The Annals of Statistics
2024, Vol. 52, No. 2, 793–816
https://doi.org/10.1214/24-AOS2369

MINIMAX RATES FOR HETEROGENEOUS CAUSAL EFFECT ESTIMATION

BY EDWARD H. KENNEDY[a], SIVARAMAN BALAKRISHNAN[b], JAMES M. ROBINS[c] AND LARRY WASSERMAN[d]

Department of Statistics and Data Science, Carnegie Mellon University, [a]*edward@stat.cmu.edu,* [b]*siva@stat.cmu.edu,* [c]*robins@hsph.harvard.edu,* [d]*larry@stat.cmu.edu*

Estimation of heterogeneous causal effects—that is, how effects of policies and treatments vary across subjects—is a fundamental task in causal inference. Many methods for estimating conditional average treatment effects (CATEs) have been proposed in recent years, but questions surrounding optimality have remained largely unanswered. In particular, a minimax theory of optimality has yet to be developed, with the minimax rate of convergence and construction of rate-optimal estimators remaining open problems. In this paper, we derive the minimax rate for CATE estimation, in a Hölder-smooth nonparametric model, and present a new local polynomial estimator, giving high-level conditions under which it is minimax optimal. Our minimax lower bound is derived via a localized version of the method of fuzzy hypotheses, combining lower bound constructions for nonparametric regression and functional estimation. Our proposed estimator can be viewed as a local polynomial R-Learner, based on a localized modification of higher-order influence function methods. The minimax rate we find exhibits several interesting features, including a nonstandard elbow phenomenon and an unusual interpolation between nonparametric regression and functional estimation rates. The latter quantifies how the CATE, as an estimand, can be viewed as a regression/functional hybrid.

1. Introduction. In this paper, we consider estimating the difference in regression functions

$$\tau(x) = \mathbb{E}(Y \mid X = x, A = 1) - \mathbb{E}(Y \mid X = x, A = 0) \tag{1.1}$$

from an i.i.d. sample of observations of $Z = (X, A, Y)$. Let Y^a denote the counterfactual outcome that would have been observed under treatment level $A = a$. Then, under the assumptions of consistency (i.e., $Y = Y^a$ if $A = a$), positivity (i.e., $\epsilon \leq \mathbb{P}(A = 1 \mid X) \leq 1 - \epsilon$ with probability one, for some $\epsilon > 0$), and no unmeasured confounding (i.e., $A \perp\!\!\!\perp Y^a \mid X$), the quantity $\tau(x)$ also equals the conditional average treatment effect (CATE)

$$\mathbb{E}\big(Y^1 - Y^0 \mid X = x\big).$$

The CATE $\tau(x)$ gives a more individualized picture of treatment effects compared to the overall average treatment effect (ATE) $\mathbb{E}(Y^1 - Y^0)$, and plays a crucial role in many fundamental tasks in causal inference, including assessing effect heterogeneity, constructing optimal treatment policies, generalizing treatment effects to new populations, finding subgroups with enhanced effects, and more. Further, these tasks have far-reaching implications across the sciences, from personalizing medicine to optimizing voter turnout. We refer to Hernán and Robins ([13], Chapter 4), Nie and Wager [26], Kennedy [17], and citations therein, for general discussion and review.

Received March 2022; revised December 2023.
MSC2020 subject classifications. 62G08, 62H12.
Key words and phrases. Causal inference, functional estimation, higher-order influence functions, nonparametric regression, optimal rates of convergence.

The simplest approach to CATE estimation would be to assume a low-dimensional parametric model for the outcome regression $\mathbb{E}(Y \mid X, A)$; then maximum likelihood estimates could be easily constructed, and under regularity conditions the resulting plug-in estimator would be minimax optimal. However, when X has continuous components, it is typically difficult to specify a correct parametric model, and under misspecification the previously described approach could lead to substantial bias. This suggests the need for more flexible methods. Early work in flexible CATE estimation employed semiparametric models, for example, partially linear models assuming $\tau(x)$ to be constant, or structural nested models in which $\tau(x)$ followed some known parametric form, but leaving other parts of the distribution unspecified [30, 32, 33, 38–40]. An important theme in this work is that the CATE can be much more structured and simple than the rest of the data-generating process. Specifically, the individual regression functions $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$ for each $a = 0, 1$ may be very complex (e.g., nonsmooth or nonsparse), even when the difference $\tau(x) = \mu_1(x) - \mu_0(x)$ is very smooth or sparse, or even constant or zero. We refer to Kennedy [17] for some recent discussion of this point.

More recently, there has been increased emphasis on incorporating nonparametrics and machine learning tools for CATE estimation. We briefly detail two especially relevant streams of this recent literature, based on so-called DR-Learner and R-Learner methods, both of which rely on doubly robust-style estimation. The DR-Learner is a model-free meta-algorithm first proposed by van der Laan ([38], Section 4.2), which essentially takes the components of the classic doubly robust estimator of the ATE, and rather than averaging, instead regresses on covariates. It has since been specialized to particular methods, for example, cross-validated ensembles [22], kernel [7, 20, 44] and series methods [34], empirical risk minimization [8] and linear smoothers [17]. On the other hand, the R-Learner is a flexible adaptation of the double-residual regression method originally built for partially linear models [33], with the first nonparametric version proposed by Robins et al. ([27], Section 5.2) using series methods. The R-Learner has since been adapted to RKHS regression [26], lasso [6, 43] and local polynomials [17]. Many flexible nondoubly robust methods have also been proposed in recent years, often based on inverse-weighting or direct regression estimation [1, 9, 12, 15, 19, 35, 41].

Despite the wide variety of methods available for flexible CATE estimation, questions of optimality have remained mostly unsolved. Gao and Han [10] studied minimax optimality, but in a specialized model where the propensity score has zero smoothness, and covariates are nonrandom; this model does not reflect the kinds of assumptions typically used in practice, for example, in the papers cited in the previous paragraph. Some but not all of these papers derive upper bounds on the error of their proposed CATE estimators; in the best case, these take the form of an oracle error rate (which would remain even if the potential outcomes $(Y^1 - Y^0)$ were observed and regressed on covariates), plus some contribution coming from having to estimate nuisance functions (i.e., outcome regressions and propensity scores). The fastest rates we are aware of come from Foster and Syrgkanis [8] and Kennedy [17]. Foster and Syrgkanis [8] studied global error rates, obtaining an oracle error plus sums of squared L_4 errors in all nuisance components. Kennedy [17] studied pointwise error rates, giving two main results; in the first, they obtain the oracle error plus a product of nuisance errors, while in the second, they obtain a faster rate via undersmoothing (described in more detail in Section 3.3). However, since these are all upper bounds on the errors of particular procedures, it is unknown whether these rates are optimal in any sense, and if they are not, how they might be improved upon. In this paper, we resolve these questions (via the minimax framework, in a nonparametric model that allows components of the data-generating process to be infinite-dimensional, yet smooth in the Hölder sense).

More specifically, in Section 3 we derive a lower bound on the minimax rate of CATE estimation, indicating the best possible (worst-case) performance of any estimator, in a model

where the CATE, regression function and propensity score are Hölder-smooth functions. Our derivation uses an adaptation of the method of fuzzy hypotheses, which is specially localized compared to the constructions previously used for obtaining lower bounds in functional estimation and hypothesis testing [4, 14, 16, 25, 29, 37]. In Section 4, we confirm that our minimax lower bound is tight (under some conditions), by proposing and analyzing a new local polynomial R-Learner, using localized adaptations of higher-order influence function methodology [27, 28, 31]. In addition to giving a new estimator that is provably optimal (under some conditions, e.g., on how well the covariate density is estimated), our results also confirm that previously proposed estimators were not generally optimal in this smooth nonparametric model. Our minimax rate also sheds light on the nature of the CATE as a statistical quantity, showing how it acts as a regression/functional hybrid, for example, the rate interpolates between nonparametric regression and functional estimation, depending on the relative smoothness of the CATE and nuisance functions (outcome regression and propensity score).

2. Setup and notation. We consider an i.i.d. sample of n observations of $Z = (X, A, Y)$ from distribution $\mathbb{P}$, where $X \in [0,1]^d$ denotes covariates, $A \in \{0,1\}$ a treatment or policy indicator and $Y \in \mathbb{R}$ an outcome of interest. We let F denote the distribution function of the covariate X (with density f as needed), and let

$$\pi(x) = \mathbb{P}(A = 1 \mid X = x),$$
$$\eta(x) = \mathbb{E}(Y \mid X = x),$$
$$\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$$

denote the propensity score, marginal and treatment-specific outcome regressions, respectively. We sometimes omit arguments from functions to ease notation, for example, note that $\tau = (\eta - \mu_0)/\pi$. We also index quantities by a distribution P when needed, for example, $\tau(x)$ under a particular distribution P is written $\tau_P(x)$; depending on context, no indexing means the quantity is evaluated at the true $\mathbb{P}$, for example, $\tau(x) = \tau_{\mathbb{P}}(x)$.

Our goal is to study estimation of the CATE $\tau(x) = \mu_1(x) - \mu_0(x)$ at a point $x_0 \in (0,1)^d$, with error quantified by mean absolute error

$$\mathbb{E}\big|\widehat{\tau}(x_0) - \tau(x_0)\big|.$$

As detailed in subsequent sections, we work in a nonparametric model $\mathcal{P}$ whose components are infinite-dimensional functions but with some smoothness. We say a function is s-smooth if it belongs to a Hölder class with index s; this essentially means it has $s-1$ bounded derivatives, and the highest-order derivative is continuous. To be more precise, let $\lfloor s \rfloor$ denote the largest integer strictly smaller than s, and let $D^\alpha = \frac{\partial^\alpha}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ denote the partial derivative operator. Then the Hölder class with index s contains all functions $g : \mathcal{X} \to \mathbb{R}$ that are $\lfloor s \rfloor$ times continuously differentiable, with derivatives up to order $\lfloor s \rfloor$ bounded, that is,

$$\big|D^\alpha g(x)\big| \le C < \infty$$

for all $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $\sum_j \alpha_j \le \lfloor s \rfloor$ and for all $x \in \mathcal{X}$, and with $\lfloor s \rfloor$-order derivatives satisyfing the Lipschitz condition

$$\big|D^\beta g(x) - D^\beta g(x')\big| \le C\|x - x'\|^{s - \lfloor s \rfloor}$$

for some $C < \infty$, for all $\beta = (\beta_1, \ldots, \beta_d)$ with $\sum_j \beta_j = \lfloor s \rfloor$ and for all $x, x' \in \mathcal{X}$, where for a vector $v \in \mathbb{R}^d$ we let $\|v\|$ denote the Euclidean norm. Sometimes Hölder classes are referenced by both the smoothness s and constant C, but we focus our discussion on the smoothness s and omit the constant, which is assumed finite and independent of n.

We write the squared $L_2(Q)$ norm of a function as $\|g\|_Q^2 = \int g(z)^2\,dQ(z)$. The sup-norm is denoted by $\|f\|_\infty = \sup_{z\in\mathcal{Z}} |f(z)|$. For a matrix A, we let $\|A\|$ and $\|A\|_2$ denote the operator/spectral and Frobenius norms, respectively, and let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A, respectively. We write $a_n \lesssim b_n$ if $a_n \le Cb_n$ for C a positive constant independent of n, and $a_n \asymp b_n$ if $a_n \le Cb_n$ and $b_n \le Ca_n$ (i.e., if $a_n \lesssim b_n$ and $b_n \lesssim a_n$). We write $a_n \sim b_n$ to mean that a_n and b_n are proportional, that is, $a_n = Cb_n$ for some C. We also use $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We use the shorthand $\mathbb{P}_n(f) = \mathbb{P}_n\{f(Z)\} = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ to write sample averages, and similarly $\mathbb{U}_n(f) = \mathbb{U}_n\{f(Z_1, Z_2)\} = \frac{1}{n(n-1)}\sum_{i\neq j} f(Z_i, Z_j)$ for the U-statistic measure.

3. Fundamental limits. In this section, we derive a lower bound on the minimax rate for CATE estimation. This result has several crucial implications, both practical and theoretical. First, it gives a benchmark for the best possible performance of any CATE estimator in the nonparametric model defined in Theorem 1. In particular, if an estimator is shown to attain this benchmark, then one can safely conclude the estimator cannot be improved, at least in terms of worst-case rates, without adding assumptions; conversely, if the benchmark is *not* shown to be attained, then one should continue searching for other better estimators (or better lower or upper risk bounds). Second, a tight minimax lower bound is important in its own right as a measure of the fundamental limits of CATE estimation, illustrating precisely how difficult CATE estimation is in a statistical sense. The main result of this section is given in Theorem 1 below. It is finally proved and discussed in detail in Section 3.3.

THEOREM 1. *For $x_0 \in (0, 1)^d$, let $\mathcal{P}$ denote the model where*:

1. *$f(x)$ is bounded above by a constant,*
2. *$\pi(x)$ is α-smooth,*
3. *$\mu_0(x)$ is β-smooth and*
4. *$\tau(x)$ is γ-smooth.*

Let $s \equiv (\alpha + \beta)/2$. Then for n larger than a constant depending on $(\alpha, \beta, \gamma, d)$, the minimax rate is lower bounded as

$$\inf_{\widehat{\tau}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\big|\widehat{\tau}(x_0) - \tau_P(x_0)\big| \gtrsim \begin{cases} n^{-1/(1+\frac{d}{2\gamma}+\frac{d}{4s})} & \text{if } s < \dfrac{d/4}{1+d/2\gamma}, \\ n^{-1/(2+\frac{d}{\gamma})} & \text{otherwise.} \end{cases}$$

REMARK 1. In Appendix A of the Supplementary Material [18], we also give results (both lower and upper bounds) for the model that puts smoothness assumptions on the marginal regression $\eta(x) = \mathbb{E}(Y \mid X = x)$, instead of the control regression $\mu_0(x) = \mathbb{E}(Y \mid X = x, A = 0)$. Interestingly, the minimax rates differ in these two models, but only in the regime where the regression function is more smooth than the propensity score (i.e., $\beta > \alpha$). Specifically, when η is β-smooth, the minimax rate from Theorem 1 holds but with $s = (\alpha + \beta)/2$ replaced by $\min(\alpha, s)$.

Crucially, Condition 4 allows the CATE $\tau(x)$ to have its own smoothness γ, which is necessarily at least the regression smoothness β, but can also be much larger, as described in the Introduction. We defer discussion of the details of the overall minimax rate of Theorem 1 to Section 3.3, moving first to a proof of the result.

REMARK 2. For simplicity, the lower bound result in Theorem 1 is given for a large model in which the covariate density is only bounded. However, as discussed in detail later in

Section 4 and Remark 11, for the stated rate to be attainable more conditions on the covariate density are required. In Section B.8 of the Appendix in the Supplementary Material [18], we give a particular submodel of $\mathcal{P}$ under which upper and lower bounds on the minimax rate match up to constants; it will be important in future work to further elucidate the role of the covariate density in CATE estimation.

The primary strategy in deriving minimax lower bounds is to construct distributions that are similar enough that they are statistically indistinguishable, but for which the parameter of interest is maximally separated; this implies no estimator can have error uniformly smaller than this separation. More specifically, we derive our lower bound using a localized version of the method of fuzzy hypotheses [4, 14, 16, 25, 29, 37]. In the classic Le Cam two-point method, which can be used to derive minimax lower bounds for nonparametric regression at a point [37], it suffices to consider a pair of distributions that differ locally; however, for nonlinear functional estimation, such pairs give bounds that are too loose. One instead needs to construct pairs of *mixture* distributions, which can be viewed via a prior over distributions in the model [4, 29, 37]. Our construction combines these two approaches via a localized mixture, as will be described in detail in the next subsection.

REMARK 3. In what follows, we focus on the lower bound in the low smoothness regime where $s < \frac{d/4}{1+d/2\gamma}$. The $n^{-1/(2+d/\gamma)}$ lower bound for the high smoothness regime matches the classic smooth nonparametric regression rate, and follows from a standard two-point argument, using the same construction as in Section 2.5 of Tsybakov [37].

The following lemma, adapted from Section 2.7.4 of Tsybakov [37], provides the foundation for the minimax lower bound result of this section.

LEMMA 1 (Tsybakov [37]). *Let P_λ and Q_λ denote distributions in $\mathcal{P}$ indexed by a vector $\lambda = (\lambda_1, \ldots, \lambda_k)$, with n-fold products denoted by P_λ^n and Q_λ^n, respectively. Let ϖ denote a prior distribution over λ. If*

$$H^2\left(\int P_\lambda^n \, d\varpi(\lambda), \int Q_\lambda^n \, d\varpi(\lambda)\right) \leq \alpha < 2$$

and

$$\psi(P_\lambda) - \psi(Q_{\lambda'}) \geq s > 0$$

for a functional $\psi : \mathcal{P} \mapsto \mathbb{R}$ and for all λ, λ', then

$$\inf_{\widehat{\psi}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\{\ell(|\widehat{\psi} - \psi(P)|)\} \geq \ell(s/2)\left(\frac{1 - \sqrt{\alpha(1-\alpha/4)}}{2}\right)$$

for any monotonic nonnegative loss function ℓ.

Lemma 1 illuminates the three ingredients for deriving a minimax lower bound, and shows how they interact. The ingredients are: (i) a pair of mixture distributions, (ii) the distance between their n-fold products, which is ideally small and (iii) the separation of the parameter of interest under the mixtures, which is ideally large. Finding the right minimax lower bound requires balancing these three ingredients appropriately: with too much distance or not enough separation, the lower bound will be too loose. In the following subsections, we describe these three ingredients in detail.

3.1. *Construction.* In this subsection, we detail the distributions P_λ and Q_λ used to construct the minimax lower bound. The main idea is to mix constructions for nonparametric regression and functional estimation, by perturbing the CATE with a bump at the point x_0, and to also use a mixture of perturbations of the propensity score and regression functions π and μ_0, locally near x_0.

For our lower bound results, we work in the setting where Y is binary; this is mostly to ease notation and calculations. Note, however, that this still yields a valid lower bound in the general continuous Y case, since a lower bound in the strict submodel where Y is binary is also a lower bound across the larger model $\mathcal{P}$. Importantly, when Y is binary, the density p of an observation Z can be indexed via either the quadruple (f, π, μ_0, μ_1) or (f, π, μ_0, τ); here, we make use of the latter parametrization (and in the Appendix we consider the (f, π, η, τ) parametrization). We first give the construction for the $\alpha \geq \beta$ case in the definition below, and then go on to discuss details (and in Appendix A [18] we give constructions for all other regimes).

DEFINITION 1 (Distributions P_λ and Q_λ). Let:

1. $B : \mathbb{R}^d \to \mathbb{R}$ denote a C^∞ function with $B(x) = 1$ for $x \in [-1/4, 1/4]^d$ and $B(x) = 0$ for $x \notin [-1/2, 1/2]^d$,
2. $\mathcal{C}_h(x_0)$ denote the cube centered at $x_0 \in (0, 1)^d$ with sides of length $h \leq 1/4$,
3. $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ denote a partition of $\mathcal{C}_h(x_0)$ into k cubes of equal size (for k an integer raised to the power d), with midpoints $(m_1, \ldots, m_k)$, so each cube $\mathcal{X}_j = \mathcal{C}_{h/k^{1/d}}(m_j)$ has side length $h/k^{1/d}$.

Then for $\lambda_j \in \{-1, 1\}$ define the functions

$$\tau_h(x) = h^\gamma B\left(\frac{x - x_0}{2h}\right),$$

$$\mu_{0\lambda}(x) = \frac{1}{2} + (h/k^{1/d})^\beta \sum_{j=1}^{k} \lambda_j B\left(\frac{x - m_j}{h/k^{1/d}}\right),$$

$$\pi_\lambda(x) = \frac{1}{2} + (h/k^{1/d})^\alpha \sum_{j=1}^{k} \lambda_j B\left(\frac{x - m_j}{h/k^{1/d}}\right),$$

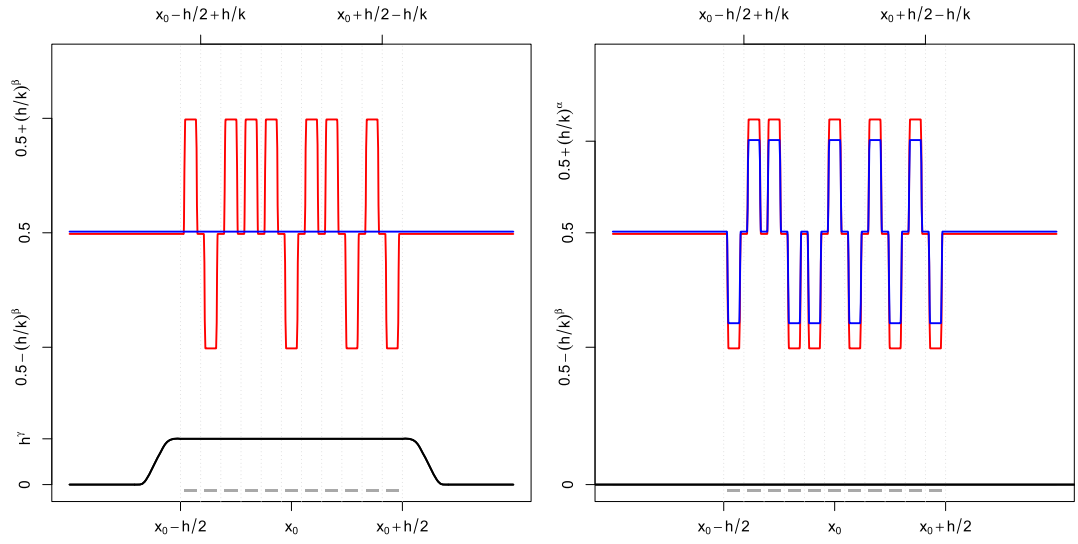
$$f(x) = \mathbb{1}(x \in \mathcal{S}_{hk}) / \left\{1 - \left(\frac{4^d - 1}{2^d}\right) h^d\right\},$$

where $\mathcal{S}_{hk} = \{\bigcup_{j=1}^{k} \mathcal{C}_{h/2k^{1/d}}(m_j)\} \cup \{[0, 1]^d \setminus \mathcal{C}_{2h}(x_0)\}$. Finally, let the distributions P_λ and Q_λ be defined via the densities

$$p_\lambda = (f, 1/2, \mu_{0\lambda} - \tau_h/2, \tau_h),$$
$$q_\lambda = (f, \pi_\lambda, \mu_{0\lambda}, 0).$$

REMARK 4. Under the (f, π, η, τ) parametrization, since $\eta = \pi\tau + \mu_0$, the above densities can equivalently be written as $p_\lambda = (f, 1/2, \mu_{0\lambda}, \tau_h)$ and $q_\lambda = (f, \pi_\lambda, \mu_{0\lambda}, 0)$.

Figure 1 shows an illustration of our construction in the $d = 1$ case. As mentioned above, the CATE is perturbed with a bump at x_0 and the nuisance functions π and μ_0 with bumps locally near x_0. The regression function μ_0 is perturbed under both P_λ and Q_λ, since it is less smooth than the propensity score in the $\alpha \geq \beta$ case. The choices of the CATE mimic those in the two-point proof of the lower bound for nonparametric regression at a point (see,



(a) Null P_λ (b) Alternative Q_λ

FIG. 1. *Minimax lower bound construction in $d = 1$ case. An example null density p_λ is displayed in panel (a) and an alternative density q_λ in panel (b). The black, red and blue lines denote the CATE,* marginal *outcome regression $\eta = \pi\tau + \mu_0$, and propensity score functions, respectively, and the gray line denotes the support of the covariate density.*

e.g., Section 2.5 of Tsybakov [37]), albeit with a particular flat-top bump function, while the choices of nuisance functions π and μ_0 are more similar to those in the lower bound for functionals such as the expected conditional covariance (cf. Section 4 of Robins et al. [29]). In this sense, our construction can be viewed as combining those for nonparametric regression and functional estimation, similar to Shen et al. [36]. In what follows, we remark on some important details.

REMARK 5. Section 3.2 of Shen et al. [36] used a similar construction for conditional variance estimation. Some important distinctions are: (i) they focused on the univariate and low smoothness setting; (ii) in that problem there is only one nuisance function, so the null can be a point rather than a mixture distribution and (iii) they use a different, arguably more complicated, approach to bound the distance between distributions. Our work can thus be used to generalize such variance estimation results to arbitrary dimension and smoothness.

First, we remark on the choice of CATE in the construction. As mentioned above, the bump construction resembles that of the standard Le Cam lower bound for nonparametric regression at a point, but differs in that we use a specialized bump function with a flat top. Crucially, this choice ensures the CATE is constant and equal to h^γ for all x in the cube $\mathcal{C}_h(x_0)$ centered at x_0 with sides of length h, and that it is equal to zero for all $x \notin \mathcal{C}_{2h}(x_0)$, that is, outside the cube centered at x_0 with side length $2h$. The fact that the CATE is constant across the top of the bump (which will be the only place where observations appear near x_0) eases Hellinger distance calculations substantially. It is straightforward to check that the CATE $\tau_h(x)$ is γ-smooth in this construction (see p. 93 of Tsybakov [37]).

REMARK 6. One example of a bump function B satisfying the conditions above is

$$B(x) = \left[1 + \frac{g\{(4x)^2 - 1\}}{g\{2 - (4x)^2\}}\right]^{-1}$$

for $g(t) = \exp(-1/t)\mathbb{1}(t > 0)$.

For the propensity score and regression functions, we similarly have

$$B\left(\frac{x-m_j}{h/k^{1/d}}\right)=\begin{cases}1 & \text{for } x\in\mathcal{C}_{h/2k^{1/d}}(m_j),\\ 0 & \text{for } x\notin\mathcal{C}_{h/k^{1/d}}(m_j)\end{cases}$$

that is, each bump equals one on the half-$h/k^{1/d}$ cube around m_j, and is identically zero outside the main larger $h/k^{1/d}$ cube around m_j. It is again straightforward to check that $\pi_\lambda(x)$ and $\mu_{0\lambda}(x)$ are α- and β-smooth, respectively.

The covariate density is chosen to be uniform, but on the set $\mathcal{S}_{hk}$ that captures the middle of all the nuisance bumps $\{\bigcup_{j=1}^k \mathcal{C}_{h/2k^{1/d}}(m_j)\}$, together with the space $\{[0,1]^d \setminus \mathcal{C}_{2h}(x_0)\}$ away from x_0. Importantly, this choice ensures there is only mass where the nuisance bumps $B(\frac{x-m_j}{h/k^{1/d}})$ are constant and nonzero (and where $\tau_h(x)=h^\gamma$), or else far away from x_0, where the densities are the same under P_λ and Q_λ. Note that, as $h\to 0$, the Lebesgue measure of the set $\mathcal{S}_{hk}$ tends to one, and the covariate density tends toward a standard uniform distribution.

3.2. *Hellinger distance.* As mentioned previously, deriving a tight minimax lower bound requires carefully balancing the distance between distributions in our construction. To this end, in this subsection we bound the Hellinger distance between the n-fold product mixtures $\int P_\lambda^n\,d\varpi(\lambda)$ and $\int Q_\lambda^n\,d\varpi(\lambda)$, for ϖ a uniform prior distribution, so that $(\lambda_1,\dots,\lambda_k)$ are i.i.d. Rademachers.

In general, these product densities can be complicated, making direct distance calculations difficult. However, Theorem 2.1 from Robins et al. [29] can be used to relate the distance between the n-fold products to those of simpler posteriors over a single observation. In the following lemma, we adapt this result to localized constructions like those in Definition 1.

LEMMA 2. *Let P_λ and Q_λ denote distributions indexed by a vector $\lambda=(\lambda_1,\dots,\lambda_k)$, and let $\mathcal{Z}=\bigcup_{j=1}^k \mathcal{Z}_j$ denote a partition of the sample space. Assume*:

1. *$P_\lambda(\mathcal{Z}_j)=Q_\lambda(\mathcal{Z}_j)=p_j$ for all λ, and*
2. *the conditional distributions $\mathbb{1}_{\mathcal{Z}_j}\,dP_\lambda/p_j$ and $\mathbb{1}_{\mathcal{Z}_j}\,dQ_\lambda/p_j$ (given an observation is in $\mathcal{Z}_j$) do not depend on λ_ℓ for $\ell\neq j$, and only differ on partitions $j\in S\subseteq\{1,\dots,k\}$.*

For a prior distribution ϖ over λ, let $\overline{p}=\int p_\lambda\,d\varpi(\lambda)$ and $\overline{q}=\int q_\lambda\,d\varpi(\lambda)$, and define

$$\delta_1=\max_{j\in S}\sup_\lambda\int_{\mathcal{Z}_j}\frac{(p_\lambda-\overline{p})^2}{p_\lambda p_j}\,d\nu,$$

$$\delta_2=\max_{j\in S}\sup_\lambda\int_{\mathcal{Z}_j}\frac{(q_\lambda-p_\lambda)^2}{p_\lambda p_j}\,d\nu,$$

$$\delta_3=\max_{j\in S}\sup_\lambda\int_{\mathcal{Z}_j}\frac{(\overline{q}-\overline{p})^2}{p_\lambda p_j}\,d\nu$$

for a dominating measure ν. If $\overline{p}/p_\lambda\le b<\infty$ and $np_j\max(1,\delta_1,\delta_2)\le b$ for all j, then $H^2(\int P_\lambda^n\,d\varpi(\lambda),\int Q_\lambda^n\,d\varpi(\lambda))$ is bounded above by

$$Cn\Big(\sum_{j\in S}p_j\Big)\Big\{n\Big(\max_{j\in S}p_j\Big)(\delta_1\delta_2+\delta_2^2)+\delta_3\Big\}$$

for a constant C only depending on b.

In the next proposition, we bound the quantities from Lemma 2 and put the results together to obtain a bound on the desired Hellinger distance between product mixtures.

PROPOSITION 1. *Assume $h \le 1/4$ and $h^\gamma + 2(h/k^{1/d})^\beta \le 1 - 4\varepsilon$ for some $\varepsilon \in (0, 1/4)$, and take $h^\gamma = 4(h/k^{1/d})^{2s}$ for $s \equiv (\alpha+\beta)/2$. Then for the distributions P_λ and Q_λ from Definition 1, with ϖ the uniform distribution over $\{-1, 1\}^k$, the quantities δ_j from Lemma 2 satisfy*

$$\delta_1 \le \left(\frac{2^{d+1}\|B\|_2^2}{\varepsilon}\right)(h/k^{1/d})^{2\beta}, \qquad \delta_2 \le \left(\frac{2^{d+1}\|B\|_2^2}{\varepsilon}\right)(h/k^{1/d})^{2\alpha}, \qquad \delta_3 = 0,$$

and $p_j \le 2(h/2)^d/k$. Further, $H^2(\int P_\lambda^n \, d\varpi(\lambda), \int Q_\lambda^n \, d\varpi(\lambda))$ is bounded above by

$$4C\left(\frac{2\|B\|_2^2}{\varepsilon}\right)^2\left(\frac{n^2h^{2d}}{k}\right)\{(h/k^{1/d})^{4s} + (h/k^{1/d})^{4\alpha}\}$$

for C a constant only depending on ε.

Before moving to the proof of Proposition 1, we briefly discuss and give some remarks. Compared to the Hellinger distance arising in the average treatment effect or expected conditional covariance lower bounds [29], there is an extra h^d factor in the numerator. Of course, one cannot simply repeat those calculations with k/h^d bins, since then, for example, the $k^{-4s/d}$ term would also be inflated to $(k/h^d)^{-4s/d}$; our carefully localized construction is crucial to obtain the right rate in this case. We also note that the choice $h^\gamma = 4(h/k^{1/d})^{2s}$ is required for ensuring that the averaged densities $\overline{p}(z)$ and $\overline{q}(z)$ are equal (implying that $\delta_3 = 0$); specifically this equalizes the CATE bump under P_λ with the squared nuisance bumps under Q_λ.

PROOF. Here, the relevant partition of the sample space $\mathcal{X} \times \mathcal{A} \times \mathcal{Y} = [0,1]^d \times \{0,1\} \times \{0,1\}$ is $\mathcal{Z}_j = \mathcal{C}_{h/2k^{1/d}}(m_j) \times \{0,1\} \times \{0,1\}$, $j = 1, \dots, k$, along with $\mathcal{Z}_j'$, which partitions the space $[0,1]^d/\mathcal{C}_{2h}(x_0)$ away from x_0 into disjoint cubes with side lengths $h/2k^{1/d}$. Thus we have

$$P_\lambda(\mathcal{Z}_j) = P_\lambda(\mathcal{Z}_j') = Q_\lambda(\mathcal{Z}_j) = Q_\lambda(\mathcal{Z}_j') = p_j,$$

where $p_j = \int \mathbb{1}\{x \in C_{h/2k^{1/d}}(m_j)\} f(x)\,dx$. In Appendix Section B.1, we show that $(h/2)^d/k \le p_j \le 2(h/2)^d/k$ when $h \le 1/4$, and so is proportional to the volume of a cube with side lengths $h/2k^{1/d}$. Further, the conditional distributions $\mathbb{1}_{\mathcal{Z}_j}\,dP_\lambda/p_j$ and $\mathbb{1}_{\mathcal{Z}_j}\,dQ_\lambda/p_j$ do not depend on λ_ℓ for $\ell \ne j$, since λ_j only changes the density in $\mathcal{Z}_j$. In what follows, we focus on the partitions $\mathcal{Z}_j$ and not $\mathcal{Z}_j'$, since the distributions are exactly equal on the latter (in the language of Lemma 1, the set S indexes only the partitions $\mathcal{Z}_j$, and note that for this set we have $\sum_{j\in S} p_j = kp_j \le 2(h/2)^d$). Note when $(\lambda_1, \dots, \lambda_k)$ are i.i.d. Rademacher random variables the marginalized densities from Lemma 2 are

$$\overline{p}(z) = f(x)\left\{\frac{1}{4} + (2a-1)(2y-1)\frac{h^\gamma}{4}B\left(\frac{x - x_0}{2h}\right)\right\},$$

$$\overline{q}(z) = f(x)\left\{\frac{1}{4} + (2a-1)(2y-1)(h/k^{1/d})^{2s}\sum_{j=1}^k B\left(\frac{x - m_j}{h/k^{1/d}}\right)^2\right\}.$$

The first step is to show that relevant densities and density ratios are appropriately bounded. We give these details in Appendix Section B.1 [18]. Next, it remains to bound the quantities δ_1, δ_2 and δ_3.

We begin with δ_3, the distance between marginalized densities $\overline{p}$ and $\overline{q}$, which is tackled somewhat differently from δ_1 and δ_2. Because we take $(h/k^{1/d})^{2s} = h^\gamma/4$, it follows that

$\overline{q}(z) - \overline{p}(z)$ equals

$$(3.1) \qquad (2a-1)(2y-1)f(x)\left\{(h/k^{1/d})^{2s}\sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 - \frac{h^\gamma}{4}B\left(\frac{x-x_0}{2h}\right)\right\} = 0,$$

since $f(x) = 0$ for $x \notin \mathcal{S}_{hk}$ and

$$B\left(\frac{x-m_j}{h/k^{1/d}}\right) = B\left(\frac{x-x_0}{2h}\right) = 0 \quad \text{for } x \in \left\{[0,1]^d \setminus \mathcal{C}_{2h}(x_0)\right\},$$

$$B\left(\frac{x-m_j}{h/k^{1/d}}\right) = B\left(\frac{x-x_0}{2h}\right) = 1 \quad \text{for } x \in \bigcup_{j=1}^{k} \mathcal{C}_{h/2k^{1/d}}(m_j).$$

We note that this result requires a carefully selected relationship between h and k, which guarantees that the squared nuisance bumps under Q_λ equal the CATE bumps under P_λ. This also exploits the flat-top bump functions we use, together with a covariate density that only puts mass at these tops, so that the squared terms are constant and no observations occur elsewhere where the bumps are not equal. Without these choices of bump function and covariate density, the expression in (3.1) would only be bounded by h^γ, and so δ_3 would only be bounded by $h^{2\gamma}$; in that case, the δ_3 term would dominate the Hellinger bound in Lemma 2, and the resulting minimax lower bound would reduce to the oracle rate $n^{-1/(2+d/\gamma)}$, which is uninformative in the low-smoothness regimes we are considering.

Now we move to the distance δ_1, which does not end up depending on h and is somewhat easier to handle. For it, we have

$$\begin{aligned}
\delta_1 &\le \left(\frac{2^d k}{h^d}\right)\max_\ell \sup_\lambda \int_{\mathcal{X}_\ell} \sum_{a,y} \frac{f(x)^2}{4p_\lambda(z)}(h/k^{1/d})^{2\beta}\sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 dx \\
&\le \left(\frac{2^d k}{h^d}\right)\left(\frac{2}{\varepsilon}\right)(h/k^{1/d})^{2\beta}\max_\ell \int_{\mathcal{X}_\ell} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 dx \\
&\le \left(\frac{2^d \|B\|_2^2}{\varepsilon/2}\right)(h/k^{1/d})^{2\beta},
\end{aligned}$$

where the first line follows by definition, and since $p_\ell \ge (h/2)^d/k$, and $B(\frac{x-m_j}{h/k^{1/d}}) = 0$ outside of the cube $\mathcal{C}_{h/k^{1/d}}(m_j)$, which implies that

$$\left\{\sum_j \lambda_j B\left(\frac{x-m_j}{h/k^{1/d}}\right)\right\}^2 = \sum_j \lambda_j^2 B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2,$$

the inequality in the second line since $p_\lambda(z)/f(x) \ge \varepsilon$ and $f(x) \le 2$ as in (B.1)and (B.2), and the last inequality since

$$\int_{\mathcal{X}_\ell} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 dx = \int_{\mathcal{X}_\ell} B\left(\frac{x-m_\ell}{h/k^{1/d}}\right)^2 dx \le \frac{h^d}{k}\int B(u)^2\,du$$

by a change of variables.

For δ_2, we use a mix of the above logic for δ_3 and δ_1. Note that $(q_\lambda - p_\lambda)^2$ equals

$$f(x)^2\left[(a-1/2)(h/k^{1/d})^\alpha \sum_{j=1}^{k} \lambda_j B\left(\frac{x-m_j}{h/k^{1/d}}\right)\right.$$

$$+ (2a-1)(2y-1)\left\{(h/k^{1/d})^{2s}\sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 - \frac{h^\gamma}{4}B\left(\frac{x-x_0}{2h}\right)\right\}\Bigg]^2$$

$$\leq (1/2)\,f(x)^2(h/k^{1/d})^{2\alpha}\sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2,$$

where we used the fact that $(a+b)^2 \leq 2(a^2+b^2)$ and $\{\sum_j \lambda_j B(\frac{x-m_j}{h/k^{1/d}})\}^2 = \sum_j B(\frac{x-m_j}{h/k^{1/d}})^2$, along with the same logic as above with δ_3 (ensuring the second term in the square equals zero). Now we have

$$\delta_2 \leq \left(\frac{2^d k}{h^d}\right)\left(\frac{2}{\varepsilon}\right)(h/k^{1/d})^{2\alpha}\max_{\ell}\int_{\mathcal{X}_\ell}\sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 dx$$

$$\leq \left(\frac{2^d\|B\|_2^2}{\varepsilon/2}\right)(h/k^{1/d})^{2\alpha}$$

using the exact same logic as for δ_1. Plugging these bounds on $(\delta_1,\delta_2,\delta_3)$ into Lemma 2, together with the fact that $p_j \leq 2(h/2)^d/k$ and $\sum_{j\in S} p_j \leq 2(h/2)^d$, yields the result. □

3.3. *Choice of parameters and final rate.* Finally, we detail how the parameters h and k can be chosen to ensure the Hellinger distance from Proposition 1 remains bounded, and use the result to finalize the proof of Theorem 1.

PROPOSITION 2. *Let*

$$\frac{h}{k^{1/d}} = \left(\frac{h^\gamma}{4}\right)^{1/2s} = \left(\frac{1}{C^* n^2}\right)^{\frac{1}{d+4s+2sd/\gamma}}$$

for $C^* = 2^{2d/\gamma+5}C(\|B\|_2^2/\varepsilon)^2$ *and* C *the constant from Proposition 1. Then under the assumptions of Proposition 1 we have*

$$H^2\left(\int P_\lambda^n\, d\varpi(\lambda), \int Q_\lambda^n\, d\varpi(\lambda)\right) \leq 1$$

and $h^\gamma = 4(\sqrt{C^*}n)^{-1/(1+\frac{d}{2\gamma}+\frac{d}{4s})}$.

The proof of Proposition 2 follows directly from Proposition 1, after plugging in the selected values of h and k. Importantly, it (together with the alternative construction for the $\beta > \alpha$ case given in Appendix A) also settles the proof of Theorem 1 via Lemma 1. This follows since, with the proposed choices of h and k, the Hellinger distance is appropriately bounded so that the term $(1-\sqrt{\alpha(1-\alpha/4)})/2 = (1-\sqrt{3/4})/2 \approx 0.067$ in Lemma 1 is a constant (greater than 1/20, e.g.), while the separation in the CATE at x_0, which equals h^γ, is proportional to $n^{-1/(1+d/2\gamma+d/4s)}$ under all P_λ and Q_λ. Therefore, this separation is indeed the minimax rate in the low smoothness regime where $s < (d/4)/(1+d/2\gamma)$. Note again that, as discussed in Remark 3, when $s > (d/4)/(1+d/2\gamma)$ the rate $n^{-1/(1+d/2\gamma+d/4s)}$ is faster than the usual nonparametric regression rate $n^{-1/(2+d/\gamma)}$, and so the standard lower bound construction as in Section 2.5 of Tsybakov [37] indicates that the slower rate $n^{-1/(2+d/\gamma)}$ is the tighter lower bound in that regime.

Figure 2 illustrates the minimax rate from Theorem 1, as a function of the average nuisance smoothness s/d (scaled by dimension), and the CATE smoothness scaled by dimension γ/d. A number of important features about the rate are worth highlighting.

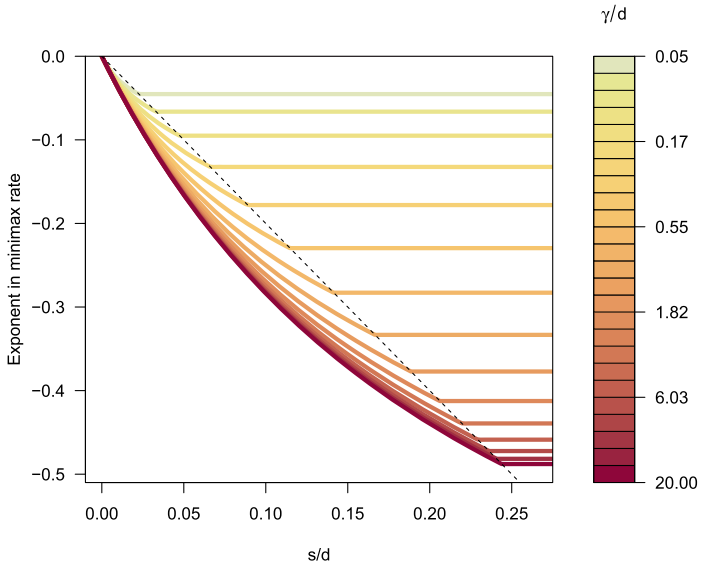


FIG. 2. *The minimax rate for CATE estimation, as a function of average nuisance smoothness s and CATE smoothness γ, each scaled by covariate dimension d. The black dotted line denotes a threshold on the nuisance smoothness s/d, below which the oracle nonparametric regression rate $n^{-1/(2+d/\gamma)}$ is unachievable (the "elbow" phenomenon).*

First, of course, the rate never slows with higher nuisance smoothness s/d, for any CATE smoothness γ/d, and vice versa. However, there is an important elbow phenomenon, akin to that found in functional estimation problems [3, 4, 29, 37]. In particular, the minimax lower bound shows that when the average nuisance smoothness is low enough that $s < \frac{d/4}{1+d/2\gamma}$, the oracle rate $n^{-1/(2+d/\gamma)}$ (which could be achieved if one actually observed the potential outcomes) is in fact unachievable. This verifies a conjecture in Kennedy [17].

Notably, though, the elbow phenomenon we find in the problem of CATE estimation differs quite substantially from that for classic pathwise differentiable functionals. For the latter, the rate is parametric (i.e., $n^{-1/2}$) above some threshold, and nonparametric ($n^{-1/(1+d/4s)}$) below. In contrast, in our setting the rate matches that of *nonparametric regression* above the threshold, and otherwise is a *combination* of nonparametric regression and functional estimation rates. Thus in this problem there are many elbows, with the threshold depending on the CATE smoothness γ. In particular, our minimax rate below the threshold,

$$n^{-1/(1+\frac{d}{2\gamma}+\frac{d}{4s})},$$

is a mixture of the nonparametric regression rate $n^{-1/(1+d/2\gamma)}$ (on the squared scale) and the classic functional estimation rate $n^{-1/(1+d/4s)}$. This means, for example, that in regimes where the CATE is very smooth, for example, $\gamma \to \infty$, the CATE estimation problem begins to resemble that of pathwise-differentiable functional estimation, where the elbow occurs at $s > d/4$, with rates approaching the parametric rate $n^{-1/2}$ above, and the functional estimation rate $n^{-1/(1+d/4s)}$ below. At the other extreme, where the CATE does not have any extra smoothness, so that $\gamma \to \beta$ (note we must have $\gamma \geq \beta$), the elbow threshold condition holds for *any* $\alpha \geq 0$. Thus, at this other extreme, there is no elbow phenomenon, and the CATE estimation problem resembles that of smooth nonparametric regression, with optimal rate $n^{-1/(2+d/\beta)}$. For the arguably more realistic setting, where the CATE smoothness γ may take intermediate values between β and ∞, the minimax rate is a mixture, interpolating between the two extremes. All of this quantifies the sense in which the CATE can be viewed as a regression/functional hybrid.

It is also worth mentioning that no estimator previously proposed in the literature (that we know of) attains the minimax rate in Theorem 1 in full generality. Some estimators have been shown to attain the oracle rate $n^{-1/(2+d/\gamma)}$, but only under stronger assumptions than the minimal condition we find here, that is, that $s > \frac{d/4}{1+d/2\gamma}$. One exception is the undersmoothed R-learner estimator analyzed in Kennedy [17], which did achieve the rate $n^{-1/(2+d/\gamma)}$ whenever $s > (d/4)/(1+d/2\gamma)$, under some conditions (including that $\alpha \geq \beta$). However, in the low-smoothness regime where $s < (d/4)/(1+d/2\gamma)$, that estimator's rate was $n^{-2s/d}$, which is slower than the minimax rate we find here. This motivates our work in the following section, where we propose and analyze a new estimator, whose error matches the minimax rate in much greater generality (under some conditions, e.g., on how well the covariate density is estimated).

REMARK 7. A slightly modified version of our construction also reveals that, when the CATE $\tau(x) = \tau$ is constant, the classic functional estimation rate $n^{-1/(1+\frac{d}{4s})}$ acts as a minimax lower bound. To the best of our knowledge, this result has not been noted elsewhere.

4. Attainability. In this section, we show that the minimax lower bound of Theorem 1 is actually attainable, via a new local polynomial version of the R-Learner [17, 26], based on an adaptation of higher-order influence functions [27, 28, 31].

4.1. *Proposed estimator and decomposition.* In this subsection, we first describe our proposed estimator, and then give a preliminary error bound, which motivates the specific bias and variance calculations in the following subsections. In short, the estimator is a higher-order influence function-based version of the local polynomial R-learner analyzed in Kennedy [17]. At its core, the R-Learner essentially regresses outcome residuals on treatment residuals to estimate a weighted average of the CATE. Early versions for a constant or otherwise parametric CATE were studied by Chamberlain [5], Robinson [33], and Robins [30], with more flexible series, RKHS and lasso versions studied more recently by Robins et al. [27], Nie and Wager [26] and Chernozhukov et al. [6], respectively. None of this previous work obtained the minimax optimal rates in Theorem 2.

DEFINITION 2 (Higher-order local polynomial R-Learner). Let $K_h(x) = \frac{1}{h^d}\mathbb{1}(\|x - x_0\| \leq h/2)$. For each covariate x_j, $j = 1, \ldots, d$, define $\rho(x_j) = \{\rho_0(x_j), \rho_1(x_j), \ldots, \rho_{\lfloor\gamma\rfloor}(x_j)\}^{\mathrm{T}}$ as the first $(\lfloor\gamma\rfloor + 1)$ terms of the Legendre polynomial series (shifted to be orthonormal on $[0, 1]$),

$$\rho_m(x_j) = \sum_{\ell=0}^{m} (-1)^{\ell+m} \sqrt{2m+1} \binom{m}{\ell} \binom{m+\ell}{\ell} x_j^{\ell}.$$

Define $\rho(x)$ to be the corresponding tensor product of all interactions of $\rho(x_1), \ldots, \rho(x_d)$ up to order $\lfloor\gamma\rfloor$, which has length $q = \binom{d+\lfloor\gamma\rfloor}{\lfloor\gamma\rfloor}$ and is orthonormal on $[0, 1]^d$, and finally define $\rho_h(x) = \rho(1/2 + (x - x_0)/h)$. The proposed estimator is then defined as

$$\widehat{\tau}(x_0) = \rho_h(x_0)^{\mathrm{T}} \widehat{Q}^{-1} \widehat{R}, \tag{4.1}$$

where $\widehat{Q}$ is a $q \times q$ matrix and $\widehat{R}$ a q-vector given by

$$\begin{aligned}
\widehat{Q} &= \mathbb{P}_n\{\rho_h(X)K_h(X)\widehat{\varphi}_{a1}(Z)\rho_h(X)^{\mathrm{T}}\} \\
&\quad + \mathbb{U}_n\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{a2}(Z_1, Z_2)K_h(X_2)\rho_h(X_1)^{\mathrm{T}}\}, \\
\widehat{R} &= \mathbb{P}_n\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y1}(Z_1)\} + \mathbb{U}_n\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y2}(Z_1, Z_2)K_h(X_2)\},
\end{aligned}$$

respectively, and

$$\begin{aligned}
\widehat{\varphi}_{a1}(Z) &= A\{A - \widehat{\pi}(X)\},\\
\widehat{\varphi}_{y1}(Z) &= \{Y - \widehat{\mu}_0(X)\}\{A - \widehat{\pi}(X)\},\\
\widehat{\varphi}_{a2}(Z_1, Z_2) &= -\{A_1 - \widehat{\pi}(X_1)\}b_{hk}(X_1)^{\mathrm{T}}\widehat{\Omega}^{-1}b_{hk}(X_2)A_2,\\
\widehat{\varphi}_{y2}(Z_1, Z_2) &= -\{A_1 - \widehat{\pi}(X_1)\}b_{hk}(X_1)^{\mathrm{T}}\widehat{\Omega}^{-1}b_{hk}(X_2)\{Y_2 - \widehat{\mu}_0(X_2)\},\\
b_{hk}(x) &= b\{1/2 + (x - x_0)/h\}\mathbb{1}\big(\|x - x_0\| \le h/2\big),\\
\widehat{\Omega} &= \int_{v\in[0,1]^d} b(v)b(v)^{\mathrm{T}}\, d\widehat{F}\big(x_0 + h(v - 1/2)\big)
\end{aligned}$$

for $b : \mathbb{R}^d \mapsto \mathbb{R}^k$ a basis of dimension k. The nuisance estimators $(\widehat{F}, \widehat{\pi}, \widehat{\mu}_0)$ are constructed from a separate training sample D^n, independent of that on which $\mathbb{U}_n$ operates.

The estimator in Definition 2 can be viewed as a *localized* higher-order estimator, and depends on two main tuning parameters: the bandwidth h, which controls how locally one averages near x_0, and the dimension k of the basis b, which controls how bias and variance are balanced in the second-order U-statistic terms in $\widehat{Q}$ and $\widehat{R}$. Specific properties of the basis b are discussed shortly, for example, just prior to Remark 7 and in (4.4). We also note that while the basis b will have dimension k growing with sample size, the dimension of the basis ρ is fixed depending on the smoothness of the CATE; for the latter, we use the Legendre series for concreteness, but expect other bases to work as well.

The U-statistic terms are important for debiasing the first-order sample average terms. In addition, our proposed estimator can be viewed as estimating a locally weighted projection parameter $\tau_h(x_0) = \rho_h(x_0)^{\mathrm{T}}\theta$, with coefficients given by

$$\arg\min_{\beta} \mathbb{E}\big[K_h(x)\pi(x)\{1 - \pi(x)\}\{\tau(x) - \beta^{\mathrm{T}}\rho_h(x)\}^2\big] = Q^{-1}R \tag{4.2}$$

for

$$\begin{aligned}
Q &= \int \rho_h(x)K_h(x)\pi(x)\{1 - \pi(x)\}\rho_h(x)^{\mathrm{T}}\, dF(x),\\
R &= \int \rho_h(x)K_h(x)\pi(x)\{1 - \pi(x)\}\tau(x)\, dF(x).
\end{aligned}$$

In other words, this projection parameter $\tau_h(x_0)$ is a $K_h(x)\pi(x)\{1 - \pi(x)\}$-weighted least squares projection of the CATE $\tau(x)$ on the scaled Legendre polynomials $\rho_h(x)$. Crucially, since $\rho_h(x)$ includes polynomials in x up to order $\lfloor\gamma\rfloor$, the projection parameter is within h^γ of the target CATE; this is formalized in the following proposition.

PROPOSITION 3. *Let $\tau_h(x) = \rho_h(x)^{\mathrm{T}}Q^{-1}R$ denote the x_0-specific projection parameter from* (4.2), *and assume*:

1. *$\tau(x)$ is γ-smooth,*
2. *the eigenvalues of Q are bounded below away from zero, and*
3. *$\int \mathbb{1}\{\|x - x_0\| \le h/2\}\, dF(x) \lesssim h^d$.*

Then for any x with $\|x - x_0\| \le Ch$, we have

$$\big|\tau_h(x) - \tau(x)\big| \lesssim h^\gamma.$$

PROOF. This proof follows from a higher-order kernel argument (e.g., Proposition 1.13 of Tsybakov [37], Proposition 4.1.5 of Giné and Nickl [11]), after noting that we can treat $K_h(x)\pi(x)\{1-\pi(x)\}$ itself as a kernel. A similar result was also proved in Kennedy [17]. A detailed proof is given in Appendix B.3 [18]. □

In Proposition S5 in the Appendix [18], we give simple sufficient conditions under which the eigenvalues of Q are bounded. In short, this holds under standard boundedness conditions on the propensity score and covariate density.

As mentioned above, our estimator (4.1) can be viewed as a modified higher-order estimator. More specifically, $\widehat{S} = \widehat{R} - \widehat{Q}\theta$ is closely related to a second-order estimator of the moment condition

$$\mathbb{E}\big[\rho_h(X)K_h(X)\{A-\pi(X)\}\{Y-\mu_0(X)-A\tau(X)\}\big] = R - Q\theta = 0 \tag{4.3}$$

under the assumption that $\tau(x) = \rho_h(x)^{\mathrm{T}}\theta$ (this is not exactly true in our case, but it is enough that it is approximately true locally near x_0, as will be proved shortly). Indeed, letting $\widehat{Q}_1 = \mathbb{P}_n\{\rho_h(X)K_h(X)\widehat{\varphi}_{a1}(Z)\rho_h(X)^{\mathrm{T}}\}$ and $\widehat{R}_1 = \mathbb{P}_n\{\rho_h(X)K_h(X)\widehat{\varphi}_{y1}(Z)\}$ denote the first terms in $\widehat{Q} = \widehat{Q}_1 + \widehat{Q}_2$ and $\widehat{R} = \widehat{R}_1 + \widehat{R}_2$, respectively, we see that $\widehat{S}_1 = \widehat{R}_1 - \widehat{Q}_1\theta$ is a usual first-order influence function-based estimator of the moment condition (4.3). Similarly, the second terms $\widehat{Q}_2$ and $\widehat{R}_1$ are akin to the second-order U-statistic corrections that would be added using the higher-order influence function methodology developed by Robins et al. [27, 28, 31]. However, these terms differ in two important ways, both relating to localization near x_0. First, the U-statistic is localized with respect to both X_1 and X_2, that is, the product $K_h(X_1)K_h(X_2)$ is included, whereas only $K_h(X_1)$ would arise if the goal were purely to estimate the moment condition (4.3) for fixed h. Second, the basis functions

$$b_{hk}(x) = b\bigg(1/2 + \frac{x-x_0}{h}\bigg)\mathbb{1}\big(\|x-x_0\| \le h/2\big)$$

appearing in $\widehat{\varphi}_{a2}$, $\widehat{\varphi}_{y2}$ and $\widehat{\Omega}$ are localized; they only operate on Xs near x_0, stretching them out so as to map the cube $[x_0 - h/2, x_0 + h/2]^d$ around x_0 to the whole space $[0,1]^d$ (e.g., $b_{hk}(x_0 - h/2) = b(0)$, $b_{hk}(x_0) = b(1/2)$, etc.). This is the same localization that is used with the Legendre basis $\rho(x)$. In this sense, these localized basis terms spend all their approximation power locally rather than globally away from x_0. (Specific approximating properties we require of b will be detailed shortly, in (4.4)). These somewhat subtle distinctions play a crucial role in appropriately controlling bias, as will be described in more detail shortly.

REMARK 8. Note again that, as with other higher-order estimators, the estimator (4.1) depends on an initial estimate of the covariate distribution F (near x_0), through $\widehat{\Omega}$. Importantly, we do not take this estimator $\widehat{F}$ to be the empirical distribution, in general, since then our optimal choices of the tuning parameter k would yield $\widehat{\Omega}$ noninvertible; this occurs with higher-order estimators of pathwise differentiable functionals as well [23]. As discussed in Remark 11, and in more detail shortly, we do give conditions under which the estimation error in $\widehat{\Omega}$ or $\widehat{F}$ does not impact the overall rate of $\widehat{\tau}(x_0)$.

Crucially, Proposition 3 allows us to focus on understanding the estimation error in $\widehat{\tau}(x_0)$ with respect to the projection parameter $\tau_h(x_0)$, treating h^γ as a separate approximation bias. The next result gives a finite-sample bound on this error, showing how it is controlled by the error in estimating the components of Q and R.

PROPOSITION 4. *Let $S = R - Q(Q^{-1}R)$ and $\widehat{S} = \widehat{R} - \widehat{Q}(Q^{-1}R)$. The estimator (4.1) satisfies*

$$\big|\widehat{\tau}(x_0) - \tau_h(x_0)\big| \le \big\|\rho(1/2)\big\|\big(\big\|Q^{-1}\big\| + \big\|\widehat{Q}^{-1} - Q^{-1}\big\|\big)\big\|\widehat{S} - S\big\|,$$

and if $\|Q^{-1}\|$ *and* $\|\widehat{Q}^{-1} - Q^{-1}\|$ *are bounded above, then*

$$\mathbb{E}|\widehat{\tau}(x_0) - \tau_h(x_0)| \lesssim \max_j \sqrt{\mathbb{E}\{\mathbb{E}(\widehat{S}_j - S_j \mid D^n)^2 + \operatorname{var}(\widehat{S}_j \mid D^n)\}}.$$

for D^n *a separate independent training sample on which* $(\widehat{F}, \widehat{\pi}, \widehat{\mu}_0)$ *are estimated.*

Thus Proposition 4 tells us that bounding the conditional bias and variance of $\widehat{S} = \widehat{R} - \widehat{Q}(Q^{-1}R)$ will also yield finite-sample bounds on the error in $\widehat{\tau}(x_0)$, relative to the projection parameter $\tau_h(x_0)$. These bias and variance bounds will be derived in the following two subsections.

4.2. *Bias.* In this subsection, we derive bounds on the conditional bias of the estimator $\widehat{S} = \widehat{R} - \widehat{Q}(Q^{-1}R)$, relative to the components of the projection parameter (4.2), given the training sample D^n. The main ideas behind the approach are to use localized versions of higher-order influence function arguments, along with a specialized localized basis construction, which results in smaller bias due to the fact that the bases only need to be used in a shrinking window around x_0.

Here, we rely on the basis $b(x)$ having optimal Hölder approximation properties, In particular, we assume the approximation error of projections in L_2 norm satisfies

$$\|(I - \Pi_b)g\|_{F^*} \lesssim k^{-s/d} \quad \text{for any } s\text{-smooth function } g, \tag{4.4}$$

where $\Pi_b g = \arg\min_{\ell = \theta^{\mathrm{T}} b} \int (g - \ell)^2 \, dF^*$ is the usual linear projection of g on b, for $dF^*(v) = dF(x_0 + h(v - 1/2))$ the distribution in $\mathcal{B}_h(x_0)$, the h-ball around x_0, mapped to $[0, 1]^d$. In a slight abuse to ease notation, we omit the dependence of $\Pi_b g$ on F^*. The approximating condition (4.4) holds for numerous bases, including spline, CDV wavelet and local polynomial partition series (and polynomial and Fourier series, up to log factors); it is used often in the literature. We refer to Belloni et al. [2] for more discussion and specific examples (see their Condition A.3 and subsequent discussion in, e.g., their Section 3.2).

PROPOSITION 5. *Assume*:

1. $\lambda_{\max}(\Omega)$ *is bounded above,*
2. *the basis* b *satisfies approximating condition* (4.4),
3. $\widehat{\pi}(x) - \pi(x)$ *is* α*-smooth,*
4. $\widehat{\mu}_0(x) - \mu_0(x)$ *is* β*-smooth.*

Then

$$\begin{aligned}|\mathbb{E}(\widehat{S}_j - S_j \mid D^n)| \lesssim{}& (h/k^{1/d})^{2s} + h^{\gamma}(h/k^{1/d})^{\alpha} \\ &+ (h^{\gamma} + \|\widehat{\mu}_0 - \mu_0\|_{F^*})(\|\widehat{\pi} - \pi\|_{F^*}\|\widehat{\Omega}^{-1} - \Omega^{-1}\|).\end{aligned}$$

Before delving into the proof, we give some brief discussion. The bias consists of three terms; the first two are the main bias terms that would result even if the covariate distribution F were known, and the third is essentially the contribution from having to estimate F. (In Lemma S2 of the Appendix, we show how the operator norm error of $\widehat{\Omega}$ is bounded above by estimation error of the distribution F itself.) We note that the second of the main bias terms $h^{\gamma}(h/k^{1/d})^{\alpha}$ will be of smaller order in all regimes we consider. Compared to the main bias term in a usual higher-order influence function analysis, which is $k^{-2s/d}$ (e.g., for the average treatment effect), our bias term is smaller; this is a result of using the localized basis $b_{hk}(x)$ defined in (4.1), which only has to be utilized locally near x_0 (this smaller bias will be partially offset by a larger variance, as discussed in the next subsection). As mentioned in Remark 11,

the contribution from having to estimate F is only a third-order term, since the estimation error of $\widehat{\Omega}$ (in terms of operator norm) is multiplied by a product of propensity score errors (in $L_2(F^*)$ norm) with the sum of regression errors and smoothing bias h^γ, which is typically of smaller order. In Proposition 6, given after the following proof of Proposition 5, we show how the bias simplifies when F is estimated accurately enough.

PROOF. By iterated expectation, the conditional mean of the first-order term in $\widehat{R}$, that is, $\mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y1}(Z) \mid D^n\}$ is equal to

$$\begin{aligned} &R + \int \rho_h(x)K_h(x)\{\pi(x) - \widehat{\pi}(x)\}\{\pi(x)\tau(x) + \mu_0(x) - \widehat{\mu}_0(x)\}\,dF(x) \\ &\quad = R + \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}\{\pi^*(v)\tau^*(v) + \mu_0^*(v) - \widehat{\mu}_0^*(v)\}\,dF^*(v), \end{aligned}$$

where we use the change of variable $v = \frac{1}{2} + \frac{x - x_0}{h}$ and again define for any function $g : \mathbb{R}^d \mapsto \mathbb{R}$ its corresponding stretched version as $g^*(v) = g(x_0 + h(v - 1/2))$. To ease notation, it is left implicit that any integral over v is only over $\{v : \|v - 1/2\| \le 1/2\}$. Similarly, for $\widehat{Q}$ we have that $\mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{a1}(Z)\rho_h(X_1)^{\mathrm{T}} \mid D^n\}$ equals

$$\begin{aligned} &Q + \int \rho_h(x)K_h(x)\{\pi(x) - \widehat{\pi}(x)\}\pi(x)\rho_h(x)^{\mathrm{T}}\,dF(x) \\ &\quad = Q + \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}\pi^*(v)\rho(v)\,dF^*(v) \end{aligned}$$

so that for the first-order term in $\widehat{S}$ (denoted $\widehat{R}_1 - \widehat{Q}_1\theta$ in discussion of the moment condition (4.3)) we have

$$\begin{aligned} &\mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y1}(Z) \mid D^n\} - \mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{a1}(Z)\rho_h(X_1) \mid D^n\}\theta \\ &\quad = R - Q\theta + \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}\{\mu_0^*(v) - \widehat{\mu}_0^*(v)\}\,dF^*(v) \\ &\qquad + \int \rho(v)\pi^*(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}\{\tau^*(v) - \tau_h^*(v)\}\,dF^*(v). \end{aligned} \tag{4.5}$$

The conditional mean of the second-order influence function term in $\widehat{R}$ is

$$\begin{aligned} &\mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y2}(Z_1, Z_2)K_h(X_2) \mid D^n\} \\ &\quad = -\int \rho(v_1)\{\pi^*(v_1) - \widehat{\pi}^*(v_1)\}\widehat{\Pi}_b(\pi^*\tau^* + \mu_0^* - \widehat{\mu}_0^*)(v_1)\,dF^*(v_1), \end{aligned} \tag{4.6}$$

where we define

$$\Pi_b g^*(u) = b(u)^{\mathrm{T}}\Omega^{-1}\int b(v)g^*(v)\,dF^*(v)$$

as the F^*-weighted linear projection of g^* on the basis b, and $\widehat{\Pi}_b g^*(u)$ as the estimated version, which simply replaces Ω with $\widehat{\Omega}$. Similarly, for $\widehat{Q}\theta$ we have

$$\begin{aligned} &\mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{a2}(Z_1, Z_2)K_h(X_2)\rho_h(X_2)^{\mathrm{T}}\theta \mid D^n\} \\ &\quad = -\int \rho(v_1)\{\pi^*(v_1) - \widehat{\pi}^*(v_1)\}\widehat{\Pi}_b(\pi^*\tau_h^*)(v_1)\,dF^*(v_1) \end{aligned} \tag{4.7}$$

so that the conditional mean (4.6) minus the conditional mean (4.7) equals

$$-\int \rho(v_1)\{\pi^*(v_1) - \widehat{\pi}^*(v_1)\}\widehat{\Pi}_b\{\pi^*(\tau^* - \tau_h^*) + (\mu_0^* - \widehat{\mu}_0^*)\}(v_1)\,dF^*(v_1). \tag{4.8}$$

Therefore, adding the first- and second-order expected values in (4.5) and (4.8), the overall bias relative to S is

$$\int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}(I - \widehat{\Pi}_b)\{\pi^*(\tau^* - \tau_h^*) + (\mu_0^* - \widehat{\mu}_0^*)\}(v)\, dF^*(v)$$

$$= \int (I - \Pi_b)\{\rho(\pi^* - \widehat{\pi}^*)\}(v)(I - \Pi_b)\{\pi^*(\tau^* - \tau_h^*) + (\mu_0^* - \widehat{\mu}_0^*)\}(v)\, dF^*(v) \tag{4.9}$$

$$+ \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}(\Pi_b - \widehat{\Pi}_b)\{\pi^*(\tau^* - \tau_h^*) + (\mu_0^* - \widehat{\mu}_0^*)\}(v)\, dF^*(v) \tag{4.10}$$

by the orthogonality of a projection with its residuals (Lemma S1(i)).

Now we analyze the bias terms (4.9) and (4.10) separately; the first is the main bias term, which would arise even if the covariate density were known, and the second is the contribution coming from having to estimate the covariate density.

Crucially, by virtue of using the localized basis b_{hk}, the projections in these bias terms are of stretched versions of the nuisance functions $(\pi^* - \widehat{\pi}^*)$ and $(\mu_0^* - \widehat{\mu}_0^*)$, on the standard *nonlocalized* basis b, with weights equal to the stretched density dF^*. This is important because stretching a function increases its smoothness; in particular, the stretched and scaled function $g^*(v)/h^\alpha$ is α-smooth whenever g is α-smooth. This follows since $|D^{\lfloor\alpha\rfloor} g^*(v) - D^{\lfloor\alpha\rfloor} g^*(v')|$ equals

$$\begin{aligned}
&|D^{\lfloor\alpha\rfloor} g(x_0 + h(v - 1/2)) - D^{\lfloor\alpha\rfloor} g(x_0 + h(v' - 1/2))| \\
&= h^{\lfloor\alpha\rfloor} |g^{(\lfloor\alpha\rfloor)}(x_0 + h(v - 1/2)) - g^{(\lfloor\alpha\rfloor)}(x_0 + h(v' - 1/2))| \\
&\lesssim h^\alpha |v - v'|,
\end{aligned}$$

where the second equality follows by the chain rule, and the third since g is α-smooth. Thus the above implies $h^{-\alpha}|D^{\lfloor\alpha\rfloor} g^*(v) - D^{\lfloor\alpha\rfloor} g^*(v')| \lesssim |v - v'|$, that is, that $g^*(v)/h^\alpha$ is α-smooth.

Therefore, if g is α-smooth, then $\|(I - \Pi_b)g^*/h^\alpha\|_{F^*} \lesssim k^{-\alpha/d}$ by the Hölder approximation properties (4.4) of the basis b, and so it follows that

$$\|(I - \Pi_b)g^*\|_{F^*} \lesssim h^\alpha k^{-\alpha/d} = (h/k^{1/d})^\alpha \tag{4.11}$$

for any α-smooth function g.

Therefore, now consider the bias term (4.9). This term satisfies

$$\begin{aligned}
&\int [(I - \Pi_b)\{\rho(\pi^* - \widehat{\pi}^*)\}(v)][(I - \Pi_b)\{\pi^*(\tau^* - \tau_h^*) + (\mu_0^* - \widehat{\mu}_0^*)\}(v)]\, dF^*(v) \\
&\le \|(I - \Pi_b)\{\rho(\pi^* - \widehat{\pi}^*)\}\|_{F^*} [\|(I - \Pi_b)\{\pi^*(\tau^* - \tau_h^*)\}\|_{F^*} \\
&\quad + \|(I - \Pi_b)(\mu_0^* - \widehat{\mu}_0^*)\|_{F^*}] \\
&\lesssim (h/k^{1/d})^\alpha \{h^\gamma + (h/k^{1/d})^\beta\} = (h/k^{1/d})^{2s} + h^\gamma (h/k^{1/d})^\alpha,
\end{aligned}$$

where the second line follows by Cauchy–Schwarz, and the third by (4.11), since $(\pi - \widehat{\pi})$ and $(\mu_0 - \widehat{\mu}_0)$ are assumed α- and β-smooth, respectively (note $\rho(v)$ is a polynomial, so the smoothness of $\rho(\pi^* - \widehat{\pi}^*)$ is the same as $(\pi^* - \widehat{\pi}^*)$), along with the fact that

$$\begin{aligned}
\|(I - \Pi_b)\pi^*(\tau^* - \tau_h^*)\|_{F^*}^2 &\le \|\pi^*(\tau^* - \tau_h^*)\|_{F^*}^2 \\
&\le \int K_h(x)\{\tau(x) - \tau_h(x)\}^2\, dF(x) \lesssim h^{2\gamma},
\end{aligned}$$

where the first inequality follows by Lemma S1(ii), the second by definition of F^* and since $\pi(x) \le 1$, and the last by Proposition 3.

Now for the term in (4.10), let $\theta_{b,g} = \Omega^{-1} \int bg\, dF^*$ denote the coefficients of the projection $\Pi_b g$, and note for any functions g_1, g_2 we have

$$\begin{aligned}\int g_1(\Pi_b - \widehat{\Pi}_b)(g_2)\, dF^* &= (\Omega^{1/2}\theta_{b,g_1})^{\mathrm{T}}\Omega^{1/2}(\Omega^{-1} - \widehat{\Omega}^{-1})\Omega^{1/2}(\Omega^{1/2}\theta_{b,g_2}) \\ &\le \|g_1\|_{F^*}\|\Omega^{1/2}(\Omega^{-1} - \widehat{\Omega}^{-1})\Omega^{1/2}\|\|g_2\|_{F^*} \\ &\le \|g_1\|_{F^*}\|g_2\|_{F^*}\|\Omega\|\|\widehat{\Omega}^{-1} - \Omega^{-1}\|,\end{aligned}$$

where the first equality follows by definition, the second line since the L_2 norm of the coefficients of a (weighted) projection is no more than the weighted $L_2(\mathbb{P})$ norm of the function itself (Lemma S1(iii)), and the last by the submultiplicative property of the operator norm, along with the fact that $\|\Omega^{1/2}\|^2 = \|\Omega\|$. □

Several of our results require the eigenvalues of Ω to be bounded above and below away from zero. Proposition S6 in the Appendix gives simple sufficient conditions for this to hold (similar to Proposition S5 for the matrix Q, which was mentioned earlier after Proposition 3).

The next result is a refined version of Proposition 5, giving high-level conditions under which estimation of F itself (rather than the matrix Ω^{-1}) does not impact the bias. We refer to Remark 11 for more detailed discussion of these conditions, and note that the result follows from Proposition 5 together with Lemma S2 in the Appendix.

PROPOSITION 6. *Under the assumptions of Proposition 6, and if*:

1. $\lambda_{\min}(\Omega)$ *is bounded below away from zero*,
2. $\|d\widehat{F}^*/dF^*\|_\infty$ *is bounded above and below away from zero*,
3. $\|(d\widehat{F}^*/dF^*) - 1\|_\infty \lesssim \frac{(h/k^{1/d})^{2s}}{\|\widehat{\pi}-\pi\|_{F^*}(\|\widehat{\mu}_0-\mu_0\|_{F^*}+h^\gamma)}$,

then, when $h^\gamma \lesssim (h/k^{1/d})^\beta$, *the bias satisfies* $|\mathbb{E}(\widehat{S}_j - S_j \mid D^n)| \lesssim (h/k^{1/d})^{2s}$.

4.3. *Variance*. In this subsection, we derive bounds on the conditional variance of the estimators $\widehat{R}_j$ and $\widehat{Q}_{j\ell}$, given the training sample D^n. The main tool used here is a localized version of second-order U-statistic variance arguments, recognizing that our higher-order estimator is, conditionally, a second-order U-statistic over nh^d observations.

PROPOSITION 7. *Assume*:

1. y^2, $\widehat{\pi}^2$, $\widehat{\mu}_0^2$, *and* $\|\widehat{\mu}_0 - \mu_0\|_{F^*}$ *are all bounded above, and*
2. $\lambda_{\max}(\Omega)$ *is bounded above.*

Then

$$\operatorname{var}(\widehat{S}_j \mid D^n) \lesssim \frac{1}{nh^d}\left(1 + \frac{k}{nh^d}(1 + \|\widehat{\Omega}^{-1} - \Omega^{-1}\|^2)\right).$$

We give the proof of Proposition 7 in Appendix B6, and so just make some comments here. First, the variance here is analogous to that of a higher-order (quadratic) influence function estimator (cf. Theorem 1 of Robins et al. [28]), except with sample size n deflated to nh^d. This is to be expected given the double localization in our proposed estimator. Another important note is that the contribution to the variance from having to estimate F is relatively minimal, compared to the bias, as detailed in Proposition 5. For the bias, nontrivial rate conditions are needed to ensure estimation of F does not play a role, whereas for the variance one only needs the operator norm of $\widehat{\Omega}^{-1} - \Omega^{-1}$ to be bounded (under regularity conditions, this amounts to the estimator $\widehat{F}$ only having bounded errors, in a relative sense, as noted in the following remark).

REMARK 9. By Lemma S2, under the assumptions of Proposition 6, it follows that:

$$\|\widehat{\Omega}^{-1} - \Omega^{-1}\| \lesssim \|(d\widehat{F}^*/dF^*) - 1\|_\infty,$$

so estimation of F will not affect the conditional variances as long as the error of $\widehat{F}$ is bounded in uniform norm.

4.4. *Overall rate.* Combining the approximation bias in Proposition 3 with the decomposition in Proposition 4, and the bias and variance bounds from Proposition 6 and Proposition 7, respectively, shows that

$$\mathbb{E}_P|\widehat{\tau}(x_0) - \tau_P(x_0)| \lesssim h^\gamma + (h/k^{1/d})^{2s} + \sqrt{\frac{1}{nh^d}\left(1 + \frac{k}{nh^d}\right)} \tag{4.12}$$

under all the combined assumptions of these results, which are compiled in the statement of Theorem 2 below. The first two terms in (4.12) are the bias, with h^γ an oracle bias that would remain even if one had direct access to the potential outcomes $(Y^1 - Y^0)$ (or equivalently, samples of $\tau(X) + \epsilon$ for some ϵ with conditional mean zero), and $(h/k^{1/d})^{2s}$ analogous to a squared nuisance bias term, but shrunken due to the stretching induced by the localized basis b_{hk}. Similarly, $1/(nh^d)$ is an oracle variance that would remain even if given access to the potential outcomes, whereas the $k/(nh^d)$ factor is a contribution from nuisance estimation (akin to the variance of a series regression on k basis terms with nh^d samples).

Balancing bias and variance in (4.12) by taking the tuning parameters to satisfy

$$h \sim n^{-(1/\gamma)/(1+\frac{d}{2\gamma}+\frac{d}{4s})} \quad \text{and} \quad k \sim n^{(\frac{d}{2s}-\frac{d}{\gamma})/(1+\frac{d}{2\gamma}+\frac{d}{4s})}$$

ensures the rate matches the minimax lower bound from Theorem 1 (in the low smoothness regime), proving that lower bound is in fact tight. (In the high smoothness regime where $s > (d/4)/(1 + d/2\gamma)$, one can simply take $k \sim nh^d$ and $h \sim n^{-1/(2\gamma+d)}$). This is formalized in the following theorem.

THEOREM 2. *Assume the regularity conditions*:

A. *The eigenvalues of Q and Ω are bounded above and below away from zero.*
B. *$\widehat{\pi}(x) - \pi(x)$ is α-smooth and $\widehat{\mu}_0(x) - \mu_0(x)$ is β-smooth.*
C. *The quantities y^2, $(\widehat{\pi}^2, \widehat{\mu}_0^2)$, $\|\widehat{\mu}_0 - \mu_0\|_{F^*}$ and $\|\widehat{Q}^{-1} - Q^{-1}\|$ are all bounded above, and $\|d\widehat{F}^*/dF^*\|_\infty$ is bounded above and below away from zero.*

Also assume the basis b satisfies Hölder approximating condition (4.4), *and*:

1. $\|(d\widehat{F}^*/dF^*) - 1\|_\infty \lesssim \frac{n^{-1/(1+\frac{d}{2\gamma}+\frac{d}{4s}\vee(1+\frac{d}{2\gamma}))}}{\|\widehat{\pi}-\pi\|_{F^*}(\|\widehat{\mu}_0-\mu_0\|_{F^*}+h^\gamma)}$,
2. *$\pi(x)$ is α-smooth, and $\epsilon \le \pi(x) \le 1 - \epsilon$ for some $\epsilon > 0$,*
3. *$\mu_0(x)$ is β-smooth,*
4. *$\tau(x)$ is γ-smooth.*

Finally let the tuning parameters satisfy

$$h \sim n^{-(1/\gamma)/(1+\frac{d}{2\gamma}+\frac{d}{4s})} \quad \text{and} \quad k \sim n^{(\frac{d}{2s}-\frac{d}{\gamma})/(1+\frac{d}{2\gamma}+\frac{d}{4s})}$$

if $s < \frac{d/4}{1+d/2\gamma}$, or $h \sim n^{-\frac{1}{2\gamma+d}}$ and $k \sim nh^d$ otherwise. Then the estimator $\widehat{\tau}$ from Definition 2 has error upper bounded as

$$\mathbb{E}_P|\widehat{\tau}(x_0) - \tau_P(x_0)| \lesssim \begin{cases} n^{-1/(1+\frac{d}{2\gamma}+\frac{d}{4s})} & \text{if } s < \dfrac{d/4}{1+d/2\gamma}, \\ n^{-1/(2+\frac{d}{\gamma})} & \text{otherwise.} \end{cases}$$

We refer to Section 3.3 for more detailed discussion and visualization of the rate from Theorem 2. Here, we give two remarks discussing the regularity conditions A–C and Condition 1 (which ensures the covariate distribution is estimated accurately enough).

REMARK 10. Condition 2 is a standard collinearity restriction used with least squares estimators; simple sufficient conditions are given in Propositions S5 and S6 in the Appendix. In Lemma S3 in the Appendix, we also prove that this condition holds for a class of densities contained in the model $\mathcal{P}$ in Theorem 1, ensuring that the upper bound holds over the same submodel. A sufficient condition for Condition 2 to hold is that the estimators $\widehat{\pi}(x)$ and $\widehat{\mu}_0(x)$ match the (known) smoothness of $\pi(x)$ and $\mu_0(x)$; this would be the case for standard minimax optimal estimators based on series or local polynomial methods. Condition 2 is a mild boundedness condition on the outcome Y (which could be weakened at the expense of adding some complexity in the analysis), as well as the nuisance estimators $(\widehat{F}^*, \widehat{\pi}, \widehat{\mu}_0)$, and even weaker, the errors $\|\widehat{\mu}_0 - \mu_0\|_{F^*}$ and $\|\widehat{Q}^{-1} - Q^{-1}\|$ (which would typically not only be bounded but decreasing to zero).

REMARK 11. First, Condition 1 of Theorem 2 will of course hold if the covariate distribution is estimated at a rate faster than that of the CATE (i.e., the numerator of the rate in Condition 1); however, it also holds under substantially weaker conditions, depending on how accurately π and μ_0 are estimated. This is because the condition really amounts to a third-order term (the covariate distribution error multiplied by a *product* of nuisance errors) being of smaller order than the CATE rate. Specifically, the result of Theorem 2 can also be written as

$$\mathbb{E}_P\big|\widehat{\tau}(x_0) - \tau_P(x_0)\big| \lesssim n^{-1/(1+\frac{d}{2\gamma}+\frac{d}{4s}\vee(1+\frac{d}{2\gamma}))} + R_{3,n}, \tag{4.13}$$

for the third-order error term

$$R_{3,n} = \big\|(d\widehat{F}^*/dF^*) - 1\big\|_\infty \|\widehat{\pi} - \pi\|_{F^*}\big(\|\widehat{\mu}_0 - \mu_0\|_{F^*} + h^\gamma\big),$$

so that Condition 1 simply requires this third-order term to be of smaller order than the first minimax optimal rate in (4.13) (note in the above that h^γ matches the overall CATE estimation error, under our tuning parameter choices, which would typically be of smaller order than the regression error $\|\widehat{\mu}_0 - \mu_0\|_{F^*}$). Second, we note that we leave the condition in terms of the $L_2(F^*)$ errors $\|\widehat{\pi} - \pi\|_{F^*}$ and $\|\widehat{\mu}_0 - \mu_0\|_{F^*}$ because, although we assume π and μ_0 are α- and β-smooth, technically, they do not need to be estimated at particular rates for any of the other results we prove to hold. Of course, under these smoothness assumptions, there are available minimax optimal estimators for which

$$\|\widehat{\pi} - \pi\|_{F^*} \asymp n^{-1/(2+d/\alpha)} \quad \text{and} \quad \|\widehat{\mu}_0 - \mu_0\|_{F^*} \asymp n^{-1/(2+d/\beta)}.$$

If in addition there exists some ζ for which $\|(d\widehat{F}^*/dF^*) - 1\|_\infty \asymp n^{-1/(2+d/\zeta)}$ (e.g., if F has a density that is ζ-smooth), then Condition 1 reduces to $\zeta > d/(1/M_{\alpha,\beta,\gamma,d} - 2)$, for

$$M_{\alpha,\beta,\gamma,d} \equiv \frac{1}{1 + d/2\gamma + d/4s \vee (1 + d/2\gamma)} - \frac{1}{2 + d/\alpha} - \frac{1}{2 + d/\beta}.$$

Exploring CATE estimation under weaker conditions on the covariate distribution is an interesting avenue for future work; we suspect the minimax rate changes depending on what is assumed about this distribution, as is the case for average effects (e.g., p. 338 of Robins et al. [27]) and conditional variance estimation [36, 42].

5. Discussion. In this paper, we have characterized the minimax rate for estimating heterogeneous causal effects in a smooth nonparametric model. We derived a lower bound on the minimax rate using a localized version of the method of fuzzy hypotheses, and a matching upper bound via a new local polynomial R-Learner estimator based on higher-order influence functions. We also characterize how the minimax rate changes depending on whether the propensity score or regression function is smoother, either when one parametrizes the control or the marginal regression function. The minimax rate has several important features. First, it exhibits a so-called elbow phenomenon: when the nuisance functions (regression and propensity scores) are smooth enough, the rate matches that of standard smooth nonparametric regression (the same that would be obtained if potential outcomes were actually observed). On the other hand, when the average nuisance smoothness is below the relevant threshold, the rate obtained is slower. This leads to a second important feature: in the latter low-smoothness regime, the minimax rate is a mixture of the minimax rates for nonparametric regression and functional estimation. This quantifies how the CATE can be viewed as a regression/functional hybrid.

There are numerous important avenues left for future work. We detail a few briefly here, based on: different error metrics, inference and testing, adaptivity and practical implementation. First, we have focused on estimation error at a point, but one could also consider global rates in L_2 or L_∞ norm, for example. We expect L_2 rates to be the same, and L_∞ rates to be the same up to log factors, but verifying this would be useful. In addition, it would also be very important to study the distribution of the proposed estimator, beyond just bias and variance, for example, for purposes of inference (L_∞ rates could also be useful in this respect). Relatedly, one could consider minimax rates for testing whether the CATE is zero, for example, versus ϵ-separated in some distance. The goal of the present work is mostly to further our theoretical understanding of the fundamental limits of CATE estimation, so there remains lots to do to make the optimal rates obtained here achievable in practice. For example, although we have specified particular values of the tuning parameters h and k to confirm attainability of our minimax lower bound, it would be practically useful to have more data-driven approaches for selection. In particular, the optimal tuning values depend on underlying smoothness, and since in practice this is often unknown, a natural next step is to study adaptivity. For example, one could study whether approaches based on Lepski's method could be used, as in Mukherjee et al. [24] and Liu et al. [21]. There are also potential computational challenges associated with constructing the tensor products in $\rho(x)$ when dimension d is not small, as well as evaluating the U-statistic terms of our estimator, and inverting the matrices $\widehat{Q}$ and $\widehat{\Omega}$. Finally, in this work we have assumed the nuisance functions are Hölder-smooth, a classic infinite-dimensional function class from which important insights can be drawn. However, it will be important to explore minimax rates in other function classes as well.

Funding. EK gratefully acknowledges support from NSF Grant DMS-1810979, NSF CAREER Award 2047444 and NIH R01 Grant LM013361-01A1, and SB and LW from NSF Grant DMS-1713003. EK also thanks Matteo Bonvini and Tiger Zeng for very helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "Minimax rates for heterogeneous causal effect estimation" (DOI: 10.1214/24-AOS2369SUPP; .pdf). Supplementary information.

REFERENCES

[1] ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

[2] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

[3] BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

[4] BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653 https://doi.org/10.1214/aos/1176324452

[5] CHAMBERLAIN, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34** 305–334. MR0888070 https://doi.org/10.1016/0304-4076(87)90015-7

[6] CHERNOZHUKOV, V., GOLDMAN, M., SEMENOVA, V. and TADDY, M. (2017). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. Preprint. Available at arXiv:1712.09988.

[7] FAN, Q., HSU, Y.-C., LIELI, R. P. and ZHANG, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *J. Bus. Econom. Statist.* **40** 313–327. MR4356575 https://doi.org/10.1080/07350015.2020.1811102

[8] FOSTER, D. J. and SYRGKANIS, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908. MR4630373 https://doi.org/10.1214/23-AOS2258

[9] FOSTER, J. C., TAYLOR, J. M. G. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880. MR2844689 https://doi.org/10.1002/sim.4322

[10] GAO, Z. and HAN, Y. (2020). Minimax optimal nonparametric estimation of heterogeneous treatment effects. Preprint. Available at arXiv:2002.06471.

[11] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics*, [40]. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[12] HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. MR4154846 https://doi.org/10.1214/19-BA1195

[13] HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. CRC, Boca Raton, FL.

[14] IBRAGIMOV, I. A., NEMIROVSKIĬ, A. S. and KHAS'MINSKIĬ, R. Z. (1986). Some problems of nonparametric estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **31** 451–466. MR0866866

[15] IMAI, K. and RATKOVIC, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7** 443–470. MR3086426 https://doi.org/10.1214/12-AOAS593

[16] INGSTER, YU. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446 https://doi.org/10.1007/978-0-387-21580-8

[17] KENNEDY, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electron. J. Stat.* **17** 3008–3049. MR4667730 https://doi.org/10.1214/23-ejs2157

[18] KENNEDY, E. H, BALAKRISHNAN, S., ROBINS, J. M and WASSERMAN, L. (2024). Supplement to "Minimax rates for heterogeneous causal effect estimation." https://doi.org/10.1214/24-AOS2369SUPP

[19] KUENZEL, S. R. (2019). *Heterogeneous Treatment Effect Estimation Using Machine Learning*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of California, Berkeley. MR4051203

[20] LEE, S., OKUI, R. and WHANG, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *J. Appl. Econometrics* **32** 1207–1225. MR3734484 https://doi.org/10.1002/jae.2574

[21] LIU, L., MUKHERJEE, R., ROBINS, J. M. and TCHETGEN TCHETGEN, E. (2021). Adaptive estimation of nonparametric functionals. *J. Mach. Learn. Res.* **22** Paper No. 99, 66. MR4279750

[22] LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Super-learning of an optimal dynamic treatment rule. *Int. J. Biostat.* **12** 305–332. MR3505699 https://doi.org/10.1515/ijb-2015-0052

[23] MUKHERJEE, R., NEWEY, W. K. and ROBINS, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. Preprint. Available at arXiv:1705.07577.

[24] MUKHERJEE, R., TCHETGEN, E. J. T. and ROBINS, J. M. (2015). Lepski's method and adaptive estimation of nonlinear integral functionals of density. Preprint. Available at arXiv:1508.00249.

[25] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640

[26] NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

[27] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A. Freedman. Inst. Math. Stat.* (*IMS*) *Collect.* **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527

[28] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. W. (2009). Quadratic semiparametric von Mises calculus. *Metrika* **69** 227–247. MR2481922 https://doi.org/10.1007/s00184-008-0214-3

[29] ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

[30] ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185 https://doi.org/10.1080/03610929408831393

[31] ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

[32] ROBINS, J. M., MARK, S. D. and NEWEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48** 479–495. MR1173493 https://doi.org/10.2307/2532304

[33] ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762 https://doi.org/10.2307/1912705

[34] SEMENOVA, V. and CHERNOZHUKOV, V. (2017). Estimation and inference about conditional average treatment effect and other structural functions. arXiv–1702, arXiv.

[35] SHALIT, U., JOHANSSON, F. D. and SONTAG, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the* 34*th International Conference on Machine Learning—Volume* 70 3076–3085. JMLR.org.

[36] SHEN, Y., GAO, C., WITTEN, D. and HAN, F. (2020). Optimal estimation of variance in nonparametric regression with random design. *Ann. Statist.* **48** 3589–3618. MR4185821 https://doi.org/10.1214/20-AOS1944

[37] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

[38] VAN DER LAAN, M. J. (2006). Statistical inference for variable importance. *Int. J. Biostat.* **2** Art. 2, 33. MR2275897 https://doi.org/10.2202/1557-4679.1008

[39] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. *Springer Series in Statistics*. Springer, New York. MR1958123 https://doi.org/10.1007/978-0-387-21700-0

[40] VANSTEELANDT, S. and JOFFE, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statist. Sci.* **29** 707–731. MR3300367 https://doi.org/10.1214/14-STS493

[41] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

[42] WANG, L., BROWN, L. D., CAI, T. T. and LEVINE, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann. Statist.* **36** 646–664. MR2396810 https://doi.org/10.1214/009053607000000901

[43] ZHAO, Q., SMALL, D. S. and ERTEFAIE, A. (2022). Selective inference for effect modification via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 382–413. MR4412991 https://doi.org/10.1111/rssb.12483

[44] ZIMMERT, M. and LECHNER, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. Preprint. Available at arXiv:1908.08779.