can actually exploit this richness in estimating CATE rather than being burdened by it.

The asymptotic theory we develop builds on previous work by Belloni et al. (2017) and

---

<sup>15</sup>The non-monotonicities in the point estimate between ages 25 and 35 could be due to undersmoothing and the quickly declining number of first time mothers toward the top of this age range.

Chernozhukov et al. (2018) on machine-learning-aided causal inference. Nevertheless, the theory requires non-trivial modifications to accommodate kernel based nonparametric regression in the second stage. Moreover, CATE is a functional parameter, and our results can be used to conduct uniform inference through a straightforward bootstrap procedure. The presented simulation results confirm that the proposed estimators and inference methods work well in finite samples. In line with Chernozhukov et al. (2018), we also advocate estimating the nuisance functions and conducting the second stage regression on different samples (i.e., using the cross-fitting approach).

Using the proposed methods, we revisited the problem of estimating the average effect of smoking during pregnancy on birthweight as a function of the mother's age. Our results fall in between Abrevaya et al. (2015) and LOW in the sense that we do find age related heterogeneity (unlike LOW), but it is less marked than in the former study. In particular, there is evidence that the negative effect of smoking becomes somewhat more pronounced with age.

## Appendix: Proofs

### A Notation

In the following, we define notation that is not used in the main text.

- Full sample estimator:

- Let  $\xi^*$  be either 1 or  $\xi$ , a random variable that satisfies Assumption 4.1. The following proof is valid for the original and bootstrap estimators with  $\xi^* = 1$  and  $\xi$ , respectively.
- Let  $\psi(1, W, \eta) = \frac{D(Y-\mu(1, X))}{\pi(X)} + \mu(1, X)$  and  $\psi(0, W, \eta) = \frac{(1-D)(Y-\mu(0, X))}{1-\pi(X)} + \mu(0, X)$  such that  $\psi(W, \eta) = \psi(1, W, \eta) - \psi(0, W, \eta)$ .
- Let  $\tau_0(j, x_1) = \mathbb{E}(\mu_0(j, X)|X_1 = x_1)$ ,  $j = 0, 1$ .
- Let  $\hat{\tau}^*(j, x_1; I) = \frac{1}{Nh^d \hat{f}^*(x_1; I)} \sum_{i \in I} \xi_i^* \psi(j, W_i, \hat{\eta}(I)) \mathcal{K}_h(X_{1i} - x_1)$ ,  $j = 0, 1$ , where

$$\hat{f}^*(x_1; I) = \frac{1}{Nh^d} \sum_{i \in I} \xi_i^* \mathcal{K}_h(X_{1i} - x_1).$$

- Split sample estimator:

- Let  $\mathbb{P}_{n,k} f = \frac{1}{n} \sum_{i \in I_k} f(W_i)$  for a generic function  $f(\cdot)$ .
- Let  $\mathbb{P}_{I_k} f = \mathbb{E}(f(W_1, \dots, W_N) | W_i, i \in I_k^c)$  for a generic function  $f$ .
- Let  $\hat{\tau}^*(j, x_1; I_k) = \frac{1}{nh^d \hat{f}^*(x_1; I_k)} \sum_{i \in I_k} \xi_i^* \psi(j, W_i, \hat{\eta}(I_k^c)) \mathcal{K}_h(X_{1i} - x_1)$ ,  $j = 0, 1$ , where

$$\hat{f}^*(x_1; I_k) = \frac{1}{nh^d} \sum_{i \in I_k} \xi_i^* \mathcal{K}_h(X_{1i} - x_1).$$

### B The proof of Theorems 3.1 and 4.1 for the full sample estimator

We first show

$$\hat{\tau}^*(1, x_1; I) - \tau_0(1, x_1) = (\mathbb{P}_N - \mathbb{P}) \frac{\xi^*}{h^d f(x_1)} [\psi(1, W, \eta_0) - \tau_0(1, x_1)] \mathcal{K}_h(X_1 - x_1) + R_N^*(1, x_1), \quad (22)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_N^*(1, x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ . By the same argument, we can show that

$$\hat{\tau}^*(0, x_1; I) - \tau_0(0, x_1) = (\mathbb{P}_N - \mathbb{P}) \frac{\xi^*}{h^d f(x_1)} [\psi(0, W, \eta_0) - \tau_0(0, x_1)] \mathcal{K}_h(X_1 - x_1) + R_N^*(0, x_1), \quad (23)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_N^*(0, x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ . When  $\xi^* = 1$ ,  $\hat{\tau}^*(1, x_1; I) - \hat{\tau}^*(0, x_1; I)$  is the original estimator  $\hat{\tau}(x_1)$ , as  $\hat{f}^*(x_1; I) = \hat{f}(x_1; I)$ . Further denote  $R_N^*(j, x_1) = R_N(j, x_1)$  when  $\xi^* = 1$ . Then, (22), (23), and the fact that  $\tau_0(x_1) = \tau_0(1, x_1) - \tau_0(0, x_1)$  imply that

$$\hat{\tau}(x_1) - \tau_0(x_1) = (\mathbb{P}_N - \mathbb{P}) \frac{1}{h^d f(x_1)} [\psi(W, \eta_0) - \tau_0(x_1)] \mathcal{K}_h(X_1 - x_1) + R_N(x_1), \quad (24)$$

where  $R_N(x_1) = R_N(1, x_1) - R_N(0, x_1)$  such that  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_N(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ . This is the desired result for Theorem 3.1(a). Similarly, if  $\xi^* = \xi$ ,  $\hat{\tau}^*(1, x_1; I) - \hat{\tau}^*(0, x_1; I)$  is the bootstrap estimator  $\hat{\tau}^b(x_1)$ , as  $\hat{f}^*(x_1; I) = \hat{f}^b(x_1; I)$ . Further denote  $R_N^*(j, x_1) = R_N^b(j, x_1)$  in this case. Then, we have

$$\hat{\tau}^b(x_1) - \tau_0(x_1) = (\mathbb{P}_N - \mathbb{P}) \frac{\xi}{h^d f(x_1)} [\psi(W, \eta_0) - \tau_0(x_1)] \mathcal{K}_h(X_1 - x_1) + \tilde{R}_N^b(x_1), \quad (25)$$

where  $\tilde{R}_N^b(x_1) = R_N^b(1, x_1) - R_N^b(0, x_1)$ . Taking the difference between (24) and (25), we have

$$\hat{\tau}^b(x_1) - \hat{\tau}(x_1) = (\mathbb{P}_N - \mathbb{P}) \frac{\xi - 1}{h^d f(x_1)} [\psi(W, \eta_0) - \tau_0(x_1)] \mathcal{K}_h(X_1 - x_1) + R_N^b(x_1),$$

where  $R_N^b(x_1) = \tilde{R}_N^b(x_1) - R_N(x_1)$  such that  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_N^b(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ . This is the Bahadur representation for the bootstrap estimator in Theorem 4.1(a).

Next, we turn to prove (22). Note that

$$\begin{aligned} & \mathbb{P}_N \xi^* h^{-d} \psi(1, W, \hat{\eta}(I)) \mathcal{K}_h(X_1 - x_1) - \tau_0(1, x_1) f(x_1) \\ &= (\mathbb{P}_N - \mathbb{P}) \xi^* h^{-d} \psi(1, W, \eta_0) \mathcal{K}_h(X_1 - x_1) + \left[ (\mathbb{P}_N - \mathbb{P}) \xi^* h^{-d} [\psi(1, W, \hat{\eta}(I)) - \psi(1, W, \eta_0)] \mathcal{K}_h(X_1 - x_1) \right] \\ & \quad + \left[ \mathbb{P} \xi^* h^{-d} [\psi(1, W, \hat{\eta}(I)) - \psi(1, W, \eta_0)] \mathcal{K}_h(X_1 - x_1) \right] \\ & \quad + \left[ \mathbb{P} \xi^* h^{-d} \psi(1, W, \eta_0) \mathcal{K}_h(X_1 - x_1) - \tau_0(1, x_1) f(x_1) \right] \\ &= I(x_1) + II(x_1) + III(x_1) + IV(x_1), \end{aligned}$$

where  $\mathbb{P} \xi^* h^{-d} \psi(1, W, \hat{\eta}(I)) \mathcal{K}_h(X_1 - x_1)$  is interpreted as  $\mathbb{P} \xi^* h^{-d} \psi(1, W, \eta) \mathcal{K}_h(X_1 - x_1)$  evaluated at  $\eta = \hat{\eta}(I)$ .

Later, we will show that

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} (|II(x_1)| + |III(x_1)| + |IV(x_1)|) = o_p((\log(N)Nh^d)^{-1/2}). \quad (26)$$

Let  $\tilde{R}_N(1, x_1) = II(x_1) + III(x_1) + IV(x_1)$ . Given (26), we have

$$\begin{aligned} & \hat{\tau}^*(1, x_1; I) - \tau_0(1, x_1) \\ &= \frac{I(x_1) + \tilde{R}_N(1, x_1) + \tau_0(1, x_1) f(x_1)}{\hat{f}^*(x_1; I)} - \tau_0(1, x_1) \\ &= \frac{\tau_0(1, x_1)(f(x_1) - \hat{f}^*(x_1; I))}{f(x_1)} + \frac{I(x_1)}{f(x_1)} + \frac{\tau_0(1, x_1)(f(x_1) - \hat{f}^*(x_1; I))^2}{f(x_1)\hat{f}^*(x_1; I)} + \frac{I(x_1)(f(x_1) - \hat{f}^*(x_1; I))}{\hat{f}^*(x_1; I)} + \frac{\tilde{R}_N(1, x_1)}{\hat{f}^*(x_1; I)} \\ &= A_1(x_1) + A_2(x_1) + A_3(x_1) + A_4(x_1) + A_5(x_1). \end{aligned}$$

It is clear that

$$A_1(x_1) = -\frac{\tau_0(1, x_1)}{h^d f(x_1)} (\mathbb{P}_N - \mathbb{P}) \xi^* \mathcal{K}_h(X_1 - x_1) + \tilde{A}_1(x_1)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |\tilde{A}_1(x_1)| = O_p(h^2) = o_p((\log(N)Nh^d)^{-1/2})$ . Second,

$$A_2(x_1) = (\mathbb{P}_N - \mathbb{P}) \xi^* h^{-d} \psi(1, W, \eta_0) \mathcal{K}_h(X_1 - x_1) / f(x_1).$$

By the usual maximal inequality and the fact that, uniformly over  $x_1 \in \bar{\mathcal{X}}_1$ ,  $\tau_0(1, x_1)$  is bounded and  $f(x_1)$  is bounded and bounded away from zero, we have

$$\begin{aligned} \sup_{x_1 \in \bar{\mathcal{X}}_1} |A_3(x_1)| &\leq O_p(1) \sup_{x_1 \in \bar{\mathcal{X}}_1} \left[ (\mathbb{P}_N - \mathbb{P}) \xi^* h^{-d} \mathcal{K}_h(X_1 - x_1) + \mathbb{E} h^{-d} \mathcal{K}_h(X_1 - x_1) - f(x_1) \right]^2 \\ &= O_p(\log(N)/(Nh^d)) = o_p((\log(N)Nh^d)^{-1/2}). \end{aligned}$$

By the usual maximal inequality again,  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |I(x_1)| = O_p(\log^{1/2}(N)(Nh^d)^{-1/2})$ . Therefore,

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |A_4(x_1)| = O_p(\log(N)/(Nh^d)) = o_p((\log(N)Nh^d)^{-1/2}).$$

Last, by (26),

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |A_5(x_1)| = o_p((\log(N)Nh^d)^{-1/2}).$$

Combining the above bounds, we obtain that

$$\hat{\tau}^*(1, x_1; I) - \tau_0(1, x_1) = \frac{1}{h^d f(x_1)} (\mathbb{P}_N - \mathbb{P}) \xi^* \left[ \psi(1, W, \eta_0) - \tau_0(1, x_1) \right] \mathcal{K}_h(X_1 - x_1) + R_N^*(1, x_1)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_N^*(x_1)| = o_p((\log(N)Nh^d)^{-1/2})$ . This is the desired result.

Therefore, the only thing left is to show (26). First note that

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |IV(x_1)| = \sup_{x_1 \in \bar{\mathcal{X}}_1} \left| h^{-d} \mathbb{E} \mu_0(1, X) \mathcal{K}_h(X_1 - x_1) - \tau_0(1, x_1) f(x_1) \right| = o((\log(N)Nh^d)^{-1/2}).$$

Next, we turn to  $III(x_1)$ . By Assumption 3.2, for any  $\varepsilon > 0$ , there exists a positive constant  $M$ , such that, with probability greater than  $1 - \varepsilon$ ,

$$\hat{\eta}(I) \equiv (\hat{\mu}(1, \cdot), \hat{\mu}(0, \cdot), \hat{\pi}(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi,$$

where

$$\mathcal{F}_n^{(j)} = \left\{ \begin{array}{l} \mu(j, \cdot) \in \mathcal{G}_n^{(j)} : \sup_{x_1 \in \bar{\mathcal{X}}_1} \left\| (\mu(j, \cdot) - \mu_0(j, \cdot)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \leq M\delta_{1N}, \\ \|\mu(j, \cdot) - \mu_0(j, \cdot)\|_{\mathbb{P}, \infty} \leq M\delta_{2N} \end{array} \right\}, \quad j = 0, 1,$$

$$\mathcal{F}_n^\pi = \left\{ \pi(\cdot) \in \mathcal{G}_n^\pi : \sup_{x_1 \in \bar{\mathcal{X}}_1} \left\| (\pi(\cdot) - \pi_0(\cdot)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \leq M\delta_{1N}, \quad \|\pi(\cdot) - \pi_0(\cdot)\|_{\mathbb{P}, \infty} \leq M\delta_{2N} \right\},$$

and the classes of functions  $(\mathcal{G}_n^{(0)}, \mathcal{G}_n^{(1)}, \mathcal{G}_n^\pi)$  are defined in Assumption 3.2,

Then, with probability greater than  $1 - \varepsilon$ ,

$$\begin{aligned} & \sup_{x_1 \in \bar{\mathcal{X}}_1} |III(x_1)| \\ &= \sup_{x_1 \in \bar{\mathcal{X}}_1, (\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi} \left| \mathbb{E} \left[ \frac{(\mu_0(1, X) - \mu(1, X))(\pi_0(X) - \pi(X))}{h^d \pi(X)} \right] \mathcal{K}_h(X_1 - x_1) \right| \\ &\lesssim \sup_{x_1 \in \bar{\mathcal{X}}_1, (\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi} h^{-d} \left\| (\mu_0(1, X) - \mu(1, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \left\| (\pi_0(X) - \pi(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \\ &\lesssim M^2 \min(h^{-d} \delta_{1N}^2, \delta_{2N}^2) = o((\log(N)Nh^d)^{-1/2}). \end{aligned} \tag{27}$$

Last,

$$\begin{aligned} & \sup_{x_1 \in \bar{\mathcal{X}}_1} |II(x_1)| \\ &= \sup_{(\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1} \left| (\mathbb{P}_n - \mathbb{P}) \xi^* \left[ \frac{D(Y - \mu(1, X))(\pi_0(X) - \pi(X))}{h^d \pi(X) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) \right| \\ &+ \sup_{(\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1} \left| h^{-d} (\mathbb{P}_n - \mathbb{P}) \xi^* \left[ \left( 1 - \frac{D}{\pi(X)} \right) (\mu(1, X) - \mu_0(1, X)) \right] \mathcal{K}_h(X_1 - x_1) \right| \\ &= II_1 + II_2. \end{aligned}$$

Uniformly over  $x \in \mathcal{X}$ ,  $\pi_0(x)$  is bounded and bounded away from zero and  $\mu_0(1, x)$  is bounded. Therefore, so are  $\mu(1, x) \in \mathcal{F}_n^{(1)}$  and  $\pi(x) \in \mathcal{F}_n^\pi$  as  $\delta_{2N} = o(1)$ . Then, we have

$$\sup_{(\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1} \left| \xi^* \left[ \frac{D(Y - \mu(1, X))(\pi_0(X) - \pi(X))}{h^d \pi(X) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) \right| \lesssim h^{-d} \delta_{2N} |\xi^* Y|. \quad (28)$$

In addition, similar to (27), we have

$$\begin{aligned} & \sup_{(\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1} \mathbb{E} \left\{ \xi^* \left[ \frac{D(Y - \mu(1, X))(\pi_0(X) - \pi(X))}{h^d \pi(X) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) \right\}^2 \\ & \lesssim h^{-2d} \sup_{\pi(\cdot) \in \mathcal{F}_n^\pi} \mathbb{E} (\pi_0(X) - \pi(X))^2 \mathcal{K}_h(X_1 - x_1) \lesssim h^{-2d} \min(\delta_{1N}^2, \delta_{2N}^2 h^d). \end{aligned} \quad (29)$$

Denote

$$\mathcal{H}_1 = \left\{ \xi^* \left[ \frac{D(Y - \mu(1, X))(\pi_0(X) - \pi(X))}{h^d \pi(X) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) : (\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1 \right\}.$$

Combining (7) and the fact that

$$\sup_Q \log N \left( \left\{ K\left(\frac{\cdot - x_1}{h}\right) : x_1 \in \mathfrak{R} \right\}, \|\cdot\|_{Q,2}, \varepsilon \right) \lesssim \log(1/\varepsilon) \vee 0,$$

we have

$$\sup_Q \log N(\mathcal{H}_1, \|\cdot\|_{Q,2}, \varepsilon \|F_1\|_{Q,2}) \lesssim \delta_{3N} (\log(A_n) + \log(1/\varepsilon) \vee 0). \quad (30)$$

Since  $\xi^*$  is either 1 or  $\eta$  which has a sub-exponential tail and  $\mathbb{E} Y^q < \infty$ , we have  $\|\max_i |\xi_i^* Y_i|\|_2 \lesssim N^{1/q}$ .

Combining this fact with (28), (29), and (30), Belloni et al. (2017, Lemma C.1) implies that

$$\mathbb{E} |II_1| \lesssim \sqrt{\frac{\delta_{3N} \log(A_n \vee N) \min(\delta_{1N}^2, h^d \delta_{2N}^2)}{N h^{2d}}} + \frac{\delta_{2N} \delta_{3N} N^{1/q} \log(A_n \vee N)}{N h^d},$$

and thus, by Assumption 3.2,

$$II_1 = o_p((\log(N) N h^d)^{-1/2}). \quad (31)$$

Similarly, for  $II_2$ , we have

$$\sup_{(\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1} \left| h^{-d} \xi^* \left[ \left(1 - \frac{D}{\pi(X)}\right) (\mu(1, X) - \mu_0(1, X)) \right] \mathcal{K}_h(X_1 - x_1) \right| \lesssim \delta_{2N} h^{-d} |\xi^*|,$$

$$\sup_{(\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1} \mathbb{E} \left\{ h^{-d} \xi^* \left[ \left(1 - \frac{D}{\pi(X)}\right) (\mu(1, X) - \mu_0(1, X)) \right] \mathcal{K}_h(X_1 - x_1) \right\}^2 \lesssim h^{-2d} \min(\delta_{1N}^2, \delta_{2N}^2 h^d),$$

and

$$\sup_Q \log N(\mathcal{H}_2, \|\cdot\|_{Q,2}, \varepsilon \|F_1\|_{Q,2}) \lesssim \delta_{3N} (\log(A_n) + \log(1/\varepsilon) \vee 0),$$

where

$$\mathcal{H}_2 = \left\{ h^{-d} \xi^* \left[ \left(1 - \frac{D}{\pi(X)}\right) (\mu(1, X) - \mu_0(1, X)) \right] \mathcal{K}_h(X_1 - x_1) : (\mu(1, \cdot), \pi(\cdot)) \in \mathcal{F}_n^{(1)} \times \mathcal{F}_n^\pi, x_1 \in \bar{\mathcal{X}}_1 \right\}.$$

Therefore, by the same argument as above,

$$II_2 = o_p((\log(N) N h^d)^{-1/2}). \quad (32)$$

(31) and (32) imply that

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |II(x_1)| = o_p((\log(N) N h^d)^{-1/2}).$$

This concludes the proof of Theorem 3.1 and 4.1.

## C The proof of Theorems 3.1 and 4.1 for the split sample estimator

Recall that  $n = N/K$ . We first show

$$\hat{\tau}^*(1, x_1; I_k) - \tau_0(1, x_1) = (\mathbb{P}_{n,k} - \mathbb{P}) \frac{\xi^*}{h^d f(x_1)} [\psi(1, W, \eta_0) - \tau_0(1, x_1)] \mathcal{K}_h(X_1 - x_1) + R_{n,k}^*(1, x_1), \quad (33)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_{n,k}^*(1, x_1)| = o_p((\log(n)nh^d)^{-1/2})$ . Similarly, we have

$$\hat{\tau}^*(0, x_1; I_k) - \tau_0(0, x_1) = (\mathbb{P}_{n,k} - \mathbb{P}) \frac{\xi^*}{h^d f(x_1)} [\psi(0, W, \eta_0) - \tau_0(0, x_1)] \mathcal{K}_h(X_1 - x_1) + R_{n,k}^*(0, x_1), \quad (34)$$

where  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |R_{n,k}^*(0, x_1)| = o_p((\log(n)nh^d)^{-1/2})$ . Similar to the proof for the full sample estimator, when  $\xi^* = 1$ ,  $\hat{\tau}^*(1, x_1; I_k) - \hat{\tau}^*(0, x_1; I_k) = \tilde{\tau}_k(x_1)$ . Further denote  $R_{n,k}^*(j, x_1) = R_{n,k}(j, x_1)$  for  $j = 0, 1, k = 1, \dots, K$ , when  $\xi^* = 1$ . Then, (33) and (34) imply

$$\tilde{\tau}(x_1) - \tau_0(x_1) = \frac{1}{K} \sum_{k=1}^K (\tilde{\tau}_k(x_1) - \tau_0(x_1)) = (\mathbb{P}_N - \mathbb{P}) \frac{\psi(W, \eta_0) - \tau_0(x_1)}{h^d f(x_1)} \mathcal{K}_h(X_1 - x_1) + \check{R}_N(x_1), \quad (35)$$

where  $\check{R}_N(x_1) = \frac{1}{K} \sum_{k=1}^K (R_{n,k}(1, x_1) - R_{n,k}(0, x_1))$  such that  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{R}_N(x_1)| = o_p((\log(n)nh^d)^{-1/2})$ . Last, note that  $\log(n)nh^d$  and  $\log(N)Nh^d$  has the same order of magnitude. This establishes Theorem 3.1(b). Similarly, when  $\xi^* = \xi$ ,  $\hat{\tau}^*(1, x_1; I_k) - \hat{\tau}^*(0, x_1; I_k) = \tilde{\tau}_k^b(x_1)$ . Further denote  $R_{n,k}^*(j, x_1) = R_{n,k}^b(j, x_1)$  for  $j = 0, 1, k = 1, \dots, K$ . Then, (33) and (34) imply

$$\hat{\tau}^b(x_1) - \tau_0(x_1) = \frac{1}{K} \sum_{k=1}^K (\tilde{\tau}_k^b(x_1) - \tau_0(x_1)) = (\mathbb{P}_N - \mathbb{P}) \frac{\xi(\psi(W, \eta_0) - \tau_0(x_1))}{h^d f(x_1)} \mathcal{K}_h(X_1 - x_1) + \check{R}'_N(x_1), \quad (36)$$

where  $\check{R}'_N(x_1) = \frac{1}{K} \sum_{k=1}^K (R_{n,k}^b(1, x_1) - R_{n,k}^b(0, x_1))$ . Taking the difference between (35) and (36), we have

$$\hat{\tau}^b(x_1) - \tilde{\tau}(x_1) = \frac{1}{K} \sum_{k=1}^K (\tilde{\tau}_k^b(x_1) - \tau_0(x_1)) = (\mathbb{P}_N - \mathbb{P}) \frac{(\xi - 1)(\psi(W, \eta_0) - \tau_0(x_1))}{h^d f(x_1)} \mathcal{K}_h(X_1 - x_1) + \check{R}^b_N(x_1),$$

where  $\check{R}^b_N(x_1) = \check{R}_N(x_1) - \check{R}'_N(x_1)$  such that  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{R}^b_N(x_1)| = o_p((\log(n)nh^d)^{-1/2})$ . This establishes Theorem 4.1(b).

Next, we turn to prove (33). Note that

$$\begin{aligned} & \mathbb{P}_{n,k} \xi^* h^{-d} \psi(1, W, \hat{\eta}(I_k^c)) \mathcal{K}_h(X_1 - x_1) - \tau_0(1, x_1) f(x_1) \\ &= (\mathbb{P}_{n,k} - \mathbb{P}) \xi^* h^{-d} \psi(1, W, \eta_0) \mathcal{K}_h(X_1 - x_1) + \left[ (\mathbb{P}_{n,k} - \mathbb{P}_{I_k}) \xi^* h^{-d} [\psi(1, W, \hat{\eta}(I_k^c)) - \psi(1, W, \eta_0)] \mathcal{K}_h(X_1 - x_1) \right] \\ & \quad + \left[ \mathbb{P}_{I_k} \xi^* h^{-d} [\psi(1, W, \hat{\eta}(I_k^c)) - \psi(1, W, \eta_0)] \mathcal{K}_h(X_1 - x_1) \right] \\ & \quad + \left[ \mathbb{P}_{I_k} \xi^* h^{-d} \psi(1, W, \eta_0) \mathcal{K}_h(X_1 - x_1) - \tau_0(1, x_1) f(x_1) \right] \\ &= I_k(x_1) + II_k(x_1) + III_k(x_1) + IV_k(x_1). \end{aligned}$$

We aim to show that

$$\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} (|II_k(x_1)| + |III_k(x_1)| + |IV_k(x_1)|) = o_p((\log(n)nh^d)^{-1/2}). \quad (37)$$

Let  $\tilde{R}_{n,k}(1, x_1) = II_k(x_1) + III_k(x_1) + IV_k(x_1)$ . Given (37), we have

$$\begin{aligned}
& \hat{\tau}^*(1, x_1; I_k) - \tau_0(1, x_1) \\
&= \frac{I(x_1) + \tilde{R}_{n,k}(1, x_1) + \tau_0(1, x_1)f(x_1)}{\hat{f}^*(x_1; I_k)} - \tau_0(1, x_1) \\
&= \frac{\tau_0(1, x_1)(f(x_1) - \hat{f}^*(x_1; I_k))}{f(x_1)} + \frac{I_k(x_1)}{f(x_1)} \\
&\quad + \frac{\tau_0(1, x_1)(f(x_1) - \hat{f}^*(x_1; I_k))^2}{f(x_1)\hat{f}^*(x_1; I_k)} + \frac{I(x_1)(f(x_1) - \hat{f}^*(x_1; I_k))}{\hat{f}^*(x_1; I_k)} + \frac{\tilde{R}_{n,k}(1, x_1)}{\hat{f}^*(x_1; I_k)} \\
&= A_{1k}(x_1) + A_{2k}(x_1) + A_{3k}(x_1) + A_{4k}(x_1) + A_{5k}(x_1).
\end{aligned}$$

It is clear that

$$A_{1k}(x_1) = -\frac{\tau_0(1, x_1)}{h^d f(x_1)} (\mathbb{P}_{n,k} - \mathbb{P}) \xi^* \mathcal{K}_h(X_1 - x_1) + \tilde{A}_{1k}(x_1),$$

where  $\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |\tilde{A}_{1k}(x_1)| = O_p(h^2) = o_p((\log(n)nh^d)^{-1/2})$ . In addition,

$$A_{2k}(x_1) = \frac{(\mathbb{P}_{n,k} - \mathbb{P}) \xi^* h^{-d} \psi(1, W, \eta_0) \mathcal{K}_h(X_1 - x_1)}{f(x_1)}.$$

Next, by the usual maximal inequality and the fact that, uniformly over  $x_1 \in \bar{\mathcal{X}}_1$ ,  $\tau_0(1, x_1)$  is bounded and  $f(x_1)$  is bounded and bounded away from zero, we have

$$\begin{aligned}
\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |A_{3k}(x_1)| &\leq O_p(1) \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} \left[ (\mathbb{P}_{n,k} - \mathbb{P}) \xi^* h^{-d} \mathcal{K}_h(X_1 - x_1) + \mathbb{E} h^{-d} \mathcal{K}_h(X_1 - x_1) - f(x_1) \right]^2 \\
&= O_p(\log(n)/(nh^d)) = o_p((\log(n)nh^d)^{-1/2}).
\end{aligned}$$

By the usual maximal inequality again,  $\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |I_k(x_1)| = O_p(\log^{1/2}(n)(nh^d)^{-1/2})$ . Therefore,

$$\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |A_{4k}(x_1)| = O_p(\log(n)/(nh^d)) = o_p((\log(n)nh^d)^{-1/2}).$$

Last, by (37),

$$\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |A_{5k}(x_1)| = o_p((\log(n)nh^d)^{-1/2}).$$

Combining the above bounds, we obtain that

$$\begin{aligned}
& \hat{\tau}^*(1, x_1; I_k) - \tau_0(1, x_1) \\
&= \frac{1}{h^d f(x_1)} (\mathbb{P}_{n,k} - \mathbb{P}) \xi^* \left[ \psi(1, W, \eta_0) - \tau_0(1, x_1) \right] \mathcal{K}_h(X_1 - x_1) + R_{n,k}^*(1, x_1)
\end{aligned}$$

where  $\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |R_{n,k}^*(1, x_1)| = o_p((\log(n)nh^d)^{-1/2})$ . We obtain the desired result.

Therefore, the only thing left is to show (37). First note that

$$\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |IV_k(x_1)| = \sup_{x_1 \in \bar{\mathcal{X}}_1} \left| h^{-d} \mathbb{E} \mu_0(1, X) \mathcal{K}_h(X_1 - x_1) - \tau_0(1, x_1) f(x_1) \right| = o((\log(n)nh^d)^{-1/2}).$$

Second, let

$$\mathcal{F}_{0n} = \left\{ \begin{array}{l} (\mu(0, X), \pi(X)) : \\ \sup_{x_1 \in \bar{\mathcal{X}}_1} \left\| (\mu(0, X) - \mu_0(0, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \left\| (\pi(X) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \\ \leq M \delta_{1n}^2, \\ \|\mu(0, X) - \mu_0(0, X)\|_{\mathbb{P}, \infty} \leq M \delta_{2n}, \quad \|\pi(X) - \pi_0(X)\|_{\mathbb{P}, \infty} \leq M \delta_{2n} \end{array} \right\},$$

$$\mathcal{F}_{1n} = \left\{ \begin{array}{l} (\mu(1, X), \pi(X)) : \\ \sup_{x_1 \in \bar{\mathcal{X}}_1} \left\| (\mu(1, X) - \mu_0(1, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \left\| (\pi(X) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \\ \leq M \delta_{1n}^2, \\ \|\mu(1, X) - \mu_0(1, X)\|_{\mathbb{P}, \infty} \leq M \delta_{2n}, \quad \|\pi(X) - \pi_0(X)\|_{\mathbb{P}, \infty} \leq M \delta_{2n} \end{array} \right\},$$

and  $\mathcal{A}_n(M) = \{(\hat{\mu}(0, \cdot; I_k^c), \hat{\pi}(\cdot; I_k^c)) \in \mathcal{F}_{0n}\} \cap \{(\hat{\mu}(1, \cdot; I_k^c), \hat{\pi}(\cdot; I_k^c)) \in \mathcal{F}_{1n}\}$ . By Assumption 3.1, for any  $\varepsilon > 0$ , there exists a positive constant  $M$ , such that  $\mathbb{P}(\mathcal{A}_n(M)) \geq 1 - \varepsilon$ . Then, on  $\mathcal{A}_n(M)$ , we have

$$\begin{aligned} & \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |III_k(x_1)| \\ & \leq \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} \left| \mathbb{P}_{I_k} \xi^* h^{-d} [\psi(1, W, \hat{\eta}(I_k^c)) - \psi(1, W, \eta_0)] \mathcal{K}_h(X_1 - x_1) \right| \\ & \leq \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1, \eta \in \mathcal{A}_n(M)} \left| \mathbb{P}_{I_k} \xi^* h^{-d} [\psi(1, W, \eta) - \psi(1, W, \eta_0)] \mathcal{K}_h(X_1 - x_1) \right| \\ & \leq \sup_{x_1 \in \bar{\mathcal{X}}_1, \eta \in \mathcal{A}_n(M)} \left| \mathbb{E} \left[ \frac{(\mu(1, X) - \mu_0(1, X))(\pi(X) - \pi_0(X))}{h^d \pi(X)} \right] \mathcal{K}_h(X_1 - x_1) \right| \\ & \leq \sup_{x_1 \in \bar{\mathcal{X}}_1, \eta \in \mathcal{A}_n(M)} h^{-d} \left\| (\mu(1, X) - \mu_0(1, X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \left\| (\pi(X) - \pi_0(X)) \mathcal{K}_h^{1/2}(X_1 - x_1) \right\|_{\mathbb{P}, 2} \\ & \lesssim h^{-d} \delta_{1n}^2. \end{aligned} \tag{38}$$

Because  $\varepsilon$  is arbitrary, we have

$$\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |III_k(x_1)| = O_p(h^{-d} \delta_{1n}^2) = o_p((\log(n)nh^d)^{-1/2}).$$

Last, note

$$\begin{aligned} & \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |II_k(x_1)| \\ & \leq \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} \left| (\mathbb{P}_{n,k} - \mathbb{P}_{I_k}) \xi^* \left[ \frac{D(Y - \hat{\mu}(1, X; I_k^c))(\pi_0(X) - \hat{\pi}(X; I_k^c))}{h^d \hat{\pi}(X; I_k^c) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) \right| \\ & \quad + \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} \left| h^{-d} (\mathbb{P}_{n,k} - \mathbb{P}_{I_k}) \xi^* \left[ \left( 1 - \frac{D}{\pi_0(X)} \right) (\hat{\mu}(1, X; I_k^c) - \mu_0(1, X)) \right] \mathcal{K}_h(X_1 - x_1) \right| \\ & = \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) + \sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} II_{2k}(x_1). \end{aligned}$$

Next, we apply the usual maximal inequality to bound  $II_1$ . The upper bound for  $II_2$  can be derived in the same manner, and thus, is omitted. We denote, for arbitrary fixed  $(\mu(1, X), \pi(X)) \in \mathcal{F}_{1n}$ ,

$$\mathcal{H}_1(\mu(1, X), \pi(X)) = \left\{ \xi^* \left[ \frac{D(Y - \mu(1, X))(\pi_0(X) - \pi(X))}{h^d \pi(X) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) : x_1 \in \bar{\mathcal{X}}_1 \right\}.$$

Then,  $\mathcal{H}_1(\mu(1, X), \pi(X))$  has an envelope function

$$H_{1i}(\mu(1, X), \pi(X)) \lesssim |\xi^* Y| h^{-d} |\pi(X) - \pi_0(X)| \lesssim |\xi_i^*| h^{-d} \delta_{2n}$$

and

$$\sup_i H_{1i}(\mu(1, X), \pi(X)) \lesssim \sup_i |\xi_i^* Y_i| h^{-d} \delta_{2n}. \tag{39}$$

Furthermore, for fixed functions  $(\mu(1, X), \pi(X))$ , we have

$$\sup_Q \log(N(\mathcal{H}_1(\mu(1, X), \pi(X)), \|\cdot\|_{Q,2}, \varepsilon) \|H_1(\mu(1, X), \pi(X))\|_{Q,2}) \lesssim \log(1/\varepsilon) \vee 0. \quad (40)$$

In addition, similar to (38), on  $\mathcal{A}_n(M)$ , we have

$$\begin{aligned} & \sup_{x_1 \in \bar{\mathcal{X}}_1, (\mu(1, X), \pi(X)) \in \mathcal{F}_{1n}} \mathbb{E} \left\{ \xi^* \left[ \frac{D(Y - \mu(1, X))(\pi_0(X) - \pi(X))}{h^d \pi(X) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) \right\}^2 \\ & \lesssim h^{-2d} \sup_{x_1 \in \bar{\mathcal{X}}_1, (\mu(1, X), \pi(X)) \in \mathcal{F}_{1n}} \mathbb{E}(\pi_0(X) - \pi(X))^2 \mathcal{K}_h(X_1 - x_1) \\ & \lesssim h^{-d} \delta_{2n}^2. \end{aligned}$$

Therefore,

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} \mathbb{P}_{I_k} \left\{ \xi^* \left[ \frac{D(Y - \hat{\mu}(1, X; I_k^c))(\pi_0(X) - \hat{\pi}(X; I_k^c))}{h^d \hat{\pi}(X; I_k^c) \pi_0(X)} \right] \mathcal{K}_h(X_1 - x_1) \right\}^2 \mathbb{1}\{\mathcal{A}_n(M)\} \lesssim \delta_{2n}^2 h^{-d}. \quad (41)$$

Last, we also note that, by construction,  $(\hat{\mu}(1, \cdot; I_k^c), \hat{\pi}(\cdot; I_k^c)) \perp\!\!\!\perp \{W_i\}_{i \in I_k}$ . This implies, conditional on  $\{W_i, i \in I_k^c\}$ , we can treat  $(\hat{\mu}(1, \cdot; I_k^c), \hat{\pi}(\cdot; I_k^c))$  as fixed functions. Therefore, Belloni et al. (2017, Lemma C.1) implies that

$$\begin{aligned} \mathbb{P}_{I_k} \sup_{x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) \mathbb{1}\{\mathcal{A}_n(M)\} &= \mathbb{E}(\sup_{x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) | W_i, i \in I^c) \mathbb{1}\{\mathcal{A}_n(M)\} \\ &\lesssim \sqrt{\frac{h^{-d} \delta_{2n}^2 \log(n)}{n}} + \frac{\|\sup_i |\xi_i^* Y_i|\|_{P,q} \delta_{2n} \log(n)}{nh^d} \\ &\lesssim \sqrt{\frac{h^{-d} \delta_{2n}^2 \log(n)}{n}} + \frac{\delta_{2n} \log(n) n^{1/q}}{nh^d}, \end{aligned} \quad (42)$$

where the second inequality holds by the fact that  $\xi_i^*$  has sub-exponential tails.<sup>16</sup> Therefore, for an arbitrary  $\varepsilon_0 > 0$ , as  $n \rightarrow \infty$ ,

$$\begin{aligned} & \mathbb{P}(\sup_{x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) \geq \varepsilon_0 (\log(n) nh^d)^{-1/2}) \\ & \leq \varepsilon + \mathbb{P}(\sup_{x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) \mathbb{1}\{\mathcal{A}_n(M)\} \geq \varepsilon_0 (\log(n) nh^d)^{-1/2}) \\ & \leq \varepsilon + \mathbb{E} \mathbb{P}_{I_k}(\sup_{x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) \mathbb{1}\{\mathcal{A}_n(M)\} \geq \varepsilon_0 (\log(n) nh^d)^{-1/2}) \\ & \leq \varepsilon + \mathbb{E} \left\{ \left[ \frac{\mathbb{P}_{I_k} \sup_{x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) (\log(n) nh^d)^{1/2}}{\varepsilon_0} \right] \mathbb{1}\{\mathcal{A}_n(M)\} \right\} \\ & \leq \varepsilon + C \left[ \delta_{2n} \log(n) + \frac{\delta_{2n} \log^{3/2}(n) n^{1/q}}{(nh^d)^{1/2}} \right] / \varepsilon_0 \\ & \leq 2\varepsilon, \end{aligned}$$

where the first inequality is due to the union bound inequality, the second inequality is due to the Bayes rule and the fact that  $\mathcal{A}_n$  belongs to the sigma field generated by  $W_i, i \in I_k^c$ , the third inequality is due to the Markov inequality, the fourth inequality is due to (42), and the last inequality holds by Assumption 3.1. Therefore,

$$\sup_{k \leq K} \sup_{x_1 \in \bar{\mathcal{X}}_1} II_{1k}(x_1) = o_p((\log(n) nh^d)^{-1/2}). \quad (43)$$

<sup>16</sup>A random variable  $\eta$  has sub-exponential tails if  $P(|\eta| > x) \leq K \exp(-Cx^2)$  for every  $x$  and some constants  $K$  and  $C$ . Note that a normal distribution satisfies this condition.

Similarly, we can show that

$$\sup_{k \leq K} \sup_{x_1 \in \bar{\mathcal{X}}_1} II_{2k}(x_1) = o_p((\log(n)nh^d)^{-1/2}). \quad (44)$$

(43) and (44) imply that  $\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |II_k(x_1)| = o_p((\log(n)nh^d)^{-1/2})$ . This concludes the proof.

## D Proof of Theorem 3.2

We focus on the split sample estimator  $\check{\sigma}_N^2(x_1)$ . The proof for the full sample estimator  $\widehat{\sigma}_N^2(x_1)$  is similar but simpler. Let

$$\sigma_k^2(x_1) = \text{Var} \left( \frac{\sqrt{nh^d}}{h^d f(x_1)} \mathbb{P}_{n,k}(\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right)$$

and recall that  $\sigma_N^2(x_1)$  is defined as

$$\sigma_N^2(x_1) = \text{Var} \left( \frac{\sqrt{Nh^d}}{h^d f(x_1)} \mathbb{P}_N(\psi(W, \eta_0) - \tau_0(x_1)) \mathcal{K}_h(X_1 - x_1) \right).$$

Then, we have

$$\frac{1}{K} \sum_{k=1}^K \sigma_k^2(x_1) = \sigma_N^2(x_1).$$

Therefore, it suffices to show that

$$\sup_{k \leq K, x_1 \in \bar{\mathcal{X}}_1} |\check{\sigma}_k^2(x_1) - \sigma_k^2(x_1)| = o_p(1).$$

Let  $\Gamma(W, x_1) = \frac{\psi(W_i, \eta_0) - \tau_0(x_1)}{h^d f(x_1)} \mathcal{K}_h(X_1 - x_1)$  and  $\check{\sigma}_k^2(x_1) = h^d \mathbb{P}_{n,k}(\Gamma(W, x_1))^2$ . We aim to show that, for  $k = 1, \dots, K$ ,

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{\sigma}_k^2(x_1) - \sigma_k^2(x_1)| = o_p(1) \quad (45)$$

and

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\check{\sigma}_k^2(x_1) - \check{\sigma}_k^2(x_1)| = o_p(1). \quad (46)$$

We first show (45). We claim that  $\sup_{x_1 \in \bar{\mathcal{X}}_1} |\text{Var}(\sqrt{nh^d} \mathbb{P}_{n,k} \Gamma(W, x_1)) - \mathbb{E}(nh^d [\mathbb{P}_{n,k} \Gamma(W, x_1)]^2)| = o_p(1)$ . Because  $\text{Var}(A) = \mathbb{E}[A^2] - \mathbb{E}[A]^2$ , it is equivalent to show that  $\mathbb{E}[\sqrt{nh^d} \mathbb{P}_{n,k} \Gamma(W, x_1)] = o(1)$  uniformly over  $x_1$ . By standard arguments and Assumption 3.1, we have, uniformly over  $x_1$

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} \mathbb{E}[\sqrt{nh^d} \mathbb{P}_{n,k} \Gamma(W, x_1)] = O(\sqrt{nh^d} h^2) = o(1),$$

Similarly, we have

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\mathbb{E}(nh^d [\mathbb{P}_{n,k} \Gamma(W, x_1)]^2) - \mathbb{E}(h^d \mathbb{P}_{n,k} \Gamma^2(W, x_1))| = O(nh^{d+4}) = o(1).$$

We next show that

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |h^d \mathbb{P}_{n,k} \Gamma^2(W, x_1) - \mathbb{E}[h^d \mathbb{P}_{n,k} \Gamma^2(W, x_1)]| = o_p(1). \quad (47)$$

By [van der Vaart and Wellner \(1996\)](#), we have

$$\mathcal{F}_K = \left\{ \mathcal{K} \left( \frac{X_1 - x_1}{h} \right) : x_1 \in \bar{\mathcal{X}}_1 \right\}$$

is of VC type with envelop function  $\bar{K} = \sup_u |\mathcal{K}(u)|$  which is bounded. This implies that

$$\mathcal{F}_{K^2} = \left\{ \mathcal{K}^2 \left( \frac{X_1 - x_1}{h} \right) : x_1 \in \bar{\mathcal{X}}_1 \right\}$$

is of VC type with envelop function  $\bar{K}^2$ . Similarly,

$$\mathcal{F}_{h^d \Gamma^2} = \left\{ h^d \Gamma^2(W, x_1) : x_1 \in \bar{\mathcal{X}}_1 \right\}$$

is of VC type with an envelop function  $Ch^{-d}\bar{K}^2 \cdot (\psi(W, \eta_0) - \tau_0(x_1))^2$  for some constant  $C > 0$ . In addition,

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} \mathbb{E} h^{2d} \Gamma^4(W, x_1) \lesssim \sup_{x_1 \in \bar{\mathcal{X}}_1} h^{-2d} \mathbb{E} \left( \mathcal{K}^4 \left( \frac{X_1 - x_1}{h} \right) \right) \lesssim h^{-d}.$$

Therefore, by [Belloni et al. \(2017, Lemma C.1\)](#), we have

$$\mathbb{E} \left[ \sup_{x_1 \in \bar{\mathcal{X}}_1} |h^d \mathbb{P}_{n,K} \Gamma^2(W, x_1) - \mathbb{E}[h^d \mathbb{P}_{n,k} \Gamma^2(W, x_1)]| \right] \lesssim \sqrt{\frac{\log(n)}{nh^d}} + \frac{\log(n)n^{1/q}}{nh^d} = o(1), \quad (48)$$

implying that (47) holds, and thus,

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\ddot{\sigma}_k^2(x_1) - \sigma_k^2(x_1)| = o_p(1).$$

This shows (45). Next, we show (46). Denote

$$\tilde{\Gamma}(W, x_1) = \frac{(\psi(W, \hat{\eta}(I_k^c)) - \tilde{\tau}_k(x_1))}{\hat{f}(x_1; I_k) h^d} \mathcal{K}_h(X_1 - x_1) \quad \text{and} \quad R_k(W_i, x_1) = \tilde{\Gamma}(W_i, x_1) - \Gamma(W_i, x_1).$$

If

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} h^d \|R_k(\cdot, x_1)\|_{\mathbb{P}_{n,k,2}}^2 = o_p(1), \quad (49)$$

then

$$\begin{aligned} & \sup_{x_1 \in \bar{\mathcal{X}}_1} |\ddot{\sigma}_k^2(x_1) - \check{\sigma}_k^2(x_1)| \\ & \leq \sup_{x_1 \in \bar{\mathcal{X}}_1} h^d [2|\mathbb{P}_{n,k} \Gamma(W, x_1) R_k(W, x_1)| + \mathbb{P}_{n,k} R_k^2(W, x_1)] \\ & \leq \sup_{x_1 \in \bar{\mathcal{X}}_1} \left[ 2h^d \|\Gamma(\cdot, x_1)\|_{\mathbb{P}_{n,k,2}} \|R_k(\cdot, x_1)\|_{\mathbb{P}_{n,k,2}} + h^d \|R_k(\cdot, x_1)\|_{\mathbb{P}_{n,k,2}}^2 \right] \\ & = 2 \sqrt{\left[ \sup_{x_1 \in \bar{\mathcal{X}}_1} |h^d \mathbb{P}_{n,K} \Gamma^2(W, x_1) - \mathbb{E}[h^d \mathbb{P}_{n,k} \Gamma^2(W, x_1)]| \right] + \sup_{x_1 \in \bar{\mathcal{X}}_1} \mathbb{E}[h^d \mathbb{P}_{n,k} \Gamma^2(W, x_1)]} \\ & \quad \times h^{d/2} \sup_{x_1 \in \bar{\mathcal{X}}_1} \|R_k(\cdot, x_1)\|_{\mathbb{P}_{n,k,2}} + h^d \|R_k(\cdot, x_1)\|_{\mathbb{P}_{n,k,2}}^2 \\ & = o_p(1), \end{aligned}$$

where the last equality holds because of (48), (49), and the fact that  $\sup_{x_1 \in \bar{\mathcal{X}}_1} \mathbb{E}[h^d \mathbb{P}_{n,k} \Gamma^2(W, x_1)]$  is bounded.

Therefore, we only need to verify (49). We have

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} h^d \|R_k(\cdot, x_1)\|_{\mathbb{P}_{n,k,2}}^2$$

$$\begin{aligned}
&\leq \sup_{x_1 \in \bar{\mathcal{X}}_1} \frac{|\hat{f}^2(x_1; I_k) - f^2(x_1)|}{\hat{f}^2(x_1; I_k) f^2(x_1)} \frac{1}{nh^d} \sum_{i \in I_k} (\psi(W, \eta_0) - \tau_0(x_1))^2 \mathcal{K}_h^2(X_{1i} - x_1) \\
&\quad + \sup_{x_1 \in \bar{\mathcal{X}}_1} \frac{1}{\hat{f}^2(x_1; I_k)} \frac{1}{nh^d} \sum_{i \in I_k} \left( \frac{D}{\hat{\pi}(X; I_k^c)} - 1 \right)^2 (\hat{\mu}(1, X; I_k^c) - \mu_0(1, X))^2 \mathcal{K}_h^2(X_{1i} - x_1) \\
&\quad + \sup_{x_1 \in \bar{\mathcal{X}}_1} \frac{1}{\hat{f}^2(x_1; I_k)} \frac{1}{nh^d} \sum_{i \in I_k} \left( \frac{1-D}{1 - \hat{\pi}(X; I_k^c)} - 1 \right)^2 (\hat{\mu}(1, X; I_k^c) - \mu_0(1, X))^2 \mathcal{K}_h^2(X_{1i} - x_1) \\
&\quad + \sup_{x_1 \in \bar{\mathcal{X}}_1} \frac{1}{\hat{f}^2(x_1; I_k)} \frac{1}{nh^d} \sum_{i \in I_k} \left( \frac{\hat{\pi}(X; I_k^c) - \pi_0(X)}{P(X) \hat{\pi}(X; I_k^c)} D(Y - \mu_0(1, X)) \right)^2 \mathcal{K}_h^2(X_{1i} - x_1) \\
&\quad + \sup_{x_1 \in \bar{\mathcal{X}}_1} \frac{1}{\hat{f}^2(x_1; I_k)} \frac{1}{nh^d} \sum_{i \in I_k} \left( \frac{\hat{\pi}(X; I_k^c) - \pi_0(X)}{\pi_0(X) \hat{\pi}(X; I_k^c)} (1-D)(Y - \mu_0(0, X)) \right)^2 \mathcal{K}_h^2(X_{1i} - x_1) \\
&\equiv \sum_{q=1}^5 A_{qk}.
\end{aligned}$$

We want to show  $A_{qk} = o_p(1)$  for  $q = 1, \dots, 5$ ,  $k = 1, \dots, K$ . Note  $A_{1k} = o_p(1)$  because

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} |\hat{f}(x_1; I_k) - f(x_1)| = O_p \left( \sqrt{\frac{\log(n)}{nh^d}} \right) = o_p(1),$$

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} \mathbb{E} \frac{1}{nh^d} \sum_{i \in I_k} (\psi(W, \eta_0) - \tau_0(x_1))^2 \mathcal{K}_h^2(X_{1i} - x_1) = O(1),$$

and

$$\sup_{x_1 \in \bar{\mathcal{X}}_1} (\mathbb{P}_{n,k} - \mathbb{P}) h^{-d} (\psi(W, \eta_0) - \tau_0(x_1))^2 \mathcal{K}_h^2(X_{1i} - x_1) = o_p(1).$$

For  $A_{2k}$ , by Assumptions 3.1(ii) and 3.3,

$$\left\| \left( \frac{D}{\hat{\pi}(X; I_k^c)} - 1 \right)^2 \right\|_{P, \infty} = O_p(1).$$

Therefore,

$$0 \leq A_{2k} \leq O_p(1) \times h^{-d} \delta_{2n}^2 \sup_{x_1 \in \bar{\mathcal{X}}_1} \left\| \mathcal{K} \left( \frac{X_1 - x_1}{h^d} \right) \right\|_{\mathbb{P}_{n,k}, 2}^2 = O_p(\delta_{2n}^2) = o_p(1).$$

Similarly,  $A_{3k} = o_p(1)$ . Next, by Assumption 3.3,

$$0 \leq A_{4k} \leq O_p(\delta_{n2}^2) \sup_{x_1 \in \bar{\mathcal{X}}_1} \mathbb{P}_{n,k} h^{-d} (D(Y - \mu_1(X)))^2 \mathcal{K}_h^2(X_{1i} - x_1) = O_p(\delta_{n2}^2) = o_p(1).$$

Similarly,  $A_{5k} = o_p(1)$ . This completes the derivation of (49), and thus, the whole proof.

## E Proof of Theorem 4.2

The proof of Theorem 4.2 is based on Chernozhukov et al. (2014, Theorem 3.2). In particular, given that the  $\eta_i^b$ 's are normally distributed, our assumptions are sufficient for Conditions H1-H4 of Theorem 3.2 in *ibid*.

## References

- Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33(4), 485–505.
- Ai, C. and X. Chen (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica* 71(6), 1795–1843.
- Almond, D. and J. Currie (2011). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives* 25(3), 153–172.
- Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation with high-dimensional data. *Econometrica* 85(1), 233–298.
- Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives* 28(2), 29–50.
- Belloni, A., V. Chernozhukov, and C. Hansen (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81(2), 608–650.
- Black, S., P. Devereux, and K. Salvanes (2007). From the cradle to the labor market? the effect of birth weight on adult outcomes. *The Quarterly Journal of Economics* 122(1), 409–439.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review Papers and Proceedings* 107(5), 261–65.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1), C1–C68.
- Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics* 42(4), 1564–1597.
- Chernozhukov, V. and V. Semenova (2019). Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions. *Working paper, Department of Economics, MIT*.

- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2), 315–331.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.
- Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature* 47(1), 5–86.
- 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.
- 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. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90(432), 1257–1270.
- van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2 ed.). MIT press.