Debiased Nonparametric Regression for Statistical Inference and Distributionally Robustness

A Preprint

Masahiro Kato

Department of Basic Science, The University of Tokyo
mkato-csecon@g.ecc.u-tokyo.ac.jp

December 31, 2024

Abstract

This study proposes a debiasing method for smooth nonparametric estimators. While machine learning techniques such as random forests and neural networks have demonstrated strong predictive performance, their theoretical properties remain relatively underexplored. Specifically, many modern algorithms lack assurances of pointwise asymptotic normality and uniform convergence, which are critical for statistical inference and robustness under covariate shift and have been well-established for classical methods like Nadaraya-Watson regression. To address this, we introduce a model-free debiasing method that guarantees these properties for smooth estimators derived from any nonparametric regression approach. By adding a correction term that estimates the conditional expected residual of the original estimator, or equivalently, its estimation error, we obtain a debiased estimator with proven pointwise asymptotic normality, uniform convergence, and Gaussian process approximation. These properties enable statistical inference and enhance robustness to covariate shift, making the method broadly applicable to a wide range of nonparametric regression problems.

1 Introduction

This paper addresses the problem of nonparametric regression. Let $X \in \mathcal{X}$ be a covariate and $Y \in \mathcal{Y}$ a target variable, where $\mathcal{X} \in [0,1]^d$ and $\mathcal{Y} \subset \mathbb{R}$ represent the covariate and target spaces, respectively. While we primarily focus on the case where X is one-dimensional for simplicity, the results presented in this paper extend to d-dimensional covariates.

Let P denote the joint distribution of (X, Y), and define the regression function under P as

$$f_P(X) = \mathbb{E}_P\left[Y \mid X\right],$$

where $\mathbb{E}_P\left[\,\cdot\,\right]$ represents the expectation operator with respect to P. We assume access to a dataset $\left\{(X_i, Y_i)\right\}_{i=1}^n$, where (X_i, Y_i) are independent and identically distributed samples from a true distribution P_0. The true regression function is denoted by $f_0 = f_{P_0}$. Our objective is to estimate f_0 based on the observations $\left\{(X_i, Y_i)\right\}_{i=1}^n$.

Notation. We denote the distribution induced by a regression function f as P_f, and the expectation operator under P_f as $\mathbb{E}_f$. For a real-valued vector z, let $\|z\|$ denote its Euclidean norm. For a measurable function $h\colon \mathcal{X} \to \mathbb{R}$, let $\|h\|_\infty := \sup_{x\in\mathcal{X}} |h(x)|$ denote the sup-norm, and $\|h\|_2 := \sqrt{\mathbb{E}\left[h(X)^2\right]}$ denote the L^2 norm. t

1.1 Content of this study

In Section 2, we define our debiased estimator in nonparametric regression. Our proposed estimator consists of the following three steps: estimate the regression function f_0 using any smooth nonparametric estimator; estimate the estimation error of the first stage estimator using the local polynomial regression; combine the first and second stage estimators. Here, we use the smoothness in the meaning that the either of the target nonparametric estimator, the true regression function, or their difference belongs to the Hölder class.

In Section 3, we show that our debiased nonparametric regression estimators have asymptotic normality and uniform convergence under mild conditions, like the smoothness. Furthermore, our estimator has the doubly robust property; that is, either of the first-stage nonparametric regression estimator or the second-stage conditional expected residual estimator is consistent, the resulting debiased estimator is consistent.

1.2 Related work

Doubly robust and debiased estimation has garnered significant attention across various fields, including statistics, economics, epidemiology, and machine learning. Doubly robust estimators are widely used in the context of semi-parametric statistics and causal inference (van der Vaart, 1998; Bang & Robins, 2005). With the development of machine learning regression models, debiasing methods utilizing doubly robust estimators have also been extensively studied (Chernozhukov et al., 2018), particularly in conjunction with sample splitting techniques (Klaassen, 1987; van der Vaart, 2002; Zheng & van der Laan, 2011). While much of the existing literature focuses on parametric or semiparametric models, this study extends these approaches to nonparametric regression.

Nonparametric regression in machine learning faces several challenges:

- For many machine learning estimators, the convergence rate to the true function is unknown, particularly with respect to minimax optimality.
- Asymptotic normality remains unproven for many machine learning estimators, including random forests (Breiman, 2001).
- Uniform convergence is not established for many estimators, such as neural networks (Schmidt-Hieber & Zamolodtchikov, 2024).
- Even for machine learning methods that theoretically satisfy these properties, verifying the required conditions can be challenging.

Minimax optimality and asymptotic normality are crucial in statistical analysis, motivating efforts to establish these properties. For example, Wager & Athey (2018) and Mourtada et al. (2020) demonstrate asymptotic normality and minimax optimality for their modified random forests, respectively, but not for the original random forest proposed by Breiman (2001). These simplifications facilitate desirable theoretical results but may degrade empirical performance.

Uniform convergence is particularly important for robustness under distributional shifts, such as covariate shift (Shimodaira, 2000), where nonparametric regression models are known to be effective. However, uniform convergence is rarely established for modern nonparametric regression estimators, including classical methods like local linear and series regression. This limitation stems from their data-adaptive nature and the prevalent use of empirical process arguments, which focus on population risk measures, such as mean squared error, rather than sup-norm bounds for uniform convergence. Schmidt-Hieber & Zamolodtchikov (2024) addresses this issue by demonstrating restricted uniform optimality for neural network regression in the one-dimensional covariate case, but their approach is not easily generalizable.

Our method shares some motivations with debiasing approaches in high-dimensional regression problems. For instance, the Lasso estimator, while widely used, introduces a bias that diminishes with sample size at a slower rate than $\sqrt{n}$, hindering asymptotic normality (Tibshirani, 1996; Bühlmann & van de Geer, 2011). To mitigate this, approaches such as van de Geer et al. (2014) incorporate bias-correction terms, yielding improved asymptotic properties like normality and efficiency (Janková & van de Geer, 2018).

This study generalizes our earlier work on doubly robust methods for nonparametric regression discontinuity design (Kato, 2024). Independently, Chernozhukov et al. (2024) develops conditional influence functions with similar motivations, building on work by Ichimura & Newey (2022).

From a technical perspective, this study is heavily influenced by Belloni et al. (2011) and Kennedy et al. (2024). Belloni et al. (2011) investigates various properties of nonparametric series estimators, which were subsequently utilized in Kennedy et al. (2024) to study minimax nonparametric estimators for conditional average treatment effects. We leverage these technical findings, alongside classical results from Stone (1980, 1982) and Tsybakov (2008).

2 The debiased estimator

We randomly split the observations $\mathcal{D}$ into two datasets, $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, satisfying $\mathcal{D} = \mathcal{D}^{(1)} \cup \mathcal{D}^{(2)}$. For simplicity, let n be an even value, and define $m = n/2$. Then, for each $\ell \in \{1, 2\}$ we denote

$$\mathcal{D}^{(\ell)} := \big\{(X_i, Y_i)\big\}_{i \in \mathcal{I}^{(\ell)}},$$

where $\mathcal{I}^{(\ell)}$ is the index set of $\mathcal{D}^{(\ell)}$; that is, $\mathcal{I}^{(1)} \cup \mathcal{I}^{(2)} = \{1, 2, \ldots, n\}$.

We consider the following three-stage estimation: for each point $x_0 \in \mathcal{X}$ of interest:

First-stage: Estimate f_0 using *any* smooth model and denote the estimator by $\widehat{f}_n$.

Second-stage: Estimate the conditional expected residual of the first-stage estimator $\mathbb{E}\big[Y - f(X) \mid X = x_0\big]$ or, equivalently, the estimation error $f_0(x_0) - \widehat{f}_n(x_0)$. We denote the estimator by

Third-stage: Summing the first and second stage estimators and obtain

$$\widetilde{f}_n(x_0) := \widehat{b}_n(x_0) + \widehat{f}_n(x_0), \tag{1}$$

as a debiased estimator of f_0.

As explained in Section 3, we can show the asymptotic normality and uniform convergence for the debiased estimator. In contrast, many of the modern machine learning methods lack those properties, even though their performances are guaranteed for the population risk.

In this section, we briefly explain the construction of the first-stage estimator $\widehat{f}_n$ and the second-stage estimator $\widehat{b}_n$.

2.1 First-stage nonparametric regression

For the first-stage estimator $\widehat{f}_n$, we can use any regression methods if $f_0 - \widehat{f}_n$ is smooth. There can be several definitions for the smoothness. In this study, for simplicity, we focus on the smoothness in the meaning of the Hölder class (Definition 3.1). We note that both f_0 and $\widehat{f}_n$ do not have to be smooth if the difference $f_0 - \widehat{f}_n$ is smooth.

We emphasize that in our analysis, we do not require specific properties for $\widehat{f}_n$ except for the smoothness. The first-stage estimator $\widehat{f}_n$ can converge to the true function f_0 with a very slow rate or even be inconsistent. Unless the smoothness conditions are satisfied, our asymptotic theoretical results hold, although the empirical finite-sample performance depends on the choice of the first stage estimator.

2.2 Second-stage conditional expected residual estimation

For each element X_j of covariates $X = (X_1\ X_2\ \cdots\ X_d)^\top \in \mathcal{X}$, let us define the Legendre polynomial series basis $\widetilde{\rho}\colon \mathbb{R} \to \mathbb{R}$ as

$$\widetilde{\rho}(X_j) := \Big(\rho_0(X_j)\ \ \rho_1(X_j)\ \ \cdots\ \ \rho_{\lfloor \ell \rfloor}(X_j)\Big)^\top,$$

where

$$\widetilde{\rho}_m(X_j) := \sum_{k=0}^{m} (-1)^{\ell+m} \sqrt{2m+1} \binom{m}{k} \binom{m+k}{k} X_j^k.$$

Then, we define $\rho(X_i)$ to be the corresponding tensor product of all interactions of $\widetilde{\rho}_m(X_{1,i}), \ldots, \widetilde{\rho}_m(X_{d,i})$.

By using the Legendre polynomial basis, we define the local linear regression as follows:

$$\begin{aligned}
\widehat{\beta}_n(x) &:= \big(\widehat{\beta}_{1,n}(x) \quad \widehat{\beta}_{2,n}(x) \quad \cdots \quad \widehat{\beta}_{\ell,n}(x)\big)^\top \\
&:= \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{\ell+1}} \sum_{i \in \mathcal{I}^{(2)}} \left(Y_i - \widehat{f}_n(X_i) - \beta^\top \rho\left(\frac{X_i - x}{h_n}\right)\right)^2 K\left(\frac{X_i - x}{h_n}\right),
\end{aligned}$$

where $K\colon \mathcal{X} \to \mathbb{R}$ is a kernel function, defined as

$$K(x) := K_h(x) := \mathbb{1}\left[\|x - x_0\| \le h\right].$$

Then, the conditional expected residual estimator corresponds to $\widehat{\beta}_{1,n}(x)$, the first element of $\widehat{\beta}_n(x)$; that is,

$$\widehat{b}_n(x) = \widehat{\beta}_{1,n}(x).$$

This second-stage estimator corresponds to the bias-correction term in the construction of semiparametric efficient estimators (Schuler & van der Laan, 2024). Such an estimator is referred to as the bias-correction or one-step estimator in the context of semiparametric statistics (van der Vaart, 2002).

2.3 Discussion

Our estimator (1) is closely related to the influence function in conditional mean estimation (Kennedy, 2023). This type of estimator is also connected to the Neyman orthogonal score (Chernozhukov et al., 2018), which is almost mathematically equivalent to the canonical gradient in the one-step bias corection (Schuler & van der Laan, 2024). Note that in conditional mean estimation, the construction of the influence function becomes complicated due to the conditioning. For the influence function including the parameter represented via the conditional expected value, see Ichimura & Newey (2022). Independently and simultaneously of us, Chernozhukov et al. (2024) investigates a related topic, with different interest and method.

3 Convergence analysis

This section presents convergence analysis of $\widetilde{f}_n$. First, we define the Hölder class as follows:

Definition 3.1 (Hölder class). *Let $\mathcal{T}$ be an interval in $\mathbb{R}^d$ and let s and L be two positive numbers. Let $D^\alpha = \frac{\partial^\alpha}{\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}}$ be the partial derivative operator. The Hölder class $\Sigma(s, L)$ on $\mathcal{T}$ is defined as the set of $\ell = \lfloor s \rfloor$ times differentiable function $f\colon \mathcal{T} \to \mathbb{R}$ satisfying the following conditions:*

- *Its derivatives up to order $\lfloor \ell \rfloor$ are bounded as*
$$|D^\alpha f(x)| \leq C < \infty \qquad \forall x \in \mathcal{X},$$
for all $\alpha = (\alpha_1, \dots, \alpha_d)$ with $\sum_j \alpha_j \leq \ell$.
- *Its ℓ-order derivatives satisfying the Lipschitz condition: there exists a constant $0 < C\infty$ such that*
$$\left|D^\beta f(x) - D^\beta f(x')\right| \leq C\|x - x'\|^{s-\ell} \qquad \forall x, x' \in \mathcal{X}$$
for all $\beta = (\beta_1, \dots, \beta_d)$ with $\sum_j \beta_j = \ell$.

Define

$$\widehat{\mathcal{B}}_n(x_0) := \frac{1}{mh_n^d} \sum_{i \in \mathcal{I}^{(2)}} \rho\left(\frac{X_i - x_0}{h_n}\right) \rho\left(\frac{X_i - x_0}{h_n}\right)^\top K\left(\frac{X_i - x_0}{h_n}\right),$$
$$\widehat{w}_h(x, x_0) := \frac{1}{h_n^d} \rho\left(0\right)^\top \widehat{\mathcal{B}}_n(x_0)^{-1} \rho\left(\frac{x - x_0}{h_n}\right) K\left(\frac{x - x_0}{h_n}\right).$$

Then, the second-stage estimator is given as

$$\widehat{b}_n(x_0) = \frac{1}{m} \sum_{i \in \mathcal{I}^{(2)}} \left(Y_i - \widehat{f}_n(X_i)\right) \widehat{w}_h(X_i, x_0).$$

3.1 Bias and variance decomposition

First, we investigate the bias and variance of $\widetilde{f}_n$. That is, we aim to analyze the bias and variance of $\widetilde{b}_n(x_0)$, denoted as

$$\text{Bias}(x_0) := \mathbb{E}\left[\widetilde{f}_n(x_0)\right] - f_0(x_0),$$
$$\text{Variance}(x_0) := \mathbb{E}\left[\left(\widetilde{f}_n(x_0) - \mathbb{E}\left[\widetilde{f}_n(x_0)\right]\right)^2\right].$$

Let us introduce the population values of $\widehat{\mathcal{B}}_{n,x}$ and $\widehat{s}_{n,x}$ as

$$\widetilde{\mathcal{B}}_n(x_0) \coloneqq \int \rho\left(\frac{x-x_0}{h_n}\right)\rho\left(\frac{x-x_0}{h_n}\right)^\top K\left(\frac{x-x_0}{h_n}\right)\mathrm{d}F(x).$$

For this value, the covariate distribution, and the smoothness of f_0 and $\widehat{f}_n$, we make the following assumptions.

Assumption 3.2. *Let $s, L, C > 0$ be constants independent of f_0 and n. Let h_n be the bandwidth of the local polynomial estimator. For a point $x_0 \in \mathcal{X}$ of interest, the following holds:*

- *$f_0 - \widehat{f}$ belongs to the Hölder class $\Sigma(s, L)$.*
- *The eigenvalues of $\widehat{\mathcal{B}}_n(x)$ is bounded above and below away from zero.*
- *There exists a constant $C > 0$ such that $\int \mathbb{1}\left[\|x - x_0\| \leq h\right]\mathrm{d}P(x) \leq Ch^d$,*

where $P(x)$ denotes a distribution of x.

Theorem 3.3 (Bias and variance decomposition)**.** *Let $s, L, C > 0$ be constants independent of f_0 and n. For the values, suppose that Assumptions 3.2 holds. Let h_n be the bandwidth of the local polynomial estimator. Then, for any $\varepsilon > 0$ and for all $x_0 \in [0,1]$, there exits $n_0 > 0$ such that for all $n \geq n_0$, with probability $1-\varepsilon$, it holds that*

$$|\mathrm{Bias}(x_0)| \leq C_1 h_n^s,$$
$$\mathrm{Variance}(x_0) \leq \frac{C_2}{nh_n^d} + C_3 h_n^{2s},$$

where $C_1, C_2, C_3 > 0$ are constants independent of f_0 and n.

The smoothness condition on $f_0 - \widehat{f}$ can be satisfied in several ways. We discuss this point in Section 4.

3.2 Pointwise convergence of the mean squared errors

Theorem 3.3 about the convergences of bias and variance terms directly indicate the convergence of the mean squared errors. This is because we have

$$\mathbb{E}\left[\left(\widehat{f}_n(x_0) - f_0(x_0)\right)^2\right] = \mathrm{Bias}(x_0)^2 + \mathrm{Variance}(x_0)$$

To minimize the mean squared error, we set the bandwidth as

$$h_n = \alpha n^{-\frac{1}{2s+d}},$$

where α is a constant independent of f_0 and n, and recall that s is the parameter of smoothness and d is the dimension of X. Then, we state the pointwise convergence of the mean squared errors as follows:

Theorem 3.4 (Pointwise mean squared error convergence)**.** *Suppose that the conditions in Theorem 3.3 holds. Let $h_n = \alpha n^{-\frac{1}{2s+d}}$. Then, for all $x_0 \in [0,1]$, it holds that*

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(s,L)} \sup_{x_0\in\mathcal{X}} \mathbb{E}_f\left[\psi_n^{-2}\left|\widetilde{f}_n(x_0) - f(x_0)\right|^2\right] \leq C < \infty,$$

where $\psi_n \coloneqq n^{-\frac{s}{2s+d}}$ is the rate of convergence and C is a constant depending only on s, L, λ_0, a_0, $\sigma^2_{\max}$, $K_{\max}$, and s.

Theorem 3.4 can be extend to the following mean squared error convergence over the distribution of X.

Corollary 3.5 (Mean squared errors over the distribution of X)**.** *Suppose that the conditions in Theorem 3.4 holds. Then, it holds that*

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(s,L)} \mathbb{E}_f\left[\psi_n^{-2}\left\|\widetilde{f}_n - f\right\|_2^2\right] \leq C < \infty,$$

where $\psi_n \coloneqq n^{-\frac{s}{2s+d}}$ is the rate of convergence and C is a constant depending only on s, L, λ_0, a_0, $\sigma^2_{\max}$, $K_{\max}$, and s.

3.3 Pointwise asymptotic normality

We now establish the asymptotic normality of $\widetilde{f}(x_0)$. This result follows directly from Theorem 3.3.

Theorem 3.6 (Asymptotic normality)**.** *Suppose one of the upper bounds in Theorem 3.4 holds. Then, it tholds that*

$$\sqrt{nh_n^d}\Big(\widetilde{f}(x_0) - f_0(x_0)\Big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V(x_0)),$$

where $V(x_0) \coloneqq nh_n^d \mathrm{Variance}(x_0) \le C_2 + o(h_n)$.

The asymptotic normality result facilitates statistical inference. However, constructing a confidence interval is challenging due to the unknown parameter s involved in h_n. To address this issue, the concept of honest confidence intervals has been developed. For example, the bootstrap method can be employed to obtain valid confidence intervals, even when certain parameters remain unspecified.

3.4 Uniform convergence

Lastly, we establish the uniform (or sup-norm) convergence of $\widetilde{f}(x_0)$. Uniform convergence bounds the supremum of the estimation error uniformly over $X \in \mathcal{X}$, without relying on the distribution of X, whereas mean squared convergence depends on the distribution of X. This theoretical guarantee is particularly useful for addressing distribution shift problems, such as covariate shift (Shimodaira, 2000). In covariate shift scenarios, the distribution of X differs between the training and test data. While mean squared convergence may be affected by such a shift, uniform convergence remains unaffected, ensuring robust performance for both training and test data. Consequently, uniform convergence is critical in discussions involving distribution shifts.

Theorem 3.7 (Uniform convergence)**.** *Suppose that f_0 belongs to a Hölder class $\Sigma(s, L)$ on $\mathcal{X}$, where $s > 0$ and $L > 0$. Let $\widehat{f}_n$ be the LP(ℓ) estimator of order $\ell = \lfloor s \rfloor$ with bandwidth*

$$h_n \coloneqq \alpha \left(\frac{\log(n)}{n}\right)^{\frac{1}{2s+d}}$$

for some $\alpha > 0$. Suppose the following:

- *Assumption 3.2 holds.*
- *ε_i is a sub-Gaussian random variable satisfying*

$$\mathbb{E}\big[\exp\big(\lambda\varepsilon_i\big)\big] \le \exp\big(K^2\lambda^2\big), \quad \forall \lambda \in \mathbb{R},$$

  for some constant $K > 0$.
- *K is a Lipschitz kernel: $K \in \Sigma(1, L_K)$ on $\mathcal{X}$ with $0 < L_K < \infty$.*

Then, there exists a constant $C < \infty$ such that

$$\limsup_{n\to\infty} \sup_{f \in \Sigma(s,L)} \mathbb{E}_f\left[\psi_n^{-2} \big\|\widetilde{f}_n - f\big\|_\infty^2\right] \le C,$$

where

$$\psi_n \coloneqq \left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+d}}.$$

Uniform convergence provides robustness against distributional shifts but has not been proven for many nonparametric regression estimators produced by machine learning methods. This limitation arises partly from their data-adaptive nature and the reliance on empirical process techniques within the empirical risk minimization framework (van de Geer, 2009). Using empirical processes and model complexity, estimation error can be evaluated in a general manner, avoiding significant dependence on specific algorithms. However, such approaches often focus on population risk, such as mean squared error, without addressing how these results translate into uniform convergence guarantees.

For example, Schmidt-Hieber & Zamolodtchikov (2024) tackle this issue by proposing an estimator based on neural networks, demonstrating restricted uniform convergence in the case of one-dimensional covariates. In contrast, our proposed method is model-free and accommodates multiple-dimensional covariates, offering broader applicability and robustness.

3.5 Double robustness

Our proposed debiased estimator possesses the property of double robustness; that is, if either $\widehat{b}_n$ or $\widehat{f}_n$ is consistent, the resulting debiased estimator $\widetilde{f}_n$ is also consistent.

Recall that the debiased estimator is defined as

$$\widetilde{f}_n(x_0) = \frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}}\Big(Y_i - \widehat{f}_n(X_i)\Big)\widehat{w}_h(X_i, x_0) + \widehat{f}_n(x_0).$$

Suppose that $\widehat{f}_n(x_0)$ converges in probability to $f^\dagger(x_0)$ and $\widehat{b}_n(x_0)$ converges in probability to $b^\dagger(x_0)$. We examine the behavior of $\widetilde{f}_n(x_0)$ under two cases: $f^\dagger(x_0) = f_0(x_0)$ and $b^\dagger(x_0) = f_0(x_0) - f^\dagger(x_0)$.

First, consider the case where $f^\dagger(x_0) = f_0(x_0)$ and $b^\dagger(x_0) \neq f_0(x_0) - f^\dagger(x_0)$. Then, it follows that

$$\widetilde{f}_n(x_0) = \frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}}\Big(Y_i - f_0(X_i)\Big)w^\dagger(X_i, x_0) + f_0(x_0) + o_p(1),$$

where $w^\dagger(X_i, x_0)$ is a function depending only on X_i and x_0. Since the first term, $\frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}}\Big(Y_i - f_0(X_i)\Big)w^\dagger(X_i, x_0)$, converges in probability to zero, $\widetilde{f}_n(x_0)$ converges in probability to $f_0(x_0)$.

Next, consider the case where $f^\dagger(x_0) \neq f_0(x_0)$ and $b^\dagger(x_0) = f_0(x_0) - f^\dagger(x_0)$. Then, it holds that

$$\widetilde{f}_n(x_0) = f_0(x_0) - f^\dagger(x_0) + f^\dagger(x_0) + o_p(1) = f_0(x_0) + o_p(1).$$

Thus, in either case, $\widetilde{f}_n(x_0)$ converges in probability to $f_0(x_0)$. We summarize this property in the following theorem.

Theorem 3.8 (Double robustness)**.** *If either $\widehat{b}_n$ or $\widehat{f}_n$ is consistent, then it holds that*

$$\widetilde{f}_n(x_0) \xrightarrow{\mathrm{p}} f_0(x_0) \qquad \text{as } n \to \infty.$$

4 Smoothness is all you need

As demonstrated above, our method requires only the smoothness of $\widehat{f} - f_0$. This assumption is satisfied when f_0 and $\widehat{f}$ are individually smooth, but it also holds in cases where f_0 and $\widehat{f}$ are not smooth, provided that their difference is smooth. Neural networks belong to this class under the analysis of (Schmidt-Hieber, 2020).

Remarkably, our results require only the smoothness of the first-stage estimator (the difference between $\widehat{f}_n$ and f_0). This assumption significantly broadens the applicability of our method. Even when f_0 and $\widehat{f}_n$ are not smooth, our method can still guarantee mean squared convergence, asymptotic normality, and uniform convergence. For instance, Imaizumi & Fukumizu (2019) discusses the convergence of nonparametric estimators using neural networks when f_0 belongs to a piecewise Hölder class. Under such a case, the liner estimators, including locally polynomial regression, are suboptimal. However, even under the case, the difference $\widehat{f} - f_0$ can be smooth, and our smoothness condition still holds.

5 Conclusion

This study developed a debiased estimator for nonparametric regression. For any smooth regression estimator, our debiased estimator gives the following desirable property: mean-squared convergence, asymptotic normality, uniform convergence, and double robustness. The asymptotic normality is useful for statistical inference, while the uniform convergence makes the nonparametric regression estimator distributionally robust. Additionally, either the first-stage or the second-stage nonparametric estimator is consistent, the resulting regression estimator is consistent; that is, doubly robust. Surprisingly, our estimator only requires the smoothness of the first-stage regression estimator. More specifically, if either (i) both the first-stage estimator $\widehat{f}$ and the true regression function f_0 are smooth or (ii) the difference $\widehat{f} - f_0$ is smooth, we can show the above desirable properties for the debiased estimator.

6 Proofs

This section provides the proofs in Section 3.

6.1 Proof of Theorem 3.6

Define

$$\begin{aligned}\text{Bias}\left(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right) &:= \mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] - f_0(x_0),\\ \text{Variance}\left(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right) &:= \mathbb{E}\left[\left(\widetilde{f}_n(x_0) - \mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right)^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right],\end{aligned}$$

$$\text{Variance}\left(\mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right) := \mathbb{E}\Big[\Big(\mathbb{E}\Big[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\Big] - \mathbb{E}\left[\mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right]\Big)^2\Big].$$

Here, the followings hold:

$$\begin{aligned}&\text{Bias}\left(x_0\right) = \mathbb{E}\left[\text{Bias}\left(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right)\right],\\ &\text{Variance}\left(x_0\right) = \mathbb{E}\left[\text{Variance}\left(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right)\right] + \text{Variance}\left(\mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right).\end{aligned}$$

We aim to show that

$$\text{Bias}\left(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right) \leq \frac{L}{m}\sum_{i\in\mathcal{I}^{(2)}}\frac{h_n^s}{\ell!}\left|\widehat{w}_h(X_i, x_0)\right|, \tag{2}$$

$$\text{Variance}\left(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right) = \frac{C_2}{nh_n^d} + o(1), \tag{3}$$

$$\text{Variance}\left(\mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right) = 2C_1^2 h^{2s}. \tag{4}$$

If (2) holds, then we have

$$\begin{aligned}\mathbb{E}\left[\frac{L}{m}\sum_{i\in\mathcal{I}^{(2)}}\frac{h_n^s}{\ell!}\left|\widehat{w}_h(X, x_0)\right|\right] &= \int L\frac{h_n^s}{\ell!}\left|\widehat{w}_h(x, x_0)\right|\mathrm{d}P(x)\mathrm{d}x\\ &= \int L\frac{h_n^s}{\ell!}h_n^d\left|\widehat{w}_h(x_0 + h_n u, x_0)\right|\mathrm{d}P(x_0 + h_n u)\mathrm{d}u,\end{aligned}$$

where $\frac{x-x_0}{h} = u$. We have

$$\begin{aligned}&\int L\frac{h_n^s}{\ell!}h_n^d\left|\widehat{w}_h(x_0 + h_n u, x_0)\right|\mathrm{d}P(x_0 + h_n u)\mathrm{d}u\\ &= L\frac{h_n^s}{\ell!}h_n^d\int\left|\frac{1}{h_n^d}\rho\left(0\right)^\top \widehat{\mathcal{B}}_n(x_0)^{-1}\rho\left(u\right)\mathbb{1}\left[\|u\| \leq 1\right]\right|\mathrm{d}P(x_0 + h_n u)\mathrm{d}u.\end{aligned}$$

6.1.1 Preliminaries

As preliminary, we first state the following lemma about the properties of the local polynomial estimator, as well as Proposition 1.12 in Tsybakov (2008) and Appendix B.3 in Kennedy et al. (2024).

Lemma 6.1 (From Proposition 1.12 in Tsybakov (2008))**.** *Let x_0 be a real number such that $\mathcal{B}_n(x_0) > 0$ and let Q be a polynomial whose degree is less than or equal to ℓ. Then, the LP(ℓ) weights $\widehat{w}_n(X_i, x_0)$ satisfy*

$$\sum_{i\in\mathcal{I}^{(2)}} Q(X_i)\widehat{w}_h(X_i, x_0) = Q(x_0).$$

for any sample $(X_1, \dots, X_n)$.

Lemma 6.2. *Under Assumption 3.2, for any ε, there exists n_0 such that for all $n \geq n_0$, $h \geq \frac{1}{2n}$, and $x_0 \in \mathcal{X}$, the weights $\widehat{w}_n(x, x_0)$ of the LP(ℓ) estimator satisfy the followings:*

- $\sup_{x,x_0}\left|\widehat{w}_h(x, x_0)\right| \leq \frac{C_*}{nh}$ *with probability* $1-\varepsilon$*;*
- $\mathbb{E}\left[\sum_{i\in\mathcal{I}^{(2)}}\left|\widehat{w}_n(X_i, x_0)\right|\right] \leq C_* + \varepsilon$*;*
- $\widehat{w}_h(x, x_0) = 0$ *if* $|x - x_0| > h$*,*

where the constant C_* depends only on λ_0, a_0, and $K_{\max}$.

6.1.2 Bounding the bias (proof of (2))

Proof. We decompose the bias term as

$$
\begin{aligned}
&\mathrm{Bias}(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n) \\
&= \mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] - f_0(x_0) \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \Big(Y_i - \widehat{f}_n(X_i)\Big)\widehat{w}_h(X_i, x_0) + \widehat{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] - f_0(x_0).
\end{aligned}
$$

We have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \Big(Y_i - \widehat{f}_n(X_i)\Big)\widehat{w}_h(X_i, x_0) + \widehat{f}_n(x_0) - f_0(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \Big(f_0(X_0) - \widehat{f}_n(X_i)\Big)\widehat{w}_h(X_i, x_0) + \widehat{f}_n(x_0) - f_0(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right].
\end{aligned}
$$

Let $\widehat{g}_n \coloneqq \widehat{f}_n - f_0$. Since $\widehat{g}_n$ belongs to the Hölder class $\Sigma(s, L)$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \Big(\widehat{g}_n(X_i) - \widehat{g}_n(x_0)\Big)\widehat{w}_h(X_i, x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \frac{\widehat{g}^{(\ell)}(x_0 + \tau_i(X_i - x_0)) - \widehat{g}^{(\ell)}(x_0)}{\ell!}\big(X_i - x_0\big)^{\ell}\widehat{w}_h(X_i, x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right].
\end{aligned}
$$

where $0 \le \tau_i \le 1$. Here, we used the Taylor expansion of $\widehat{g}$, and Lemma 6.1.

With probability one, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \frac{\widehat{g}^{(\ell)}(x_0 + \tau_i(X_i - x_0)) - \widehat{g}^{(\ell)}(x_0)}{\ell!}\big(X_i - x_0\big)^{\ell}\widehat{w}_h(X_i, x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&\le \frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \frac{L\big\|X_i - x_0\big\|^s}{\ell!}\Big|\widehat{w}_h(X_i, x_0)\Big| \\
&= \frac{1}{m}\sum_{i\in\mathcal{I}^{(2)}} \frac{L\big\|X_i - x_0\big\|^s}{\ell!}\Big|\widehat{w}_h(X_i, x_0)\Big|\mathbb{1}\big[|X_i - x_0| \le h\big] \\
&\le \frac{L}{m}\sum_{i\in\mathcal{I}^{(2)}} \frac{h_n^s}{\ell!}\Big|\widehat{w}_h(X_i, x_0)\Big|.
\end{aligned}
$$

□

6.1.3 Bounding the variance: part I (proof of (3))

Proof. We have

$$
\begin{aligned}
&\mathrm{Variance}(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n) \\
&= \mathbb{E}\left[\Big(\widetilde{f}_n(x_0) - f_0(x_0)\Big)^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&= \mathbb{E}\left[\Big(\widetilde{f}_n(x_0) - \mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\Big)^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&= \mathbb{E}\Bigg[\Bigg(\sum_{i\in\mathcal{I}^{(2)}} \Big(Y_i - \widehat{f}_n(X_i)\Big)\widehat{w}_h(X_i, x_0) + \widehat{f}_n(X_i)
\end{aligned}
$$

$$
- \mathbb{E}\left[\sum_{i\in\mathcal{I}^{(2)}}\left(Y_i - \widehat{f}_n(X_i)\right)\widehat{w}_h(X_i,x_0) + \widehat{f}_n(X_i) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\Bigg)^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\Bigg]
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\sum_{i\in\mathcal{I}^{(2)}} \xi_i \widehat{w}_h(X_i,x_0)\right)^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&= \sum_{i\in\mathcal{I}^{(2)}} \left(\widehat{w}_h(X_i,x_0)\right)^2 \mathbb{E}\big[\xi_i^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\big].
\end{aligned}
$$

From Assumption 3.2, we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{Variance}(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n)\right] \\
&\leq \frac{1}{m^2}\mathbb{E}\left[\sum_{i\in\mathcal{I}^{(2)}} \left(\widehat{w}_h(X,x_0)\right)^2 \mathbb{E}\big[\xi_i^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\big]\right] \\
&\leq \sigma_{\max}^2 \frac{1}{m}\mathbb{E}\left[\sup_{x,x_0}\left|\widehat{w}_h(x,x_0)\right|\right]\mathbb{E}\left[\left|\widehat{w}_h(X,x_0)\right|\right] \\
&\leq \frac{\sigma_{\max}^2 C_*^2}{m h_n^d} + o(1) \\
&= \frac{C_2}{n h_n^d} + o(1).
\end{aligned}
$$

□

6.1.4 Bounding the variance: part II (proof of (4))

Proof. We have

$$
\begin{aligned}
&\mathrm{Variance}\left(\mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right) \\
&= \mathbb{E}\Big[\Big(\mathbb{E}\Big[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\Big] - f_0(x_0) + f_0(x_0) - \mathbb{E}\left[\mathbb{E}\left[\widetilde{f}_n(x_0) \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right]\Big)^2\Big] \\
&= \mathbb{E}\Big[\Big(\mathrm{Bias}(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n) - \mathrm{Bias}(x_0)\Big)^2\Big] \\
&\leq 2C_1^2 h_n^{2s}.
\end{aligned}
$$

□

6.2 Proof of Theorem 3.7

Proof. We decompose the sup-norm as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{f}_n - f_0\right\|_\infty^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&\leq \mathbb{E}\left[2\left\|\widetilde{f}_n - \mathbb{E}\left[\widetilde{f}_n \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right\|_\infty^2 + 2\left\|\mathbb{E}\left[\widetilde{f}_n \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] - f_0\right\|_\infty^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] \\
&\leq 2\mathbb{E}\left[\left\|\widetilde{f}_n - \mathbb{E}\left[\widetilde{f}_n \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right\|_\infty^2 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] + 2\left(\sup_{x\in\mathcal{X}}\left|\mathrm{Bias}(x_0 \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n)\right|_\infty\right)^2 \\
&\leq 2\mathbb{E}\left[\left\|\widetilde{f}_n - \mathbb{E}\left[\widetilde{f}_n \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right\|_\infty \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right] + 2C_1^2 h_n^{2s}.
\end{aligned}
$$

Here, note that

$$
\mathbb{E}\left[\left\|\widetilde{f}_n - \mathbb{E}\left[\widetilde{f}_n \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]\right\|_\infty \mid \{X_i\}_{i\in\mathcal{I}^{(2)}}, \widehat{f}_n\right]
$$

$$= \mathbb{E}\left[\sup_{x_0 \in \mathcal{X}} \left| \sum_{i \in \mathcal{I}^{(2)}} \xi_i \widehat{w}_h(X_i, x_0) \right|^2 \mid \{X_i\}_{i \in \mathcal{I}^{(2)}}, \widehat{f}_n \right].$$

Hereafter, the proof step is the same as the one of Theorem 1.8 of Tsybakov (2008). We use the condition that ε_i is sub-Gaussian with Corollary 1.3 and Lemma 1.6 in Tsybakov (2008). In the original Corollary 1.3 and Lemma 1.6, Tsybakov (2008) assumes that ε_i is Gaussian, but we can generalize the results for sub-Gaussian random variables with minor modifications.

Finally, we have

$$\mathbb{E}\left[\left\|\widetilde{f}_n - \mathbb{E}\left[\widetilde{f}_n \mid \{X_i\}_{i \in \mathcal{I}^{(2)}}, \widehat{f}_n\right]\right\|_\infty \mid \{X_i\}_{i \in \mathcal{I}^{(2)}}, \widehat{f}_n\right] \leq C_3 \frac{\log(n)}{nh_n^d},$$

where $C_3 > 0$ is a constant independent of f and n.

We have

$$\mathbb{E}\left[\left\|\widetilde{f}_n - f_0\right\|_\infty^2 \mid \{X_i\}_{i \in \mathcal{I}^{(2)}}, \widehat{f}_n\right] \leq C_3 \frac{\log(n)}{nh} + 2C_1^2 h_n^{2s}.$$

By choosing the bandwidth as

$$h_n = \alpha \left(\frac{\log(n)}{n}\right)^{\frac{1}{2s+d}},$$

we complete the proof. □

References

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. 2

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models, 2011. 2

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. 2

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. 2

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018. 2, 4

Victor Chernozhukov, Whitney K. Newey, and Vasilis Syrgkanis. Conditional influence functions, 2024. arXiv:2412.18080. 2, 4

Hidehiko Ichimura and Whitney K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022. 2, 4

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. 7

Jana Janková and Sara van de Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336 – 2359, 2018. 2

Masahiro Kato. Doubly robust regression discontinuity designs, 2024. arXiv:2411.07978. 2

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review, 2023. arXiv: 2203.06469. 4

Edward H. Kennedy, Sivaraman Balakrishnan, James M. Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793 – 816, 2024. 2, 8

Chris A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics*, 15, 1987. 2

Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for Mondrian trees and forests. *The Annals of Statistics*, 48(4):2253 – 2276, 2020. doi: 10.1214/19-AOS1886. 2

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4), 2020. 7

Johannes Schmidt-Hieber and Petr Zamolodtchikov. Local convergence rates of the nonparametric least squares estimator with applications to transfer learning. *Bernoulli*, 30(3):1845 – 1877, 2024. 2, 6

Alejandro Schuler and Mark van der Laan. Introduction to modern causal inference, 2024. URL https://alejandroschuler.github.io/mci/introduction-to-modern-causal-inference.html. 4

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. 2, 6

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980. 2

Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10 (4):1040 – 1053, 1982. 2

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996. 2

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. 2, 8, 11

S.A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2009. 6

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202, 2014. 2

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. 2

A.W. van der Vaart. Semiparametric statistics, 2002. 2, 4

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. 2

Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer New York, NY, 2011. 2