Minimax rates for heterogeneous causal effect estimation

Edward H. Kennedy, Sivaraman Balakrishnan, Larry Wasserman

Department of Statistics & Data Science,
Carnegie Mellon University

{edward, siva, larry} @ stat.cmu.edu

Abstract

Estimation of heterogeneous causal effects – i.e., how effects of policies and treatments vary across subjects – is a fundamental task in causal inference, playing a crucial role in optimal treatment allocation, generalizability, subgroup effects, and more. Many flexible 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 nonparametric model where distributional components are Hölder-smooth, and present a new local polynomial estimator, giving high-level conditions under which it is minimax optimal. More specifically, 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; it is shown to be minimax optimal under a condition on how accurately the covariate distribution is estimated. The minimax rate we find exhibits several interesting features, including a non-standard 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. We conclude with some discussion of a few remaining open problems.

Keywords: causal inference, functional estimation, higher order influence functions, nonparametric regression, optimal rates of convergence.

1 Introduction

In this paper we consider estimating the difference in regression functions

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

from an iid 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)

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

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.

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 [Robins, 1994, Robins et al., 1992, Robinson, 1988, van der Laan, 2006, van der Laan and Robins, 2003, Vansteelandt and Joffe, 2014]. 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., non-smooth or non-sparse), 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 [2020] 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 [2006] (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, e.g., cross-validated ensembles [Luedtke and van der Laan, 2016], kernel [Fan et al., 2019, Lee et al., 2017, Zimmert and Lechner, 2019] and series methods [Semenova and Chernozhukov, 2017], empirical risk minimization [Foster and Syrgkanis, 2019], and linear smoothers [Kennedy, 2020]. On the other hand, the R-Learner is a flexible adaptation of the double-residual regression method originally built for partially linear models [Robinson, 1988], with the first nonparametric version

proposed by Robins et al. [2008] (Section 5.2) using series methods. The R-Learner has since been adapted to RKHS regression [Nie and Wager, 2021], lasso [Chernozhukov et al., 2017, Zhao et al., 2017], and local polynomials [Kennedy, 2020]. Many flexible non-doubly robust methods have also been proposed in recent years, often based on inverse-weighting or direct regression estimation [Athey and Imbens, 2016, Foster et al., 2011, Hahn et al., 2020, Imai and Ratkovic, 2013, Künzel et al., 2019, Shalit et al., 2017, Wager and Athey, 2018].

Despite the wide variety of methods available for flexible CATE estimation, questions of optimality have remained mostly unsolved. Gao and Han [2020] studied minimax optimality, but in a specialized model where the propensity score has zero smoothness, and covariates are non-random; this model does not reflect the kinds of assumptions typically used in practice, e.g., 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 [2019] and Kennedy [2020]. Foster and Syrgkanis [2019] studied global error rates, obtaining an oracle error plus sums of squared L_4 errors in all nuisance components. Kennedy [2020] 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, with the propensity score at least as smooth as the regression function. 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 [Birgé and Massart, 1995, Ibragimov et al., 1987, Ingster et al., 2003, Nemirovski, 2000, Robins et al., 2009b, Tsybakov, 2009]. 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 [Robins et al., 2008, 2009a, 2017]. 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 & Notation

We consider an iid 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 $F(x)$ denote the distribution function of the covariate X (with density $f(x)$ as needed), and let

$$\pi(x) = \mathbb{P}(A = 1 \mid X = x)$$
$$\mu(x) = \mathbb{E}(Y \mid X = x)$$

denote the propensity score, and marginal outcome regression functions, respectively. We sometimes omit arguments from functions to ease notation, e.g., note that $\tau = (\mu - \mu_0)/\pi$ for $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$. We also index functions by a distribution P when needed, e.g., $\tau(x)$ under a particular distribution P is written $\tau_P(x)$; depending on context, no indexing means the function is evaluated at the true $\mathbb{P}$, e.g., $\tau(x) = \tau_{\mathbb{P}}(x)$.

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

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

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, which we denote $\mathcal{H}(s)$; 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 $\mathcal{H}(s)$ 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, i.e.,

$$|D^\alpha g(x)| \leq C$$

for all $\alpha = (\alpha_1, ..., \alpha_d)$ with $\sum_j \alpha_j \leq \lfloor s \rfloor$ and for all $x \in \mathcal{X}$, and with $\lfloor s \rfloor$-order derivatives Hölder continuous, i.e.,

$$\left|D^\beta g(x) - D^\beta g(x')\right| \leq C\|x - x'\|^{s - \lfloor s \rfloor}$$

for all $\beta = (\beta_1, ..., \beta_d)$ 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, as in $\mathcal{H}(s, C)$, but we focus our discussion on the smoothness s and omit the constant.

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 $a_n \leq C b_n$ for C a positive constant independent of n, and $a_n \asymp b_n$ if $a_n \leq C b_n$ and $b_n \leq C a_n$ (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, i.e., $a_n = C b_n$ 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 $x_0 \in (0,1)$, let $\mathcal{P}$ denote the model where:*

1. *$dF(x)$ is known, satisfies $\int \mathbb{1}\{\|x - x_0\| \leq h/2\}\, dF(x) \asymp h^d$, and has local support $\{x \in \mathbb{R}^d : dF(x) > 0, \|x - x_0\| \leq h/2\}$ on a union of no more than k disjoint cubes all with proportional volume, for h and k defined in Proposition 3,*
2. *$\pi(x)$ is α-smooth, and $\epsilon \leq \pi(x) \leq 1 - \epsilon$ for some $\epsilon > 0$,*
3. *$\mu(x)$ is β-smooth, with $\beta \leq \alpha$, 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

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

First we remark on some details about the model we consider. 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. In Condition 3 we also assume the propensity score is at least as smooth as the regression function, i.e., $\alpha \geq \beta$. This can be motivated by practical settings where the treatment process is more simple or structured than the outcome process; for example, treatment may be based on relatively simple human decision-making, whereas the outcome may be some complex physiological response. One can also view this as a nonparametric analog of semiparametric models that employ parametric assumptions on the treatment but not outcome processes [Tsiatis, 2006, van der Laan and Robins, 2003]. Further, when $\alpha < \beta$, we expect our proof techniques would need to change substantially, and so leave this avenue to future work; this is detailed further in the Discussion. Condition 1 of our model does not impose any smoothness on the covariate distribution F, but ensures it is sufficiently dense and that sufficiently many samples are observed near the target point x_0. The condition would be satisfied whenever F has a density bounded above and below away from zero, with support $[0,1]^d$, for example. We also note that, although F is taken to be known in the model, of course the derived lower bound equally

applies to larger models where F is unknown and needs to be estimated. 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.

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 [Birgé and Massart, 1995, Ibragimov et al., 1987, Ingster et al., 2003, Nemirovski, 2000, Robins et al., 2009b, Tsybakov, 2009]. In the classic Le Cam two-point method, which can be used to derive minimax lower bounds for nonparametric regression at a point [Tsybakov, 2009], 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 [Birgé and Massart, 1995, Robins et al., 2009b, Tsybakov, 2009]. Our construction combines these two approaches via a localized mixture, as will be described in detail in the next subsection.

Remark 1. 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 [2009].

The following lemma, adapted from Section 2.7.4 of Tsybakov [2009], provides the foundation for the minimax lower bound result of this section.

Lemma 1 (Tsybakov [2009]). *Let P_λ and Q_λ denote distributions in $\mathcal{P}$ indexed by a vector $\lambda = (\lambda_1, ..., \lambda_k)$, 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, then

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

for any monotonic non-negative 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 perturb the CATE with a bump at the point x_0, and to also perturb the propensity score and regression functions π and μ, but only 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) for $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$, or (f, π, μ, τ); we make use of the latter parametrization. We first give the construction in the definition below, and then go on to discuss the details.

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 $x \in [-1/2, 1/2]^d$, and $B(x) = 0$ for $x \notin [-1, 1]^d$,
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. $(\mathcal{X}_1, ..., \mathcal{X}_k)$ denote a partition of $\mathcal{C}_h(x_0)$ into k cubes of equal size, with midpoints $(m_1, ..., m_k)$, 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

$$\begin{aligned}
\tau_h(x) &= h^\gamma B\left(\frac{x - x_0}{h}\right) \\
\mu_\lambda(x) &= \frac{1}{2} + k^{-\beta/d} \sum_{j=1}^{k} \lambda_j B\left(\frac{x - m_j}{h/k^{1/d}}\right) \\
\pi_\lambda(x) &= \frac{1}{2} + k^{-\alpha/d} \sum_{j=1}^{k} \lambda_j B\left(\frac{x - m_j}{h/k^{1/d}}\right) \\
f(x) &= \mathbb{1}(x \in \mathcal{S}_{hk}) \Big/ \left\{1 - \left(\frac{4^d - 1}{2^d}\right) h^d\right\}
\end{aligned}$$

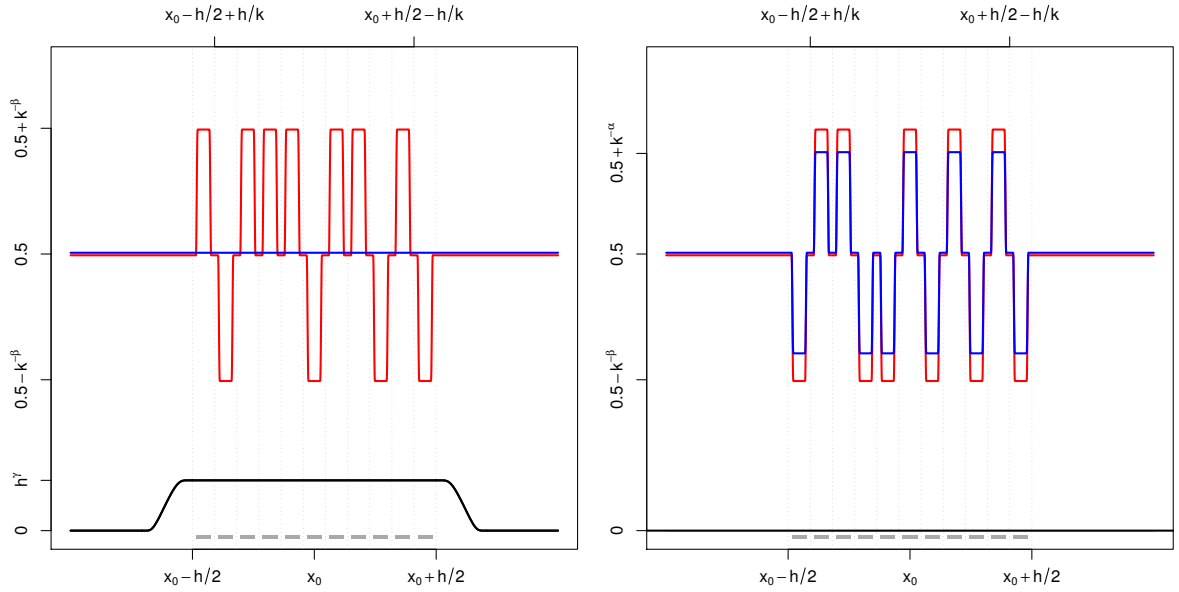
where $\mathcal{S}_{hk} = \left\{\bigcup_{j=1}^{k} \mathcal{C}_{h/2k^{1/d}}(m_j)\right\} \bigcup \left\{[0, 1]^d \setminus \mathcal{C}_{2h}(x_0)\right\}$. Finally take the distributions P_λ and Q_λ to be defined via the densities

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

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 with bumps

locally near x_0. The regression function μ is perturbed under both P_λ and Q_λ, since it is less smooth than the propensity score in our model. The choices of the CATE mimic those in the two-point proof of the lower bound for nonparametric regression at a point (see, e.g., Section 2.5 of Tsybakov [2009]), albeit with a particular flat-top bump function, while the choices of nuisance functions π and μ are more similar to those in the lower bound for the expected conditional covariance (cf. Section 4 of Robins et al. [2009b]). In this sense our construction can be viewed as combining those for nonparametric regression and functional estimation, similar to Shen et al. [2020]. In what follows we remark on some important details.

Remark 2. Section 3.2 of Shen et al. [2020] used a similar construction for deriving the minimax lower bound 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.



(a) Null P_λ (b) Alternative Q_λ

Figure 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, outcome regression, and propensity score functions, respectively, and the gray line denotes the support of the covariate density.

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 and that it is equal to zero for all $x \notin \mathcal{C}_{2h}(x_0)$, i.e.,

outside the cube centered at x_0 with side length $2h$. It is straightforward to check that the CATE function $\tau_h(x)$ is γ-smooth in this construction (see page 93 of Tsybakov [2009]).

Remark 3. One example of a bump function B satisfying the conditions above is

$$B(x) = \begin{cases} 1 & \text{if } |x| \leq 1/2 \\ \frac{\exp\left(\frac{1}{4x^2-1}\right)}{\exp\left(\frac{1}{4x^2-1}\right)+\exp\left(\frac{1}{2-4x^2}\right)} & \text{if } |x| \in (1/2, \sqrt{2}/2) \\ 0 & \text{if } |x| \geq \sqrt{2}/2. \end{cases}$$

For the propensity score and regression functions, we similarly have

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

i.e., 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 $\mu_\lambda(x)$ 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 $\left\{\bigcup_{j=1}^k \mathcal{C}_{h/2k^{1/d}}(m_j)\right\}$, together with the space $\left\{[0,1]^d \setminus \mathcal{C}_{2h}(x_0)\right\}$ away from x_0. Importantly, this choice ensures there is only mass where the nuisance bumps $B\left(\frac{x-m_j}{h/k^{1/d}}\right)$ are constant and non-zero (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 towards a standard uniform distribution. It is also straightforward to check that this density satisfies the required denseness in Condition 1 of Theorem 1.

The following proposition gives an expression for the densities under P_λ and Q_λ, which is important for deriving the relevant distances in the next subsection.

Proposition 1. *The densities under P_λ and Q_λ from Definition 1 are given by*

$$p_\lambda(z) = f(x)\left\{\frac{1}{4} + (y-1/2)k^{-\beta/d}\sum_{j=1}^k \lambda_j B\left(\frac{x-m_j}{h/k^{1/d}}\right) + (2a-1)(2y-1)\frac{h^\gamma}{4}B\left(\frac{x-x_0}{h}\right)\right\}$$

$$q_\lambda(z) = f(x)\left[\frac{1}{4} + \left\{(a-1/2)k^{-\alpha/d} + (y-1/2)k^{-\beta/d}\right\}\sum_{j=1}^k \lambda_j B\left(\frac{x-m_j}{h/k^{1/d}}\right) + (2a-1)(2y-1)k^{-2s/d}\sum_{j=1}^k B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2\right]$$

where $s \equiv (\alpha+\beta)/2$.

We note that the densities are both equal to $1/4$ for all $x \notin \mathcal{C}_{2h}(x_0)$ away from x_0, since $B\left(\frac{x-m_j}{h/k^{1/d}}\right) = 0$ for $x \notin \mathcal{C}_{h/k^{1/d}}(m_j)$ and $\mathcal{C}_{h/k^{1/d}}(m_j) \subseteq \mathcal{C}_h(x_0) \subseteq \mathcal{C}_{2h}(x_0)$, and since $B\left(\frac{x-x_0}{h}\right) = 0$ for $x \notin \mathcal{C}_{2h}(x_0)$.

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 $(\lambda_1, ..., \lambda_k)$ are iid Rademachers.

In general these product densities can be complicated, making direct distance calculations difficult. Fortunately the following lemma from Robins et al. [2009b] can be used to relate the distance between the n-fold products to those of simpler posteriors over a single observation.

Lemma 2 (Robins et al. [2009b]). *Let P_λ and Q_λ denote distributions indexed by a vector $\lambda = (\lambda_1, ..., \lambda_k)$, and let $\mathcal{Z} = \cup_{j=1}^k \mathcal{Z}_j$ 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 $\mathbb{1}_{\mathcal{Z}_j} dQ_\lambda / p_j$ do not depend on λ_ℓ for $\ell \neq j$.*

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

$$\delta_1 = \max_j \sup_\lambda \int_{\mathcal{Z}_j} \frac{(p_\lambda - \overline{p})^2}{p_\lambda p_j} \, d\nu$$
$$\delta_2 = \max_j \sup_\lambda \int_{\mathcal{Z}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda p_j} \, d\nu$$
$$\delta_3 = \max_j \sup_\lambda \int_{\mathcal{Z}_j} \frac{(\overline{q} - \overline{p})^2}{p_\lambda p_j} \, d\nu$$

for a dominating measure ν. If $\overline{p}/p_\lambda \leq b < \infty$ and $\max_j p_j \leq b/n$, then

$$H^2\left(\int P_\lambda^n \, d\varpi(\lambda), \int Q_\lambda^n \, d\varpi(\lambda)\right) \leq Cn^2\left(\max_j p_j\right)\left(\delta_1\delta_2 + \delta_2^2\right) + Cn\delta_3$$

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

$$\delta_1 \leq \left(\frac{2^{d+1}\|B\|_2^2}{\epsilon}\right) k^{-2\beta/d}, \;\; \delta_2 \leq \left(\frac{2^{d+1}\|B\|_2^2}{\epsilon}\right) k^{-2\alpha/d}, \;\; \delta_3 = 0,$$

and $p_j = (h/2)^d/k$. Further

$$H^2\left(\int P_\lambda^n \, d\varpi(\lambda), \int Q_\lambda^n \, d\varpi(\lambda)\right) \leq C\left(\frac{2^{d+2}\|B\|_2^2}{\epsilon}\right)\left(\frac{n^2h^d}{k}\right)\left(k^{-4s/d} + k^{-4\alpha/d}\right)$$

for C a constant only depending on ϵ.

Before moving to the proof of Proposition 2, we briefly discuss and give some remarks. Compared to the Hellinger distance arising in the average treatment effect or expected conditional covariance lower bounds [Robins et al., 2009b], 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 $h^\gamma = 4k^{-2s/d}$ 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 $\mathcal{Z}_j = \mathcal{C}_{h/2k^{1/d}}(m_j) \times \{0,1\} \times \{0,1\}$, $j = 1, ..., k$, 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 $h/2k^{1/d}$. Therefore

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

where $p_j = (h/2)^d/k$ is 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 $\ell \neq j$, since λ_j only changes the density in $\mathcal{Z}_j$. Note when $(\lambda_1, ..., \lambda_k)$ are iid Rademacher random variables the marginalized densities are

$$\overline{p}(z) \equiv \int p_\lambda(z)\ d\nu(\lambda) = f(x) \left\{ \frac{1}{4} + (2a-1)(2y-1) \frac{h^\gamma}{4} B\left(\frac{x - x_0}{h} \right) \right\}$$

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

First we show that relevant densities and density ratios are appropriately bounded. In particular, when $h \leq 1/4$ then it follows that on $\mathcal{S}_{hk}$ we have

$$1 \leq f(x) = \left\{ 1 - \left(\frac{4^d - 1}{2^d} \right) h^d \right\}^{-1} \leq 2. \tag{2}$$

Further, since $B(x) \leq \mathbb{1}(x \in [-1,1]^d)$, $a, y \in \{0,1\}$, and $\lambda \in \{-1,1\}$, we have on $\mathcal{S}_{hk}$ that

$$\left(\frac{1}{4} - \frac{k^{-\beta/d}}{2} - \frac{h^\gamma}{4} \right) \leq \frac{p_\lambda(z)}{f(x)} \leq \left(\frac{1}{4} + \frac{k^{-\beta/d}}{2} + \frac{h^\gamma}{4} \right),$$

regardless of the values of $h, k \geq 0$. Therefore when $h^\gamma + 2k^{-\beta/d} \leq 1 - 4\epsilon$, the above bound implies

$$\frac{p_\lambda(z)}{f(x)} \geq \left(\frac{1}{4} - \frac{k^{-\beta/d}}{2} - \frac{h^\gamma}{4} \right) \geq \epsilon. \tag{3}$$

Similarly, when $h^\gamma + 2k^{-\beta/d} \leq 1 - 4\epsilon$ (which implies $h^\gamma \leq 1$) we also have

$$\frac{\overline{p}(z)}{p_\lambda(z)} \leq \frac{\frac{1}{4} + \frac{h^\gamma}{4}}{\frac{1}{4} - \frac{k^{-\beta/d}}{2} - \frac{h^\gamma}{4}} \leq \frac{1/2}{\epsilon}.$$

Note that, although Robins et al. [2009b] assume $p_\lambda(z)$ is uniformly lower bounded away from zero in their version of Lemma 2, they only use a bound on $\overline{p}/p_\lambda$ to ensure their quantity c is

bounded (see page 1319). Therefore this condition also holds in our case. Now it remains to bound the quantities δ_1, δ_2, and δ_3.

We begin with δ_3, which is tackled somewhat differently from δ_1 and δ_2, as it is a distance between the marginalized densities $\overline{p}$ and $\overline{q}$. For it notice that if we take $k^{-2s/d} = h^\gamma/4$ then

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

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

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

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.

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

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

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

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

the inequality in the second line since $p_\lambda(z)/f(x) \geq \epsilon$ and $f(x) \leq 2$ as in (3) and (2), and the last equality since

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

by a change of variables.

For δ_2 we use a mix of the above logic for δ_3 and δ_1. Note that

$$
\begin{aligned}
(q_\lambda - p_\lambda)^2 &= f(x)^2 \Bigg[(a-1/2)k^{-\alpha/d} \sum_{j=1}^{k} \lambda_j B\left(\frac{x-m_j}{h/k^{1/d}}\right) \\
&\qquad + (2a-1)(2y-1)\left\{ k^{-2s/d} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 - \frac{h^\gamma}{4} B\left(\frac{x-x_0}{h}\right) \right\} \Bigg]^2 \\
&\leq 2f(x)^2 \left[\frac{k^{-2\alpha/d}}{4} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 + \left\{ k^{-2s/d} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 - \frac{h^\gamma}{4} B\left(\frac{x-x_0}{h}\right) \right\}^2 \right] \\
&= (1/2) f(x)^2 k^{-2\alpha/d} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2
\end{aligned}
$$

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

$$
\begin{aligned}
\delta_2 &= \left(\frac{2^d k}{h^d}\right) \max_\ell \sup_\lambda \int_{\mathcal{X}_\ell} \sum_{a,y} \frac{f(x)^2}{4p_\lambda(z)} k^{-2\alpha/d} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 \, dx \\
&\leq \left(\frac{2^d k}{h^d}\right)\left(\frac{2}{\epsilon}\right) k^{-2\alpha/d} \max_\ell \int_{\mathcal{X}_\ell} \sum_{j=1}^{k} B\left(\frac{x-m_j}{h/k^{1/d}}\right)^2 \, dx = \left(\frac{2^d \|B\|_2^2}{\epsilon/2}\right) k^{-2\alpha/d}
\end{aligned}
$$

using the exact same logic as for δ_1. □

3.3 Choice of Parameters & Final Rate

Finally we detail how the parameters h and k can be chosen to ensure the Hellinger distance from Proposition 2 remains bounded, and use the result to finalize the proof of Theorem 1.

Proposition 3. *Let*

$$
h = (4k^{-2s/d})^{1/\gamma} \quad \text{and} \quad k = (C^* n^2)^{d/(4s+d+2sd/\gamma)}
$$

for $C^* = 2^{2d/\gamma+d+3} C \|B\|_2^2/\epsilon$ *and* C *the constant from Proposition 2. Then under the assumptions of Proposition 2 we have*

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

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

The proof of Proposition 3 follows directly from Proposition 2, after plugging in the selected values of h and k. Importantly, it 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 $\frac{1-\sqrt{\alpha(1-\alpha/4)}}{2} = \frac{1-\sqrt{3/4}}{2} \approx 0.067$ in Lemma 1 is a constant (greater than $1/20$ for example), while the separation in the CATE at x_0, which equals h^γ, is proportional to $n^{-1/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\right)}$ under all P_λ and Q_λ. Therefore this separation is indeed the minimax rate in the low smoothness regime where $s < \frac{d/4}{1+d/2\gamma}$. Note again that, as discussed in Remark 1, when $s > \frac{d/4}{1+d/2\gamma}$ the rate $n^{-1/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\right)}$ 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 [2009] 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.

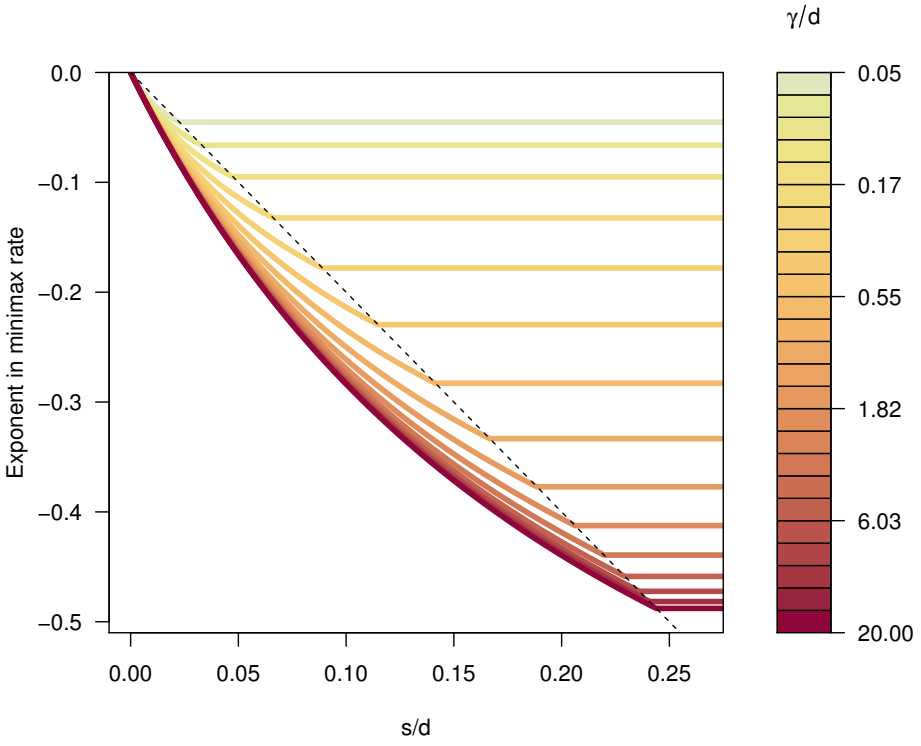


Figure 2: The minimax rate for CATE estimation, as a function of average nuisance smoothness s and CATE smoothness γ, each scaled by covariate dimension d. The black dotted line denotes a threshold on the nuisance smoothness s/d, below which the oracle nonparametric regression rate $n^{-1/(2+d/\gamma)}$ 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. In other words, more smoothness can never hurt. However, there is an important elbow phenomenon, akin to that found in functional estimation problems

[Bickel and Ritov, 1988, Birgé and Massart, 1995, Robins et al., 2009b, Tsybakov, 2009]. 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 [2020].

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,

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

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, e.g., $\gamma \to \infty$, 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 approaches

$$s > \frac{d/4}{1+d/2\beta} \iff \alpha > \frac{-\beta}{1+d/2\beta}$$

which 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, i.e., that $s > \frac{d/4}{1+d/2\gamma}$. One exception is the undersmoothed R-learner estimator analyzed in Kennedy [2020], which did achieve the rate $n^{-1/(2+d/\gamma)}$ whenever $s > \frac{d/4}{1+d/2\gamma}$, under some conditions. However, in the low-smoothness regime where $s < \frac{d/4}{1+d/2\gamma}$, 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 4. A slightly modified version of our construction also reveals that, when the CATE $\tau(x) = \tau$ is constant, the classic functional estimation rate $n^{-1/\left(1+\frac{d}{4s}\right)}$ 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 [Kennedy, 2020, Nie and Wager, 2021], based on an adaptation of higher-order influence functions [Robins et al., 2008, 2009a, 2017].

4.1 Proposed Estimator & 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 following subsections. In short, the estimator is a higher-order influence function-based version of the local polynomial R-learner analyzed in Kennedy [2020]. 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 [1987], Robinson [1988], and Robins [1994], with more flexible series, RKHS, and lasso versions studied more recently by Robins et al. [2008], Nie and Wager [2021], and Chernozhukov et al. [2017], respectively. This previous work did not obtain 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, $j = 1, ..., d$, define $\rho(x_j) = \{\rho_0(x_j), \rho_1(x_j), ..., \rho_{\lfloor\gamma\rfloor}(x_j)\}^{\mathrm{T}}$ as the first $(\lfloor\gamma\rfloor + 1)$ terms of the Legendre polynomial series (shifted to be orthonormal on $[0, 1]$),

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

Define $\rho(x)$ to be the corresponding tensor product of all interactions of $\rho(x_1), ..., \rho(x_d)$ 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 $\rho_h(x) = \rho\left(1/2 + (x - x_0)/h\right)$. The proposed estimator is then defined as

$$\widehat{\tau}(x_0) = \rho_h(x_0)^{\mathrm{T}}\widehat{Q}^{-1}\widehat{R} \tag{4}$$

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

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

respectively, and

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

for $b : \mathbb{R}^d \mapsto \mathbb{R}^k$ a basis of dimension k. The nuisance estimators $(\widehat{F}, \widehat{\pi}, \widehat{\mu})$ 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 basis dimension k, which controls how bias and variance are balanced in the second-order U-statistic terms in $\widehat{Q}$ and $\widehat{R}$. The latter 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

$$\arg\min_{\beta} \mathbb{E}\left[K_h(x)\pi(x)\{1-\pi(x)\}\Big\{\tau(x) - \beta^{\mathrm{T}}\rho_h(x)\Big\}^2\right] = Q^{-1}R \tag{5}$$

for

$$Q = \int \rho_h(x)K_h(x)\varphi_{a1}(z)\rho_h(x)^{\mathrm{T}}\ d\mathbb{P}(z) = \int \rho_h(x)K_h(x)\pi(x)\{1-\pi(x)\}\rho_h(x)^{\mathrm{T}}\ dF(x)$$
$$R = \int \rho_h(x)K_h(x)\varphi_{y1}(z)\ d\mathbb{P}(z) = \int \rho_h(x)K_h(x)\pi(x)\{1-\pi(x)\}\tau(x)\ dF(x).$$

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 4. *Let $\tau_h(x_0) = \rho_h(x_0)^{\mathrm{T}}Q^{-1}R$ denote the projection parameter from (5), 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

$$|\tau_h(x_0) - \tau(x_0)| \lesssim h^\gamma.$$

Proof. This proof follows from a higher-order kernel argument (e.g., Proposition 1.13 of Tsybakov [2009], Proposition 4.1.5 of Giné and Nickl [2021]), 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 [2020]. To ease notation we prove the result in the $d = 1$ case but the logic is the same when $d > 1$.

First note that the local polynomial projection operator $Lg(x) \equiv \int g(x)w_h(x)\ dx$ for

$$w_h(x) \equiv \rho_h(x_0)^{\mathrm{T}}Q^{-1}\rho_h(x)K_h(x)\pi(x)\{1-\pi(x)\}f(x)$$

reproduces polynomials, in the sense that, for any polynomial of the form $g(x) = a^{\mathrm{T}}\rho_h(x)$, $a \in \mathbb{R}^q$, we have

$$\begin{aligned} Lg(x) &= \int \Big\{a^{\mathrm{T}}\rho_h(x)\Big\}w_h(x)\ dx \\ &= \rho_h(x_0)^{\mathrm{T}}Q^{-1}\int \rho_h(x)K_h(x)\pi(x)\{1-\pi(x)\}\rho_h(x)^{\mathrm{T}}f(x)\ dx\ a \\ &= \rho_h(x_0)^{\mathrm{T}}Q^{-1}Qa = a^{\mathrm{T}}\rho_h(x_0) = g(x_0). \end{aligned}$$

Therefore

$$\begin{aligned}
\int w_h(x)\tau(x)\ dx - \tau(x_0) &= \int w_h(x)\Big\{\tau(x) - \tau(x_0)\Big\}\ dx \\
&= \int w_h(x) \left\{ \sum_{j=1}^{\lfloor\gamma\rfloor-1} \frac{D^j\tau(x_0)}{j!}(x-x_0)^j + \frac{D^{\lfloor\gamma\rfloor}\tau(x^*)}{\lfloor\gamma\rfloor!}(x-x_0)^{\lfloor\gamma\rfloor} \right\} dx \\
&= 0 + \int w_h(x) \left\{ \frac{D^{\lfloor\gamma\rfloor}\tau(x_0+\epsilon(x-x_0)) - D^{\lfloor\gamma\rfloor}\tau(x_0)}{\lfloor\gamma\rfloor!} \right\} (x-x_0)^{\lfloor\gamma\rfloor}\ dx \\
&\leq \int |w_h(x)| \frac{C\|x-x_0\|^{\gamma-\lfloor\gamma\rfloor}}{\lfloor\gamma\rfloor!} \|x-x_0\|^{\lfloor\gamma\rfloor}\ dx \\
&= \frac{Ch^\gamma}{\lfloor\gamma\rfloor!} \left\{ h^d \int \|u\|^\gamma\ |w_h(x_0+hu)|\ du \right\} \qquad (6)
\end{aligned}$$

where the second line follows by Taylor expansion (with $x^* = x_0 + \epsilon(x - x_0)$ for some $\epsilon \in [0, 1]$), the third by the polynomial reproducing property, the fourth since τ is γ-smooth, and the last by a change of variable with $u = (x - x_0)/h$ (so that $dx = h^d\ du$).

Now the result follows since we show the term on the right in (6) is bounded under the stated assumptions. Specifically

$$\begin{aligned}
h^d \int \|u\|^\gamma\ |w_h(x_0+hu)|\ du &= \int \|u\|^\gamma \left|\rho_h(x_0)^{\mathrm{T}} Q^{-1} \rho_h(x_0+hu)\right| \mathbb{1}(\|u\| \leq 1/2) \\
&\qquad \times \pi(x_0+hu)\{1-\pi(x_0+hu)\} f(x_0+hu)\ du \\
&\leq \left(\frac{1/4}{2^\gamma}\right) \|\rho(1/2)\| \left\|Q^{-1}\right\| \int_{-1/2}^{1/2} \|\rho(1/2+u)\| f(x_0+hu)\ du \\
&= \frac{Cq^2/4}{2^\gamma} \left\|Q^{-1}\right\| h^{-d} \int \mathbb{1}(\|x-x_0\| \leq h/2)\ dF(x) \lesssim 1
\end{aligned}$$

where the second line follows from the submultiplicative property of the operator norm and since $\pi(1-\pi) \leq 1/4$, and the last from Assumptions 2 and 3 and since

$$\|\rho(x)\|^2 \leq \binom{d+\lfloor\gamma\rfloor}{\lfloor\gamma\rfloor}(2\lfloor\gamma\rfloor+1) \leq Cq^2$$

for all x (Belloni et al. [2015], Example 3.1), since each Legendre term satisfies $|\rho_m(x_j)| \leq \sqrt{2m+1} \leq \sqrt{2\lfloor\gamma\rfloor+1}$ for $m \leq \lfloor\gamma\rfloor$, and the length of ρ is $q = \binom{d+\lfloor\gamma\rfloor}{\lfloor\gamma\rfloor}$, i.e., the maximum number of monomials in a polynomial in d variables with degree up to $\lfloor\gamma\rfloor$. □

Before continuing, we first 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.

Proposition 5. *If (i) $\epsilon \leq \pi(x) \leq 1-\epsilon$ and (ii) the density $dF(x)$ is bounded above and below away from zero on $\{x : \|x - x_0\| \leq h/2\}$, then the eigenvalues of Q are bounded above and below away from zero.*

Proof. Define the stretched function $g^*(v) = g(x_0 + h(v - 1/2))$ for any $g : \mathbb{R}^d \mapsto \mathbb{R}$. This maps values of g in the small cube $[x_0 - h/2, x_0 + h/2]^d$ around x_0 to the whole space $[0,1]^d$. Then note that, with the change of variables $v = \frac{1}{2} + \frac{x - x_0}{h}$,

$$\begin{aligned} Q &= \int \rho_h(x) K_h(x) \pi(x) \{1 - \pi(x)\} \rho_h(x)^{\mathrm{T}} \, dF(x) \\ &= \int_{\|v - 1/2\| \leq 1/2} \rho(v) \rho(v)^{\mathrm{T}} \pi^*(v) \{1 - \pi^*(v)\} \, dF^*(v). \end{aligned}$$

Next note that $\epsilon(1 - \epsilon) \leq \pi(1 - \pi) \leq 1/4$, so the eigenvalues of Q will be bounded if those of the matrix

$$\int \rho(v) \rho(v)^{\mathrm{T}} \, dF(x_0 + h(v - 1/2))$$

are. But $\int \rho(x) \rho(x)^{\mathrm{T}} \, dx = I$ by orthonormality of the Legendre polynomials on $[0,1]^d$, and the local boundedness of dF ensures $dF^*/d\mu$ is bounded above and below away from zero, for μ the uniform measure. Therefore Proposition 2.1 of Belloni et al. [2015] yields the result. □

As mentioned above, the estimator (4) can be viewed as a modified higher-order (specifically, second-order) estimator of the projection parameter. To see this, first note that the first term in $\widehat{R}$, i.e.,

$$\mathbb{P}_n \Big\{ \rho_h(X) K_h(X) \widehat{\varphi}_{y1}(Z) \Big\},$$

is the usual first-order influence function-based estimator of R. Kennedy [2020] analyzed an undersmoothed version of this estimator (where the nuisance estimates $\widehat{\pi}$ and $\widehat{\mu}$ themselves are undersmoothed linear smoothers), calling it the local polynomial R-learner. The second term

$$\mathbb{U}_n \Big\{ \rho_h(X_1) K_h(X_1) \widehat{\varphi}_{y2}(Z_1, Z_2) K_h(X_2) \Big\}$$

is similar to the second-order U-statistic correction that would be added using the higher-order influence function methodology developed by Robins et al. [2008, 2009a, 2017]. However, this term differs 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, i.e., 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 parameter R in (5). Second, the basis functions

$$b_h(x) = b\left(1/2 + \frac{x - x_0}{h}\right) \mathbb{1}(\|x - x_0\| \leq h/2)$$

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., $b_h(x_0 - h/2) = b(0)$, $b_h(x_0) = b(1/2)$, 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). These somewhat subtle distinctions play a crucial role in appropriately controlling bias, as will be described in more detail shortly.

Remark 5. Note again that, as with other higher-order estimators, the estimator (4) 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 $\widehat{\Omega}$ non-invertible; this occurs with higher-order estimators of pathwise differentiable functionals as well [Mukherjee et al., 2017]. As discussed in Remark 7, 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 4 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 6. *The estimator (4) satisfies*

$$|\widehat{\tau}(x_0) - \tau_h(x_0)| \;\leq\; \|\rho(1/2)\| \left(\|\|Q^{-1}\| + \|\widehat{Q}^{-1} - Q^{-1}\|\right) \left(\left\|\widehat{R} - R\right\| + \left\|Q - \widehat{Q}\right\|_2 \|Q^{-1}R\|\right),$$

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

$$\begin{aligned}\mathbb{E}\,|\widehat{\tau}(x_0) - \tau_h(x_0)| \;\lesssim\; &\max_j \sqrt{\mathbb{E}\left\{\mathbb{E}(\widehat{R}_j - R_j \mid D^n)^2 + var(\widehat{R}_j \mid D^n)\right\}} \\ &+ \max_{j,\ell} \sqrt{\mathbb{E}\left\{\mathbb{E}(\widehat{Q}_{j\ell} - Q_{j\ell} \mid D^n)^2 + var(\widehat{Q}_{j\ell} \mid D^n)\right\}}.\end{aligned}$$

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

Proof. We have

$$\begin{aligned}|\widehat{\tau}(x_0) - \tau_h(x_0)| &= \left|\rho_h(x_0)^{\mathrm{T}} \widehat{Q}^{-1} \left\{\left(\widehat{R} - R\right) + \left(Q - \widehat{Q}\right) Q^{-1} R\right\}\right| \\ &\leq \|\rho(1/2)\| \left\|\widehat{Q}^{-1}\right\| \left(\left\|\widehat{R} - R\right\| + \left\|Q - \widehat{Q}\right\| \|Q^{-1}R\|\right) \\ &\leq \|\rho(1/2)\| \left(\|Q^{-1}\| + \left\|\widehat{Q}^{-1} - Q^{-1}\right\|\right) \left(\left\|\widehat{R} - R\right\| + \left\|Q - \widehat{Q}\right\|_2 \|Q^{-1}R\|\right)\end{aligned}$$

by the sub-multiplicative and triangle inequalities of the operator norm, along with the fact that $\|A\| \leq \|A\|_2$. Together with the bounds on $\|Q^{-1}\|$, $\|\widehat{Q}^{-1} - Q^{-1}\|$, and $\|Q^{-1}R\|$, this yields the first inequality. For the second inequality, first note $\|\rho(x)\| \leq Cq$, as described in the proof of Proposition 4. The second inequality now follows since

$$\begin{aligned}\mathbb{E}\|\widehat{R} - R\| \leq \sqrt{\mathbb{E}\|\widehat{R} - R\|^2} &= \sqrt{\sum_j \mathbb{E}\left[\mathbb{E}\left\{(\widehat{R}_j - R_j)^2 \mid D^n\right\}\right]} \\ &= \sqrt{\sum_j \mathbb{E}\left\{\mathrm{bias}(\widehat{R}_j \mid D^n)^2 + \mathrm{var}(\widehat{R}_j \mid D^n)\right\}} \\ &\leq \sqrt{\binom{d + \lfloor\gamma\rfloor}{\lfloor\gamma\rfloor}} \max_j \sqrt{\mathbb{E}\left\{\mathrm{bias}(\widehat{R}_j \mid D^n)^2 + \mathrm{var}(\widehat{R}_j \mid D^n)\right\}}\end{aligned}$$

using Jensen's inequality, iterated expectation, and the usual bias-variance decomposition. The last line follows since the length of R is $\binom{d+\lfloor\gamma\rfloor}{\lfloor\gamma\rfloor}$. The logic is exactly the same for $\mathbb{E}\|\widehat{Q} - Q\|_2 = \mathbb{E}\sqrt{\sum_{j,\ell}(\widehat{Q}_{j\ell} - Q_{j\ell})^2}$. □

Thus Proposition 6 tells us that bounding the conditional bias and variance of the components of $\widehat{R}$ and $\widehat{Q}$ 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 estimators $\widehat{R}_j$ and $\widehat{Q}_{j\ell}$, relative to the components of the projection parameter (5), 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 for approximation in a shrinking window around x_0.

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

$$\|(I - \Pi_b)g\|_{F^*} \lesssim k^{-s/d} \text{ for any } s\text{-smooth function } g \tag{7}$$

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 $[0,1]^d$. The approximating condition (7) 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. [2015] for more discussion and specific examples (see their Condition A.3 and subsequent discussion in, for example, their Section 3.2).

Proposition 7. *Assume:*

1. $\lambda_{\max}(\Omega)$ *is bounded above,*
2. *the basis* b *satisfies approximating condition* (7)*,*
3. $\pi(x) - \widehat{\pi}(x)$ *is* α*-smooth,*
4. $\mu(x) - \widehat{\mu}(x)$ *is* β*-smooth.*

Then

$$|\mathbb{E}(\widehat{R}_j - R_j \mid D^n)| \lesssim \left(\frac{k}{h^d}\right)^{-2s/d} + \|\widehat{\pi} - \pi\|_{F^*} \|\widehat{\mu} - \mu\|_{F^*} \|\widehat{\Omega}^{-1} - \Omega^{-1}\|$$
$$|\mathbb{E}(\widehat{Q}_{j\ell} - Q_{j\ell} \mid D^n)| \lesssim \left(\frac{k}{h^d}\right)^{-2\alpha/d} + \|\widehat{\pi} - \pi\|_{F^*}^2 \|\widehat{\Omega}^{-1} - \Omega^{-1}\|.$$

Before delving into the proof, we give some brief discussion. The bias consists of two terms; the first is the main bias term that would result even if the covariate distribution F were known, and the second is essentially the contribution from having to estimate F. 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 $b_h(x)$ in Definition (4), 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 7, 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 and square of nuisance errors for R_j and $Q_{j\ell}$, respectively (in $L_2(F^*)$ norm). In Proposition 9, given after the proof of Proposition 7, we show how the operator norm error of $\widehat{\Omega}$ translates to estimation error in the distribution F itself.

Proof. We only prove the result for $\widehat{R}_j$, since the logic is the same for $\widehat{Q}_{j\ell}$. By iterated expectation, the conditional mean of the first-order term is

$$\begin{aligned}\mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y1}(Z) \mid D^n\} &= R + \int \rho_h(x)K_h(x)\{\pi(x) - \widehat{\pi}(x)\}\{\mu(x) - \widehat{\mu}(x)\}\, dF(x) \\ &= R + \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}\{\mu^*(v) - \widehat{\mu}^*(v)\}\, dF^*(v) \qquad (8)\end{aligned}$$

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 $\{v : \|v - 1/2\| \leq 1/2\}$. Similarly, the conditional mean of the second-order influence function term is

$$\begin{aligned}&\mathbb{E}\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y2}(Z_1, Z_2)K_h(X_2) \mid D^n\} \\ &= -\iint \rho_h(x_1)K_h(x_1)\{\pi(x_1) - \widehat{\pi}(x_1)\}b_h(x_1)^{\mathrm{T}}\widehat{\Omega}^{-1}b_h(x_2)\{\mu(x_2) - \widehat{\mu}(x_2)\}K_h(x_2)\, dF(x_2)\, dF(x_1) \\ &= -\iint \rho(v_1)\{\pi^*(v_1) - \widehat{\pi}^*(v_1)\}b(v_1)^{\mathrm{T}}\widehat{\Omega}^{-1}b(v_2)\{\mu^*(v_2) - \widehat{\mu}^*(v_2)\}\, dF^*(v_2)\, dF^*(v_1) \\ &= -\int \rho(v_1)\{\pi^*(v_1) - \widehat{\pi}^*(v_1)\}\widehat{\Pi}_b(\mu^* - \widehat{\mu}^*)(v_1)\, dF^*(v_1) \qquad (9)\end{aligned}$$

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 $\widehat{\Omega}$. Therefore adding the first- and second-order expected values in (8) and (9), the overall bias relative to R is

$$\begin{aligned}&\int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}(I - \widehat{\Pi}_b)(\mu^* - \widehat{\mu}^*)(v)\, dF^*(v) \\ &\quad = \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}(I - \Pi_b)(\mu^* - \widehat{\mu}^*)(v)\, dF^*(v) \\ &\qquad + \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}(\Pi_b - \widehat{\Pi}_b)(\mu^* - \widehat{\mu}^*)(v)\, dF^*(v) \\ &\quad = \int (I - \Pi_b)\{\rho(\pi^* - \widehat{\pi}^*)\}(v)(I - \Pi_b)(\mu^* - \widehat{\mu}^*)(v)\, dF^*(v) \qquad (10) \\ &\qquad + \int \rho(v)\{\pi^*(v) - \widehat{\pi}^*(v)\}(\Pi_b - \widehat{\Pi}_b)(\mu^* - \widehat{\mu}^*)(v)\, dF^*(v) \qquad (11)\end{aligned}$$

where the last line follows from orthogonality of a projection with its residuals (Lemma 3(i)).

Now we analyze the bias terms (10) and (11) 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_h, the projections in these bias terms are of stretched versions of the nuisance functions $(\pi^* - \widehat{\pi}^*)$ and $(\mu^* - \widehat{\mu}^*)$, on the standard *non-localized* 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

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

where the second equality follows by the chain rule, and the third since g is α-smooth. Thus the above implies $h^{-\alpha}\left|D^{\lfloor\alpha\rfloor}g^*(v) - D^{\lfloor\alpha\rfloor}g^*(v')\right| \lesssim |v - v'|$, i.e., 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 (7) of the basis b, and so it follows that

$$\|(I - \Pi_b)g^*/h^\alpha\|_{F^*} \;\lesssim\; h^\alpha k^{-\alpha/d} = (k/h^d)^{-\alpha/d} \tag{12}$$

for any α-smooth function g.

Therefore now consider the bias term (10). This term satisfies

$$\begin{aligned}&\int \Big[(I - \Pi_b)\{\rho(\pi^* - \widehat{\pi}^*)\}(v)\Big]\Big\{(I - \Pi_b)(\mu^* - \widehat{\mu}^*)(v)\Big\}\, dF^*(v) \\ &\quad\le \|(I - \Pi_b)\{\rho(\pi^* - \widehat{\pi}^*)\}\|_{F^*}\|(I - \Pi_b)(\mu^* - \widehat{\mu}^*)\|_{F^*} \\ &\quad\lesssim\; (k/h^d)^{-2s/d}\end{aligned}$$

where the second line follows by Cauchy-Schwarz, and the third by (12), since $(\pi - \widehat{\pi})$ and $(\mu - \widehat{\mu})$ 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 $(\pi^* - \widehat{\pi}^*)$).

Now for the term in (11), 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

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

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 3(iii)), and the last by the sub-multiplicative property of the operator norm, along with the fact that $\|\Omega^{1/2}\|^2 = \|\Omega\|$. □

Several of our results require the eigenvalues of Ω be bounded above and below away from zero. The next proposition gives simple sufficient conditions for this to hold, just as in Proposition 5 for the matrix Q.

Proposition 8. *If (i) the basis $b(v)$ is orthonormal with respect to the uniform measure, and (ii) the density $dF(x)$ is bounded above and below away from zero on $\{x : \|x - x_0\| \leq h/2\}$, then the eigenvalues of Ω are bounded above and below away from zero.*

Proof. The proof is similar to that of Proposition 5. First note that

$$\Omega = \int b_h(x) K_h(x) b_h(x)^{\mathrm{T}} dF(x) = \int b(v) b(v)^{\mathrm{T}} \, dF^*(v)$$

by the change of variables $v = \frac{1}{2} + \frac{x - x_0}{h}$, and where $dF^*(v) = dF(x_0 + h(v - 1/2))$ as before. Further $\int b(v) b(v)^{\mathrm{T}} \, dv = I$ by the assumed orthonormality, and the local boundedness of dF ensures $dF^*/d\mu$ is bounded above and below away from zero, for μ the uniform measure. Therefore Proposition 2.1 of Belloni et al. [2015] yields the result. □

The next result is a refined version of Proposition 7, 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 7 for more detailed discussion of these conditions, and note that the result follows directly from Proposition 7 together with Lemma 4.

Proposition 9. *Under the assumptions of Proposition 7, if $\alpha \geq \beta$ and additionally*

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. *$\|(d\widehat{F}^*/dF^*) - 1\|_\infty \lesssim \frac{(k/h^d)^{-2s/d}}{\|\widehat{\pi} - \pi\|_{F^*}(\|\widehat{\pi} - \pi\|_{F^*} + \|\widehat{\mu} - \mu\|_{F^*})}$,*

then the bias satisfies

$$|\mathbb{E}(\widehat{R}_j - R_j \mid D^n)| \vee |\mathbb{E}(\widehat{Q}_{j\ell} - Q_{j\ell} \mid D^n)| \;\lesssim\; \left(\frac{k}{h^d}\right)^{-2s/d}.$$

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 10. *Assume:*

1. *y^2, $\widehat{\pi}^2$, $\widehat{\mu}^2$, and $\|\widehat{\mu} - \mu\|_{F^*}$ are all bounded above, and*

2. $\lambda_{\max}(\Omega)$ *is bounded above.*

Then

$$var(\widehat{R}_j \mid D^n) \vee var(\widehat{Q}_{j\ell} \mid D^n) \lesssim \frac{1}{nh^d}\left(1 + \frac{k}{nh^d}\left(1 + \|\widehat{\Omega}^{-1} - \Omega^{-1}\|^2\right)\right).$$

Before giving the proof, we make a few brief comments. First, the variance here is analogous to that of a higher-order (quadratic) influence function estimator (cf. Theorem 1 of Robins et al. [2009a]), 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 7. For the bias, non-trivial 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 will be noted in a remark shortly after the proof).

Proof. As with the bias, we prove the result for $\widehat{R}_j$, since the logic is exactly the same for $\widehat{Q}_{j\ell}$. Given the training data D^n, the estimator $\widehat{R}_j$ is a second-order U-statistic with kernel

$$\xi(Z_1, Z_2) = \rho_h(X_1)K_h(X_1)\Big\{\widehat{\varphi}_{y1}(Z_1) + \widehat{\varphi}_{y2}(Z_1, Z_2)K_h(X_2)\Big\}$$

Thus its conditional variance is given by the usual variance of a second-order U-statistic (e.g., Lemma 6 of Robins et al. [2009a]), which is

$$\begin{aligned}\operatorname{var}\{\mathbb{U}_n(\xi) \mid D^n\} &= \left(\frac{4(n-2)}{n(n-1)}\right)\operatorname{var}\{\xi_1(Z_1) \mid D^n\} + \left(\frac{2}{n(n-1)}\right)\operatorname{var}\{\xi(Z_1, Z_2) \mid D^n\} \\ &\leq \frac{4\mathbb{E}\{\xi_1(Z_1)^2 \mid D^n\}}{n} + \frac{4\mathbb{E}\{\xi(Z_1, Z_2)^2 \mid D^n\}}{n^2} \qquad (13)\end{aligned}$$

for $\xi_1(z_1) = \int \xi(z_1, z_2)\, d\mathbb{P}(z_2)$, if $n \geq 2$. In our case ξ_1 equals

$$\rho_h(X_1)K_h(X_1)\Big\{\widehat{\varphi}_{y1}(Z_1) + \{A_1 - \widehat{\pi}(X_1)\}\widehat{\Pi}_b(\mu^* - \widehat{\mu}^*)(X_1)\Big\}.$$

Therefore for the first term in (13) we have

$$\begin{aligned}\int \xi_1^2\, d\mathbb{P} &\leq 2\left(\int \rho_h^2 K_h^2 \widehat{\varphi}_{y1}^2\, d\mathbb{P} + \int \rho_h^2 K_h^2 \{\pi(1-\pi) + (\pi - \widehat{\pi})^2\}\{\widehat{\Pi}_b(\mu^* - \widehat{\mu}^*)\}^2\, dF\right) \\ &\lesssim \frac{1}{h^d}\left(\int \rho(v)^2\, dF^*(v) + \int \rho(v)^2\{\widehat{\Pi}_b(\mu^* - \widehat{\mu}^*)(v)\}^2\, dF^*(v)\right) \\ &\lesssim \frac{1}{h^d}\left(1 + \|\widehat{\mu} - \mu\|_{F^*}^2\right)\end{aligned}$$

where the second inequality follows by the change of variables $v = \frac{1}{2} + \frac{x - x_0}{h}$ and since $(\widehat{\varphi}_{y1}, \pi, \widehat{\pi})$ are uniformly bounded and $0 \leq K_h(x) \lesssim h^{-d}$, and the third since $\sup_v \|\rho(v)\| \lesssim q$ (Belloni et al. [2015], Example 3.1) and the weighted L_2 norm of a projection is no more than the

weighted $L_2(\mathbb{P})$ norm of the function itself (Lemma 3(ii)). Now, for the second term in (13), letting $\bar{b}_j(x) \equiv \Omega^{-1/2} b_j(x)$ and $M \equiv \Omega^{1/2}\widehat{\Omega}^{-1}\Omega^{1/2}$, we have

$$\begin{aligned}
\mathbb{E}\{\xi(Z_1,Z_2)^2 \mid D^n\} &= \int \left[\rho_h(x_1)\{a_1 - \widehat{\pi}(x_1)\}\Big\{b_h(x_1)^{\mathrm{T}}\widehat{\Omega}^{-1}b_h(x_2)\Big\}\{y_2 - \widehat{\mu}(x_2)\}K_h(x_1)K_h(x_2)\right]^2 \, d\mathbb{P}(z_1)\, d\mathbb{P}(z_2) \\
&\lesssim \frac{1}{h^{2d}} \int \rho_h(x_1)^2 \Big\{b_h(x_1)^{\mathrm{T}}\widehat{\Omega}^{-1}b_h(x_2)\Big\}^2 K_h(x_1)K_h(x_2)\, dF(x_1)\, dF(x_2) \\
&\lesssim \frac{1}{h^{2d}} \int \Big\{\bar{b}(v_1)^{\mathrm{T}} M \bar{b}(v_2)\Big\}^2 \, dF^*(v_1)\, dF^*(v_2) \\
&= \frac{\|M\|_2^2}{h^{2d}} = \left(\frac{1}{h^{2d}}\right) \|\Omega^{1/2}\Omega^{-1}\Omega^{1/2} + \Omega^{1/2}(\widehat{\Omega}^{-1} - \Omega^{-1})\Omega^{1/2}\|_2^2 \\
&\leq 2\left(\frac{1}{h^{2d}}\right)\left(\|I\|_2^2 + \|\Omega^{1/2}(\widehat{\Omega}^{-1} - \Omega^{-1})\Omega^{1/2}\|_2^2\right) \\
&\leq 2\left(\frac{1}{h^{2d}}\right) k\left(1 + \|\Omega\|^2 \|\widehat{\Omega}^{-1} - \Omega^{-1}\|^2\right)
\end{aligned}$$

where the first line follows by definition, the second since $(a - \widehat{\pi})$ and $(y - \widehat{\mu})$ are uniformly bounded and $0 \leq K_h(x) \lesssim h^{-d}$, the third by a change of variables $v = \frac{1}{2} + \frac{x - x_0}{h}$, by definition of $\bar{b}$ and M, and since $\rho(v)$ is bounded, the fourth by definition since $\int \bar{b}\bar{b}^{\mathrm{T}}\, dF^* = I$, and the last two by basic properties of the Frobenius norm (e.g., triangle inequality and $\|A\|_2 \leq \sqrt{k}\|A\|$) and $(a+b)^2 \leq 2(a^2+b^2)$. □

Remark 6. By Lemma 4, under the assumptions of Proposition 9, 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 4 with the decomposition in Proposition 6, and the bias and variance bounds from Proposition 9 and Proposition 10, respectively, shows that

$$\mathbb{E}_P|\widehat{\tau}(x_0) - \tau_P(x_0)| \;\lesssim\; h^\gamma + \left(\frac{k}{h^d}\right)^{-2s/d} + \sqrt{\frac{1}{nh^d}\left(1 + \frac{k}{nh^d}\right)} \tag{14}$$

under all the combined assumptions of these results, which are compiled in the statement of Theorem 2 below. The first two terms in (14) 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 $(k/h^d)^{-2s/d}$ analogous to a squared nuisance bias term, but shrunken due to the stretching induced by the localized basis b_h. 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 (14) by taking the tuning parameters to satisfy

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

ensures the rate matches the minimax lower bound from Theorem 1, proving that lower bound is in fact tight. This is formalized in the following theorem, which we present after compiling and briefly discussing the necessary regularity conditions.

Condition A1. The eigenvalues of Q and Ω are bounded above and below away from zero.

Condition A2. $\pi(x) - \widehat{\pi}(x)$ is α-smooth and $\mu(x) - \widehat{\mu}(x)$ is β-smooth.

Condition A3. The quantities y^2, $(\widehat{\pi}^2, \widehat{\mu}^2)$, $\|\widehat{\mu} - \mu\|_{F^*}$, and $\|\widehat{Q}^{-1} - Q^{-1}\|$ are all bounded above, and $\|d\widehat{F}^*/dF^*\|_\infty$ is bounded above and below away from zero.

Condition A1 is a standard collinearity restriction used with least squares estimators; simple sufficient conditions were given earlier in Propositions 5 and 8. In Lemma 5 in the appendix we also prove that this condition holds for the class of densities used in the model $\mathcal{P}$ in Theorem 1, ensuring that the upper bound holds over the same model. A sufficient condition for Condition A2 to hold is that the estimators $\widehat{\pi}(x)$ and $\widehat{\mu}(x)$ match the (known) smoothness of $\pi(x)$ and $\mu(x)$; this would be the case for standard minimax optimal estimators based on series or local polynomial methods. Condition A3 is a mild boundedness condition on the outcome Y (which could be weakened at the expense of adding some complexity via tail conditions), as well as the nuisance estimators $(\widehat{F}^*, \widehat{\pi}, \widehat{\mu})$, and even weaker, the errors $\|\widehat{\mu} - \mu\|_{F^*}$ and $\|\widehat{Q}^{-1} - Q^{-1}\|$ (which would typically not only be bounded but decreasing to zero).

Theorem 2. *Assume regularity conditions A1–A3, that the basis b satisfies Hölder approximating condition (7), and:*

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

Then if the tuning parameters satisfy $h \sim n^{-(1/\gamma)/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\right)}$ and $k \sim n^{-\left(\frac{d}{\gamma}+\frac{d}{2s}\right)/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\right)}$, the estimator $\widehat{\tau}$ from Definition 2 has error upper bounded as

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

We refer to Section 3.3 for more detailed discussion and visualization of the rate from Theorem 2. However, in the following remark we discuss some details of Condition 1 of Theorem 2, which ensures the covariate distribution is estimated accurately enough in uniform norm.

Remark 7. 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 are estimated. This is because the condition really amounts to a third-order term (the covariate distribution error multiplied by the *squared* nuisance error) being of smaller order than the CATE rate. Specifically, the result of Theorem 2 can also be written as

$$\mathbb{E}_P|\widehat{\tau}(x_0) - \tau_P(x_0)| \lesssim n^{-1/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\vee\left(1+\frac{d}{2\gamma}\right)\right)} + R_{3,n}, \tag{15}$$

for the third-order error term

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

so that Condition 1 simply requires this third-order term to be smaller order than the first minimax optimal rate in (15). Second, we note that we leave the condition in terms of the $L_2(F^*)$ errors $\|\widehat{\pi} - \pi\|_{F^*}$ and $\|\widehat{\mu} - \mu\|_{F^*}$ because, although we assume π and μ are $\alpha-$ 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

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

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

$$1/\left(2+\frac{d}{\zeta}\right) + 2/\left(2+\frac{d}{\alpha}\right) > 1/\left(1+\frac{d}{2\gamma}+\frac{d}{4s}\vee\left(1+\frac{d}{2\gamma}\right)\right).$$

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., page 338 of Robins et al. [2008]) and conditional variance estimation [Shen et al., 2020, Wang et al., 2008].

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. 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. First, the goal of the present work is mostly to further our theoretical understanding of the fundamental limits of CATE estimation; thus 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. We expect that approaches based on Lepski's method could be used, as in Mukherjee et al. [2015] and Liu et al. [2021], but how well these work in practice is unclear. 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 $\widehat{\Omega}$.

Second, in this work we have assumed the propensity score is smoother than the regression function. Interestingly, there seems to be some important asymmetry in the role of these two nuisance functions for the CATE, indicating that it may differ in this respect from the ATE or expected partially missing outcome [Robins et al., 2009b]. In particular, our lower bound construction breaks down when the regression function is smoother, and our upper bound rate is then driven by the lesser propensity score smoothness, rather than the average smoothness. We conjecture that the upper bound could be improved in this regime.

Acknowledgements

EK gratefully acknowledges support from NSF DMS Grant 1810979, NSF CAREER Award 2047444, and NIH R01 Grant LM013361-01A1, and SB and LW from NSF Grant DMS1713003. EK also thanks Matteo Bonvini for very helpful discussions.

References

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: pointwise and uniform results. *Journal of Econometrics*, 186(2): 345–366, 2015.

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā*, pages 381–393, 1988.

L. Birgé and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, 23(1):11–29, 1995.

A. Carbery and J. Wright. Distributional and l^q norm inequalities for polynomials over convex bodies in r^n. *Mathematical Research Letters*, 8(3):233–248, 2001.

G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

V. Chernozhukov, M. Goldman, V. Semenova, and M. Taddy. Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*, 2017.

Q. Fan, Y.-C. Hsu, R. P. Lieli, and Y. Zhang. Estimation of conditional average treatment effects with high-dimensional data. *arXiv preprint arXiv:1908.02399*, 2019.

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

Z. Gao and Y. Han. Minimax optimal nonparametric estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2002.06471*, 2020.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2021.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

I. A. Ibragimov, A. S. Nemirovskii, and R. Khas'minskii. Some problems on nonparametric estimation in gaussian white noise. *Theory of Probability & Its Applications*, 31(3):391–406, 1987.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Y. Ingster, J. I. Ingster, and I. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2003.

E. H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arxiv*, 2020.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

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

L. Liu, R. Mukherjee, J. M. Robins, and E. J. Tchetgen Tchetgen. Adaptive estimation of nonparametric functionals. *Journal of Machine Learning Research*, 22(99):1–66, 2021.

A. R. Luedtke and M. J. van der Laan. Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, 12(1):305–332, 2016.

R. Mukherjee, E. J. Tchetgen Tchetgen, and J. M. Robins. Lepski's method and adaptive estimation of nonlinear integral functionals of density. *arXiv preprint arXiv:1508.00249*, 2015.

R. Mukherjee, W. K. Newey, and J. M. Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

A. Nemirovski. Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

J. M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412, 1994.

J. M. Robins, S. D. Mark, and W. K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495, 1992.

J. M. Robins, L. Li, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421, 2008.

J. M. Robins, L. Li, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Quadratic semiparametric von mises calculus. *Metrika*, 69(2-3):227–247, 2009a.

J. M. Robins, E. J. Tchetgen Tchetgen, L. Li, and A. W. van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009b.

J. M. Robins, L. Li, R. Mukherjee, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Minimax estimation of a functional on a structured high dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

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

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

Y. Shen, C. Gao, D. Witten, and F. Han. Optimal estimation of variance in nonparametric regression with random design. *The Annals of Statistics*, 48(6):3589–3618, 2020.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

M. J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

S. Vansteelandt and M. M. Joffe. Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

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

L. Wang, L. D. Brown, T. T. Cai, and M. Levine. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664, 2008.

Q. Zhao, D. S. Small, and A. Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.

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

6 Technical Lemmas

Lemma 3. *Let* $\Pi_w f(t) = b(t)^{\mathrm{T}}\Omega^{-1}\int b(x)w(x)f(x)\ d\mathbb{P}(x)$ *denote a* w*-weighted projection with* $\Omega = \int bwb^T d\mathbb{P}$ *for* $w \geq 0$*. And let* $\theta_{w,f} = \Omega^{-1}\int b(x)w(x)f(x)\ d\mathbb{P}(x)$ *denote the coefficients of the projection. Then:*

(i) projections are orthogonal to residuals, i.e.,

$$\int (\Pi_w f)(I - \Pi_w)g\ wd\mathbb{P} = 0,$$

(ii) the $L_2(w\mathbb{P})$ *norm of a projection is no more than that of the function, i.e.,*

$$\int (\Pi_w f)^2 w\ d\mathbb{P} \leq \int f^2\ wd\mathbb{P},$$

(iii) the L_2 *norm of the scaled coefficients is no more than the* $L_2(w\mathbb{P})$ *norm of the function, i.e.,*

$$\|\Omega^{1/2}\theta_{w,f}\|^2 \leq \int f^2\ wd\mathbb{P}.$$

Proof. For (i) let $b^*(x) = \Omega^{-1/2}b(x)$ so that $\int b^* w (b^*)^{\mathrm{T}} d\mathbb{P} = I$ and note that

$$\begin{aligned}
\int (\Pi_w f)(I - \Pi_w)g\ wd\mathbb{P} &= \int\int b(x)^{\mathrm{T}}\Omega^{-1}b(t)f(t)g(x)\ wd\mathbb{P}(t)\ wd\mathbb{P}(x) \\
&\qquad - \int b(x)^{\mathrm{T}}\Omega^{-1}b(t)f(t)b(x)^{\mathrm{T}}\Omega^{-1}b(u)g(u)\ wd\mathbb{P}(t)\ wd\mathbb{P}(x)\ wd\mathbb{P}(u) \\
&= \int b^*(x)^{\mathrm{T}}b^*(t)f(t)g(x)\ wd\mathbb{P}(t)\ wd\mathbb{P}(x) \\
&\qquad - \int f(t)b^*(t)^{\mathrm{T}}b^*(x)b^*(x)^{\mathrm{T}}b^*(u)g(u)\ wd\mathbb{P}(t)\ wd\mathbb{P}(x)\ wd\mathbb{P}(u) \\
&= \int b^*(x)^{\mathrm{T}}b^*(t)f(t)g(x)\ wd\mathbb{P}(t)\ wd\mathbb{P}(x) \\
&\qquad - \int f(t)b^*(t)^{\mathrm{T}}b^*(u)g(u)\ wd\mathbb{P}(t)\ wd\mathbb{P}(u) = 0
\end{aligned}$$

where the last line follows since $\int b^*(x)b^*(x)^{\mathrm{T}}\ wd\mathbb{P}(x) = I$.

For (ii) we have by definition that

$$\begin{aligned}\int \{\Pi_w f(x)\}^2 w(x)\ dF(x) &= \int \left\{ \int f(u)w(u)b(u)^{\mathrm{T}}\ d\mathbb{P}(u)\Omega^{-1}b(x) \right\} \left\{ b(x)^{\mathrm{T}}\Omega^{-1} \int b(t)w(t)f(t)\ d\mathbb{P}(t) \right\} w(x) d\mathbb{P}(x) \\ &= \left\{ \int f(u)w(u)b(u)^{\mathrm{T}}\ d\mathbb{P}(u)\Omega^{-1} \right\} \left\{ \int b(t)w(t)f(t)\ d\mathbb{P}(t) \right\} \qquad (16) \\ &= \int f(t)\Pi_w f(t)w(t)\ d\mathbb{P}(t)\end{aligned}$$

and so

$$\begin{aligned}\int (f - \Pi_w f)^2 w\ d\mathbb{P} &= \int f^2 w\ d\mathbb{P} - 2\int f\Pi_w f w\ d\mathbb{P} + \int (\Pi_w f)^2 w\ d\mathbb{P} \\ &= \int f^2 w\ d\mathbb{P} - \int (\Pi_w f)^2 w\ d\mathbb{P}\end{aligned}$$

which implies the result as long as $w \geq 0$, so that the far left side is non-negative.

For (iii) we have

$$\begin{aligned}\|\Omega^{1/2}\theta_{w,f}\|^2 &= (\Omega^{1/2}\theta_{w,f})^{\mathrm{T}}(\Omega^{1/2}\theta_{w,f}) = \theta_{w,f}^{\mathrm{T}}\Omega\theta_{w,f} \\ &= \left(\int b^{\mathrm{T}} w f\ d\mathbb{P} \right) \Omega^{-1} \left(\int bwf\ d\mathbb{P} \right) \\ &= \int \{\Pi_w f(x)\}^2 w(x)\ d\mathbb{P}(x)\end{aligned}$$

where the last equality holds by (16), and so the result follows from Lemma 3(ii). □

Lemma 4. *Assume:*

1. $0 < b \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq B < \infty$
2. $0 < c \leq \|d\widehat{F}^*/dF^*\|_\infty \leq C < \infty$

Then

$$bc \leq \lambda_{\min}(\widehat{\Omega}) \leq \lambda_{\max}(\widehat{\Omega}) \leq BC$$

and

$$\|\widehat{\Omega}^{-1} - \Omega^{-1}\| \leq \left(\frac{B}{b^2 c} \right) \|(d\widehat{F}^*/dF^*) - 1\|_\infty.$$

Proof. The logic mirrors that of the proof of Proposition 2.1 in Belloni et al. [2015]. Note that

$$\begin{aligned}a^{\mathrm{T}}\widehat{\Omega}a &= \int \left\{ a^{\mathrm{T}}b(v) \right\}^2 d\widehat{F}^*(v) \\ &\leq \|d\widehat{F}^*/dF^*\|_\infty \int \left\{ a^{\mathrm{T}}b(v) \right\}^2 dF^*(v) \\ &= \|d\widehat{F}^*/dF^*\|_\infty\ a^{\mathrm{T}}\Omega a\end{aligned}$$

and by the same logic, $a^{\mathrm{T}}\Omega a \leq \|dF^*/d\widehat{F}^*\|_\infty \; a^{\mathrm{T}}\widehat{\Omega}a$. Therefore

$$\lambda_{\max}(\widehat{\Omega}) = \max_{\|a\|=1} a^{\mathrm{T}}\widehat{\Omega}a \leq \|d\widehat{F}^*/dF^*\|_\infty \lambda_{\max}(\Omega)$$
$$\lambda_{\min}(\widehat{\Omega}) = \min_{\|a\|=1} a^{\mathrm{T}}\widehat{\Omega}a \geq \|dF^*/d\widehat{F}^*\|_\infty^{-1} \lambda_{\min}(\Omega)$$

by the min-max theorem, which gives the first inequality. For the second, note

$$\begin{aligned}\|\widehat{\Omega}^{-1} - \Omega^{-1}\| &= \|\widehat{\Omega}^{-1}(\Omega - \widehat{\Omega})\Omega^{-1}\| \\ &\leq \|\widehat{\Omega}^{-1}\|\|\Omega - \widehat{\Omega}\|\|\Omega^{-1}\|\end{aligned}$$

by the sub-multiplicative property of the operator norm, and then

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

by the same logic as above. □

Lemma 5. *Assume:*

1. *$dF(x)$ satisfies $\int \mathbb{1}\{\|x - x_0\| \leq h/2\} \; dF(x) \asymp h^d$,*
2. *$dF(x)$ is bounded above and below away from zero on its local support $\{x \in \mathbb{R}^d : dF(x) > 0, \|x - x_0\| \leq h/2\}$, which is a union of no more than k disjoint cubes all with proportional volume, for h and k defined in Proposition 3.*

Let $dF^*(v) = dF(x_0 + h(v - 1/2))$ denote the distribution in $\mathcal{B}_h(x_0)$, the h-ball around x_0, mapped to $[0,1]^d$, and similarly for π^*. Then the eigenvalues of

$$Q = \int \rho(v)\rho(v)^{\mathrm{T}}\pi^*(v)\{1 - \pi^*(v)\} \; dF^*(v)$$

are bounded above and below away from zero, and there exists a basis b with Hölder approximation property (7), for which the eigenvalues of

$$\Omega = \int b(v)b(v)^{\mathrm{T}} \; dF^*(v)$$

are bounded above and below away from zero.

Proof. Note that since $\epsilon(1-\epsilon) \leq \pi(1-\pi) \leq 1/4$ and since $dF(x)$ is bounded above and below on its support, the eigenvalues of Q will be bounded if those of the matrix

$$\sum_{j=1}^{k} \int \rho(v)\rho(v)^{\mathrm{T}}\mathbb{1}(v \in M_j) \; dv$$

are, where M_j indicates the jth disjoint cube making up the local support of F. By the min-max theorem, the eigenvalues are bounded by the min and max of

$$a^{\mathrm{T}} \int \rho(v)\rho(v)^{\mathrm{T}} \sum_j \mathbb{1}(v \in M_j) \; dv \; a = \int \left\{a^{\mathrm{T}}\rho(v)\right\}^2 \sum_j \mathbb{1}(v \in M_j) \; dv$$

over all $a \in \mathbb{R}^q$ with $\|a\| = 1$. First consider lower bounding the eigenvalues. Note $g(v) = a^{\mathrm{T}}\rho(v)$ is a polynomial of degree at most q. Therefore

$$\begin{aligned}
\sum_j \int g(v)^2 \mathbb{1}(v \in M_j)\, dv &\geq \int g(v)^2 \mathbb{1}\{g(v)^2 \geq \epsilon\} \sum_j \mathbb{1}(v \in M_j)\, dv \\
&\geq \epsilon \int \mathbb{1}\{g(v)^2 \geq \epsilon\} \sum_j \mathbb{1}(v \in M_j)\, dv \\
&\geq \epsilon \left\{ \sum_j \int \mathbb{1}(v \in M_j)\, dv - \int \mathbb{1}\{g(v)^2 < \epsilon\}\, dv \right\} \\
&\geq \epsilon \left(C^* - C\epsilon^{1/2q} \right) = \left(\frac{C^*}{\sqrt{C}} \right)^{4q} > 0
\end{aligned}$$

where the last line follows since

$$\begin{aligned}
\int \sum_j \mathbb{1}(v \in M_j)\, dv &\asymp \int dF^*(v) = \int dF(x_0 + h(v - 1/2)) \\
&= h^{-d} \int \mathbb{1}\{\|x - x_0\| \leq h/2\}\, dF(x) \geq C^*
\end{aligned}$$

from a change of variables $x = x_0 + h(v - 1/2)$, so that $v = 1/2 + (x - x_0)/h$ and $h^d\, dv = dx$, together with Assumption 1 of the lemma, and by Theorem 4 of Carbery and Wright [2001]. The last equality follows if we choose $\epsilon = (C^*/2C)^{2q}$. To upper-bound the eigenvalues, note that

$$\int \left\{ a^{\mathrm{T}}\rho(v) \right\}^2 \sum_j \mathbb{1}(v \in M_j)\, dv \leq \int \left\{ a^{\mathrm{T}}\rho(v) \right\}^2 dv = \|a\|^2 = 1$$

since $\int \rho(v)\rho(v)^{\mathrm{T}}\, dv = I$ by the orthonormality of ρ.

Now consider the eigenvalues of Ω. We will construct a basis of order k for which eigenvalues of Ω are bounded and for which the Hölder approximation property (7) holds.

Remark 8. For simplicity we only consider the case where there are exactly k cubes in the local support of F; if there are finitely many cubes, say $M < \infty$, the arguments are more straightforward, orthonormalizing a standard series M times, once per cube, and taking k/M terms from each. We omit details in the intermediate case where the number of cubes scales at a rate slower than k, but we expect similar arguments to those below can be used.

First, let $\rho(x)$ denote the tensor products of a Legendre basis of order s, orthonormal on $[-1, 1]$, i.e., tensor products of

$$\rho_j(x_\ell) = \frac{\sqrt{j + 1/2}}{2^j j!} \frac{d^j}{dx_\ell^j}(x_\ell^2 - 1)^j$$

for $j \in \{1, ..., s\}$ and $\ell \in \{1, ..., d\}$. This basis satisfies

$$\int \rho(x)\rho(x)^{\mathrm{T}}\, dx = I_s.$$

Now shift and rescale so that the basis is orthonormal on M_j with

$$b_j(x) = 2^d\sqrt{k}\rho\left(4k^{1/d}\left(x - \frac{1}{2} - \frac{m_j - x_0}{h}\right)\right)$$

and so satisfies

$$\begin{aligned}\int b_j(x)b_j(x)^{\mathrm{T}}\mathbb{1}(x \in M_j)\ dx &= 4^d k \int \rho\left(4k^{1/d}\left(x - \frac{1}{2} - \frac{m_j - x_0}{h}\right)\right)\mathbb{1}(x \in M_j)\ dx \\ &= \int_{-1}^{1} \rho(v)\rho(v)^{\mathrm{T}}\ dv = I_s\end{aligned}$$

where we used the change of variable $v = 4k^{1/d}\left(x - \frac{1}{2} - \frac{m_j - x_0}{h}\right)$ so that $dv = 4^d k\ dx$ and $x = v/4k^{1/d} + 1/2 + (m_j - x_0)/h \in M_j$ when $v \in [-1, 1]^d$. Finally define the basis

$$b(v) = \{b_1(v)^{\mathrm{T}}\mathbb{1}(v \in M_1), ..., b_k(v)^{\mathrm{T}}\mathbb{1}(v \in M_k)\}^{\mathrm{T}}$$

of length sk. Then

$$b(v)b(v)^{\mathrm{T}} = \begin{pmatrix} b_1(v)b_1(v)^{\mathrm{T}}\mathbb{1}(v \in M_1) & 0 & \cdots & 0 \\ 0 & b_2(v)b_2(v)^{\mathrm{T}}\mathbb{1}(v \in M_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_k(v)b_k(v)^{\mathrm{T}}\mathbb{1}(v \in M_k) \end{pmatrix}$$

Therefore since $dF^*(x)$ is bounded above and below on its support, the j-th diagonal block of Ω is proportional to

$$\int b_j(v)b_j(v)^{\mathrm{T}}\mathbb{1}(v \in M_j)\ dv = I_s.$$

Therefore the eigenvalues of Ω are all proportional to one and bounded as desired. Further, by the same higher-order kernel arguments for local polynomials as in Proposition 4, the basis satisfies the Hölder approximation property (7). □