A Simple and General Debiased Machine Learning Theorem with Finite Sample Guarantees

Victor Chernozhukov
MIT Economics
vchern@mit.edu

Whitney K. Newey
MIT Economics
wnewey@mit.edu

Rahul Singh
MIT Economics
rahul.singh@mit.edu

Abstract

Debiased machine learning is a meta algorithm based on bias correction and sample splitting to calculate confidence intervals for functionals, i.e. scalar summaries, of machine learning algorithms. For example, an analyst may desire the confidence interval for a treatment effect estimated with a neural network. We provide a nonasymptotic debiased machine learning theorem that encompasses any global or local functional of any machine learning algorithm that satisfies a few simple, interpretable conditions. Formally, we prove consistency, Gaussian approximation, and semiparametric efficiency by finite sample arguments. The rate of convergence is $n^{-1/2}$ for global functionals, and it degrades gracefully for local functionals. Our results culminate in a simple set of conditions that an analyst can use to translate modern learning theory rates into traditional statistical inference. The conditions reveal a general double robustness property for ill posed inverse problems.

1 Introduction

The goal of this paper is to provide a useful technical result for analysts who desire confidence intervals for functionals, i.e. scalar summaries, of machine learning algorithms. For example, the functional of interest could be the average treatment effect of a medical intervention, and the machine learning algorithm could be a neural network trained on medical scans. Alternatively, the functional of interest could be the price elasticity of consumer demand, and the machine learning algorithm could be a kernel ridge regression trained on economic transactions. Treatment effects and price elasticities for a specific demographic are examples of localized functionals. In these various applications, confidence intervals are essential.

We provide a simple set of conditions that can be verified using the kind of rates provided by statistical learning theory. Unlike previous work, we provide a finite sample analysis for any global or local functional of any machine learning algorithm, without bootstrapping, subject to these simple and interpretable conditions. The machine learning algorithm may be estimating a nonparametric regression, a nonparametric instrumental variable regression, or some other nonparametric quantity. We provide conceptual and statistical contributions for the rapidly growing literature on debiased machine learning.

Conceptually, our result unifies, refines, and extends existing debiased machine learning theory for a broad audience. We unify finite sample results that are specific to particular functionals or machine learning algorithms. General asymptotic theory with abstract conditions already exists, which we refine to finite sample theory with simple conditions. In doing so, we uncover a new notion of double robustness for exactly identified ill posed inverse problems. A virtue of finite sample analysis is that it handles the case where the functional involves localization. We show how learning theory delivers inference.

Preprint. Under review.

Statistically, we provide results for the class of global functionals that are mean square continuous, and their local counterparts, using algorithms that have sufficiently fast finite sample learning rates. Formally, we prove (i) consistency, Gaussian approximation, and semiparametric efficiency for global functionals; and (ii) consistency and Gaussian approximation for local functionals. The analysis explicitly accounts for each source of error in any finite sample size. The rate of convergence is the parametric rate of $n^{-1/2}$ for global functionals, and it degrades gracefully to nonparametric rates for local functionals.

2 Related work

By focusing on functionals of nonparametric quantities, this paper continues the tradition of classic semiparametric statistics [22, 34, 9, 28, 3, 33, 2]. Whereas classic semiparametric theory studies functionals of densities or regressions over low dimensional domains, we study functionals of machine learning algorithms over arbitrary domains. In classic semiparametric theory, an object called the Riesz representer appears in efficient influence functions and asymptotic variance calculations [28]. For the same reasons, it appears in debiased machine learning confidence intervals.

In asymptotic inference, the Riesz representer is inevitable. A growing literature directly incorporates the Riesz representer into estimation, which amounts to debiasing known estimators. Doubly robust estimating equations serve this purpose [33]. A geometric perspective emphasizes Neyman orthogonality: by debiasing, the learning problem for the functional becomes orthogonal to the learning problem for the nonparametric object [14, 13, 21]. An analytic perspective emphasizes the mixed bias property: by debiasing, the functional has bias equal to the product of certain learning rates [13, 35]. In this work, we focus on debiased machine learning with doubly robust estimating equations.

With debiasing alone, a key challenge remains: for inference, the function class in which the nonparametric quantity is learned must be Donsker [44, 27, 43, 32], or it must have slowly increasing entropy [6, 7, 45, 24, 41]. However, popular nonparametric settings in machine learning may not satisfy this property. A solution to this challenging issue is to combine debiasing with sample splitting [26]. The targeted [46] and debiased [5, 14, 13] machine learning literatures provide this insight. In particular, debiased machine learning delivers sufficient conditions for asymptotic inference on functionals in terms of learning rates of the underlying nonparametric quantity and the Riesz representer. We complement prior results with a finite sample analysis.

This paper subsumes [38, Section 4].

3 Framework and examples

The general inference problem is to find a confidence interval for some scalar θ_0 in $\mathbb{R}$ where $\theta_0 = E\{m(W, \gamma_0)\}$, γ_0 is in Γ, and $m : \mathcal{W} \times \mathbb{L}_2 \to \mathbb{R}$ is an abstract formula. W in $\mathcal{W}$ is a concatenation of random variables in the model excluding the outcome Y in $\mathcal{Y} \subset \mathbb{R}$. $\mathbb{L}_2$ is the space of functions of the form $\gamma : \mathcal{W} \to \mathbb{R}$ that are square integrable with respect to measure pr. Γ is a linear subset of $\mathbb{L}_2$ known by the analyst, which may be $\mathbb{L}_2$ itself.

Note that γ_0 may be the conditional expectation function $\gamma_0(w) = E(Y \mid W = w)$ or some other nonparametric quantity. For example, it could be the function defined as the solution to the ill posed inverse problem $E(Y \mid W_2 = w_2) = E\{\gamma(W_1) \mid W_2 = w_2\}$ where $W_1, W_2 \subset W$. Such a function is called a nonparametric instrumental variable regression in econometrics [30]. We study the exactly identified case, which amounts to assuming completeness when $\Gamma = \mathbb{L}_2$ [12]. If $W_1 = W_2$ then nonparametric instrumental variable regression simplifies into nonparametric regression.

A local functional $\theta_0^{\lim}$ in $\mathbb{R}$ is a scalar that takes the form

$$\theta_0^{\lim} = \lim_{h \to 0} \theta_0^h, \quad \theta_0^h = E\{m_h(W, \gamma_0)\} = E\{\ell_h(W_j) m(W, \gamma_0)\}, \quad \gamma_0 \text{ in } \Gamma,$$

where ℓ_h is a Nadaraya Watson weighting with bandwidth h and W_j is a scalar component of W. $\theta_0^{\lim}$ is a nonparametric quantity. However, it can be approximated by the sequence (θ_0^h). Each θ_0^h can be analyzed like θ_0 above as long as we keep track of how certain quantities depend on h. By this logic, finite sample semiparametric theory for θ_0^h translates to finite sample nonparametric

theory for $\theta_0^{\lim}$ up to some approximation error. In this sense, our analysis encompasses both global and local functionals.

To illustrate, we consider some classic functionals.

Example 3.1 (Heterogeneous treatment effect estimated by neural network)**.** *Let Y be a health outcome. Let $W = (D, V, X)$ concatenate binary treatment D, covariate of interest V such as age, and other covariates X such as medical scans. Let $\gamma_0(d, v, x) = E(Y \mid D = d, V = v, X = x)$ be a function estimated by a neural network. Under the assumption of selection on observables, the heterogeneous treatment effect is*

$$\text{CATE}(v) = E\{\gamma_0(1, V, X) - \gamma_0(0, V, X) \mid V = v\} = \lim_{h \to 0} E[\ell_h(V)\{\gamma_0(1, V, X) - \gamma_0(0, V, X)\}],$$

where $\ell_h(V) = (h\omega)^{-1} K\{(V - v)/h\}$, $\omega = E[h^{-1} K\{(V - v)/h\}]$, and K is a bounded and symmetric kernel that integrates to one.

The heterogeneous treatment effect is defined with respect to some interpretable, low dimensional characteristic V such as age, race, or gender [1]. The same functional without the localization ℓ_h is the classic average treatment effect. See [8] and [19] for other meaningful localizations of average treatment effect.

Example 3.2 (Regression discontinuity design estimated by random forest)**.** *Let Y be an educational outcome. Let $W = (D, X)$ concatenate test score variable D and covariates X. Let $\gamma_0(d, x) = E(Y \mid D = d, X = x)$ be a function estimated by a random forest. Suppose the cutoff for a scholarship is the test score $D = 0$. The regression discontinuity design parameter is*

$$\text{RDD} = \lim_{d \downarrow 0} E\{\gamma_0(d, X)\} - \lim_{d \uparrow 0} E\{\gamma_0(d, X)\} = \lim_{h \to 0} E\{\ell_h^+(D)\gamma_0(D, X) - \ell_h^-(D)\gamma_0(D, X)\},$$

where $\ell_h^+(D) = (h\omega^+)^{-1} K\{(2D - h)/(2h)\}$, $\omega^+ = E\left[h^{-1} K\{(2D - h)/(2h)\}\right]$, $\ell_h^-(D) = (h\omega^-)^{-1} K\{(-2D - h)/(2h)\}$, $\omega^- = E\left[h^{-1} K\{(-2D - h)/(2h)\}\right]$, and K vanishes outside of the interval $(-1/2, 1/2)$.

The expressions for fuzzy regression discontinuity, exact kink, and fuzzy kink designs are similar.

Example 3.3 (Demand elasticity estimated by kernel instrumental variable regression)**.** *Let Y be log quantity demanded of some good. Let $W = (D, X, Z)$ concatenate log price D, covariates X, and cost shifter Z. Let $\gamma_0(d, x)$ be defined as the solution to $E(Y \mid X = x, Z = z) = E\{\gamma(D, X) \mid X = x, Z = z\}$ estimated by a kernel instrumental variable regression [39]. The demand elasticity is*

$$\text{ELASTICITY} = E\left\{\frac{\partial}{\partial d}\gamma_0(D, X)\right\}.$$

In Supplement 2, we present the additional example of heterogeneous average derivative estimated by lasso, which is useful when an analyst has access to data on household spending behavior.

For our simple and general theorem, we require that the formula m is mean square continuous.

Assumption 3.1 (Linearity and mean square continuity)**.** *Assume that the functional $\gamma \mapsto \mathbb{E}\{m(W, \gamma)\}$ is linear, and that there exist $\bar{Q} < \infty$ and $q > 0$ such that $E\{m(W, \gamma)^2\} \leq \bar{Q}[E\{\gamma(W)^2\}]^q$ for all γ in Γ.*

This condition will be key in Section 5, where we reduce the problem of inference for θ_0 into the problem of learning $(\gamma_0, \alpha_0^{\min})$, where $\alpha_0^{\min}$ is introduced below. It is a powerful condition satisfied by many functionals of interest, or at least satisfied by their approximating sequences. Though the local functional $\theta_0^{\lim}$ does not satisfy Assumption 3.1, each approximating θ_0^h does. In particular, for each m_h there exists some $\bar{Q}_h$ that depends on h. We keep track of $\bar{Q}$ in our analysis and subsequently consider $\bar{Q} = \bar{Q}_h$. See Theorem 5.2 below for conditions that characterize $\bar{Q}_h$ in local functionals, including Examples 3.1 and 3.2.

The restriction that γ_0 is in $\Gamma \subset \mathbb{L}_2$, where Γ is some linear function space, is called a restricted model in semiparametric statistical theory. In learning theory, mean square rates are adaptive to the smoothness of γ_0, encoded by γ_0 in Γ. We quote a general Riesz representation theorem for restricted models.

Lemma 3.1 (Riesz representation [16]). *Suppose Assumption 3.1 holds. Further suppose γ_0 is in Γ. Then there exists a Riesz representer α_0 in $\mathbb{L}_2$ such that for all γ in Γ, $E\{m(W,\gamma)\} = E\{\alpha_0(W)\gamma(W)\}$. There exists a unique minimal Riesz representer $\alpha_0^{\min}$ in $closure(\Gamma)$ that satisfies this equation, obtained by projecting any α_0 onto Γ. Moreover, denoting by $\bar{M}$ the operator norm of $\gamma \mapsto E\{m(W,\gamma)\}$, we have that $[E\{\alpha_0^{\min}(W)^2\}]^{1/2} = \bar{M} \leq \bar{Q}^{1/2} < \infty$.*

The condition $\bar{M} < \infty$ is enough for the conclusions of Lemma 3.1 to hold. Since $\bar{M} \leq \bar{Q}^{1/2}$, $\bar{Q} < \infty$ in Assumption 3.1 is a sufficient condition. Nonetheless, we assume $\bar{Q} < \infty$ because mean square continuity plays a central role in the main results of Section 5. In Examples 3.1 and 3.2, with propensity score $\pi_0(v,x)$,

$$\alpha_0(d,v,x) = \ell_h(v)\left\{\frac{d}{\pi_0(v,x)} - \frac{1-d}{1-\pi_0(v,x)}\right\}; \quad \alpha_0^+(d,x) = \ell_h^+(d), \quad \alpha_0^-(d,x) = \ell_h^-(d).$$

Riesz representation delivers a doubly robust formulation of the target θ_0 in $\mathbb{R}$. For the case where $\gamma_0(w)$ is defined as a nonparametric regression in Γ or projection onto Γ, consider the estimating equation

$$\theta_0 = E[m(W,\gamma_0) + \alpha_0^{\min}(W)\{Y - \gamma_0(W)\}].$$

This formulation is doubly robust since it remains valid if either γ_0 or $\alpha_0^{\min}$ is correct: for all (γ,α) in Γ,

$$\theta_0 = E[m(W,\gamma_0) + \alpha(W)\{Y - \gamma_0(W)\}] = E[m(W,\gamma) + \alpha_0^{\min}(W)\{Y - \gamma(W)\}].$$

The term $\alpha(w)\{y - \gamma(w)\}$ serves as a bias correction for the term $m(w,\gamma)$. We view $(\gamma_0, \alpha_0^{\min})$ as nuisance parameters that we must learn in order to learn and infer θ_0. While any Riesz representer α_0 will suffice for valid learning and inference of $\theta_0 = E\{m(W,\gamma_0)\}$ under correct specification of γ_0 as the regression $E(Y \mid W = w)$ in Γ, the minimal Riesz representer $\alpha_0^{\min}$ confers specification robust inference and semiparametric efficiency for estimating $\theta_0 = E\{m(W,\gamma_0)\}$ when γ_0 is only the projection of $E(Y \mid W = w)$ onto Γ; see [16, Theorem 4.2].

If $\gamma_0(w)$ is defined as the solution to an ill posed inverse problem, then the appropriate Riesz representer is defined as the solution to another ill posed inverse problem [36, 23]. The relevant nuisance parameters are $(\gamma_0, \alpha_0^{\min})$ defined as unique solutions (γ,α) to

$$E(Y \mid W_2 = w_2) = E\{\gamma(W_1) \mid W_2 = w_2\}, \quad \eta_0^{\min}(w_1) = E\{\alpha(W_2) \mid W_1 = w_1\},$$

where $\eta_0^{\min}$ is the minimal Riesz representer satisfying $E\{m(W_1,\gamma)\} = E\{\eta_0(W_1)\gamma(W_1)\}$ for all γ in Γ from Lemma 3.1. Uniqueness is due to the assumption of exact identification, which amounts to completeness when $\Gamma = \mathbb{L}_2$. In Example 3.3, $w_1 = (d,x)$, $w_2 = (z,x)$, and $\eta_0(d,x) = -\partial_d \log f(d \mid x)$ where $f(d \mid x)$ is a conditional density. This abuse of notation allows us to state unified results. The estimating equation is

$$\theta_0 = E[m(W_1,\gamma_0) + \alpha_0^{\min}(W_2)\{Y - \gamma_0(W_1)\}].$$

A new insight of this work is that, for any mean square continuous functional, $n^{-1/2}$ Gaussian approximation is still possible if either γ_0 or $\alpha_0^{\min}$ is the solution to a mildly, rather than severely, ill posed inverse problem; the doubly robust formulation confers double robustness to ill posedness.

4 Algorithm

Our goal is general purpose learning and inference for the target parameter θ_0 in $\mathbb{R}$ that is a mean square continuous functional of γ_0 in Γ. Lemma 3.1 demonstrates that any such θ_0 has a unique minimal representer $\alpha_0^{\min}$ in Γ. In this section, we describe a meta algorithm to turn estimators $\hat{\gamma}$ of γ_0 and $\hat{\alpha}$ of $\alpha_0^{\min}$ into an estimator $\hat{\theta}$ of θ_0 such that $\hat{\theta}$ has a valid and practical confidence interval. Recall that $\hat{\gamma}$ may be any machine learning algorithm. To preserve this generality, we do not instantiate a choice of $\hat{\gamma}$; we treat it as a black box. In subsequent analysis, we will only require that $\hat{\gamma}$ converges to γ_0 in mean square error. This mean square rate is guaranteed by existing statistical learning theory.

The target estimator $\hat{\theta}$ as well as its confidence interval will depend on nuisance estimators $\hat{\gamma}$ and $\hat{\alpha}$. We refrain from instantiating the estimator $\hat{\alpha}$ for $\alpha_0^{\min}$. As we will see in subsequent analysis, the general theory only requires that $\hat{\alpha}$ converges to $\alpha_0^{\min}$ in mean square error. A recent literature provides $\hat{\alpha}$ estimators with fast rates inspired by the Dantzig selector [16], lasso [18, 40, 4], adversarial neural networks [17, 25], and kernel ridge regression [38].

Algorithm 4.1 (Debiased machine learning)**.** *Given a sample* (Y_i, W_i) $(i = 1, ..., n)$*, partition the sample into folds* (I_ℓ) $(\ell = 1, ..., L)$*. Denote by* I_ℓ^c *the complement of* I_ℓ*.*

1. *For each fold* ℓ*, estimate* $\hat{\gamma}_\ell$ *and* $\hat{\alpha}_\ell$ *from observations in* I_ℓ^c*.*
2. *Estimate* θ_0 *as* $\hat{\theta} = n^{-1} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} [m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_i)\{Y_i - \hat{\gamma}_\ell(W_i)\}]$*.*
3. *Estimate its* $(1 - a)100\%$ *confidence interval as* $\hat{\theta} \pm c_a \hat{\sigma} n^{-1/2}$*, where* c_a *is the* $1 - a/2$ *quantile of the standard Gaussian and* $\hat{\sigma}^2 = n^{-1} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} [m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_i)\{Y_i - \hat{\gamma}_\ell(W_i)\} - \hat{\theta}]^2$*.*

This meta algorithm can be seen as an extension of classic one step corrections [31] amenable to the use of modern machine learning, and it has been termed debiased machine learning [13]. It departs from targeted machine learning inference with a finite sample [42, 11] in a few ways. On the one hand, it avoids iteration and bootstrapping, thereby simplifying computation. On the other hand, it does not involve substitution, which would ensure that the estimator obeys additional meaningful constraints. See [18] for an algorithm that combines the two approaches.

5 Validity of confidence interval

We write this section at a high level of generality so it can be used by analysts working on a variety of problems. We assume a few simple and interpretable conditions and consider black box estimators $(\hat{\gamma}, \hat{\alpha})$. We prove by finite sample arguments that $\hat{\theta}$ defined by Algorithm 4.1 is consistent, and that its confidence interval is valid and semiparametrically efficient. Towards this end, define the oracle moment function

$$\psi_0(w) = \psi(w, \theta_0, \gamma_0, \alpha_0^{\min}), \quad \psi(w, \theta, \gamma, \alpha) = m(w, \gamma) + \alpha(w)\{y - \gamma(w)\} - \theta.$$

Its moments are $\sigma^2 = E\{\psi_0(W)^2\}$, $\kappa^3 = E\{|\psi_0(W)|^3\}$, and $\zeta^4 = E\{\psi_0(W)^4\}$. Write the Berry Esseen constant as $c^{BE} = 0.4748$ [37]. The result will be in terms of abstract mean square rates.

Definition 5.1 (Mean square error)**.** *Write the mean square error* $\mathcal{R}(\hat{\gamma}_\ell)$ *and the projected mean square error* $\mathcal{P}(\hat{\gamma}_\ell)$ *of* $\hat{\gamma}_\ell$ *trained on observations indexed by* I_ℓ^c *as*

$$\mathcal{R}(\hat{\gamma}_\ell) = E[\{\hat{\gamma}_\ell(W) - \gamma_0(W)\}^2 \mid I_\ell^c], \quad \mathcal{P}(\hat{\gamma}_\ell) = E([E\{\hat{\gamma}_\ell(W_1) - \gamma_0(W_1) \mid W_2, I_\ell^c\}]^2 \mid I_\ell^c).$$

Likewise define $\mathcal{R}(\hat{\alpha}_\ell)$ *and* $\mathcal{P}(\hat{\alpha}_\ell)$*.*

Statistical learning theory provides rates of this form, where I_ℓ^c is a training set and W is a test point. In the case of nonparametric regression, $\mathcal{R}(\hat{\gamma}_\ell)$ or $\mathcal{R}(\hat{\alpha}_\ell)$ typically has a fast rate between $n^{-1/2}$ and n^{-1}. In the case of nonparametric instrumental variable regression, $\mathcal{R}(\hat{\gamma}_\ell)$ and $\mathcal{R}(\hat{\alpha}_\ell)$ typically have rates slower than $n^{-1/2}$ due to ill posedness, but $\mathcal{P}(\hat{\gamma}_\ell)$ or $\mathcal{P}(\hat{\alpha}_\ell)$ may have a fast rate [10, 39, 20]. Our main result is a finite sample Gaussian approximation.

Theorem 5.1 (Finite sample Gaussian approximation)**.** *Suppose Assumption 3.1 holds,* $E[\{Y - \gamma_0(W)\}^2 \mid W] \leq \bar{\sigma}^2$*, and* $\|\alpha_0^{\min}\|_\infty \leq \bar{\alpha}$*. Then with probability* $1 - \epsilon$*,*

$$\sup_{z \in \mathbb{R}} \left| \text{pr} \left\{ \frac{n^{1/2}}{\sigma} (\hat{\theta} - \theta_0) \leq z \right\} - \Phi(z) \right| \leq c^{BE} \left(\frac{\kappa}{\sigma} \right)^3 n^{-1/2} + \frac{\Delta}{(2\pi)^{1/2}} + \epsilon,$$

where $\Phi(z)$ *is the standard Gaussian cumulative distribution function and*

$$\Delta = \frac{3L}{\epsilon \sigma} \left[(\bar{Q}^{1/2} + \bar{\alpha}) \{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} + \bar{\sigma} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} + \{n \mathcal{R}(\hat{\gamma}_\ell) \mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \right].$$

If in addition $\|\hat{\alpha}_\ell\|_\infty \leq \bar{\alpha}'$ *then the same result holds updating* Δ *to be*

$$\frac{4L}{\epsilon^{1/2} \sigma} \left[(\bar{Q}^{1/2} + \bar{\alpha} + \bar{\alpha}') \{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} + \bar{\sigma} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \right] + \frac{1}{\sigma} [\{n \mathcal{P}(\hat{\gamma}_\ell) \mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n \mathcal{R}(\hat{\gamma}_\ell) \mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}].$$

For local functionals, further suppose approximation error of size $\Delta_h = n^{1/2} \sigma_h^{-1} |\theta_0^h - \theta_0^{\lim}|$*. Then the same result holds replacing* $(\hat{\theta}, \theta_0, \Delta)$ *with* $(\hat{\theta}^h, \theta_0^{\lim}, \Delta + \Delta_h)$*.*

Theorem 5.1 is a finite sample Gaussian approximation for debiased machine learning with black box $(\hat{\gamma}_\ell, \hat{\alpha}_\ell)$. It degrades gracefully if the parameters $(\bar{Q}, \bar{\sigma}, \bar{\alpha}, \bar{\alpha}')$ diverge relative to n and the learning rates. Note that $\bar{\alpha}'$ is a bound on the chosen estimator $\hat{\alpha}_\ell$ that can be imposed by censoring extreme evaluations. Theorem 5.1 is a finite sample refinement of the asymptotic black box result in [14].

In the bound Δ, the expression $\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}$ allows a tradeoff: one of the learning rates may be slow, as long as the other is sufficiently fast to compensate. It is easily handled in the case of nonparametric regression, where $\mathcal{R}(\hat{\gamma}_\ell)$ or $\mathcal{R}(\hat{\alpha}_\ell)$ typically has a fast rate. However, the expression may diverge in the case of nonparametric instrumental variable regression, where both rates may be slow due to ill posedness.

The refined bound provides an alternative path to Gaussian approximation, replacing $\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}$ with the minimum of $\{n\mathcal{P}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}$ and $\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}$. Importantly, the projected mean square error $\mathcal{P}(\hat{\gamma}_\ell)$ can have a fast rate even when the mean square error $\mathcal{R}(\hat{\gamma}_\ell)$ has a slow rate because its definition sidesteps ill posedness. Moreover, the analyst only needs $\mathcal{P}(\hat{\gamma}_\ell)$ fast enough to compensate for the ill posedness encoded in $\mathcal{R}(\hat{\alpha}_\ell)$, or $\mathcal{P}(\hat{\alpha}_\ell)$ fast enough to compensate for the ill posedness encoded in $\mathcal{R}(\hat{\gamma}_\ell)$. This general and finite sample characterization of double robustness to ill posedness appears to be new. In independent work, [25] document an asymptotic special case of this result for a specific global functional and specific nuisance estimators; see Supplement 3.

By Theorem 5.1, the neighborhood of Gaussian approximation scales as $\sigma n^{-1/2}$. If σ is a constant, then the rate of convergence is $n^{-1/2}$, i.e. the parametric rate. If σ is a diverging sequence, then the rate of convergence degrades gracefully to nonparametric rates. A precise characterization of σ is possible, which we provide in Supplement 2 and summarize here. It turns out that global functionals have σ that is constant, while local functionals have $\sigma = \sigma_h$ that is a diverging sequence. We emphasize which quantities are diverging sequences for local functionals by indexing with the bandwidth h.

Theorem 5.2 (Characterization of key quantities)**.** *If noise has finite variance then $\bar{\sigma}^2 < \infty$. Suppose bounded moment and heteroscedasticity conditions defined in Supplement 2 hold. Then for global functionals $\kappa/\sigma \lesssim \sigma \asymp \bar{M} < \infty$; $\kappa, \zeta \lesssim \bar{M}^2 \leq \bar{Q} < \infty$; and $\bar{\alpha} < \infty$. Suppose bounded moment, heteroscedasticity, density, and derivative conditions defined in Supplement 2 hold. Then for local functionals $\kappa_h/\sigma_h \lesssim h^{-1/6}$, $\sigma_h \asymp \bar{M}_h \asymp h^{-1/2}$, $\kappa_h \lesssim h^{-2/3}$, $\zeta_h \lesssim h^{-3/4}$, $\bar{Q}_h \lesssim h^{-2}$, $\bar{\alpha}_h \lesssim h^{-1}$, and $\Delta_h \lesssim n^{1/2}h^{\mathsf{v}+1/2}$ where v is the order of differentiability defined in Supplement 2.*

For global functionals, $(\bar{Q}, \bar{\alpha})$ are finite constants that depend on the problem at hand. For example, for treatment effects a sufficient condition is that the propensity score is bounded away from zero and one. For derivatives, a sufficient condition is that Γ satisfies Sobolev conditions. For local functionals, we handle $(\bar{Q}_h, \bar{\alpha}_h)$ on a case by case basis. See Supplement 2 for interpretable and complete characterizations.

Observe that the finite sample Gaussian approximation in Theorem 5.1 is in terms of the true asymptotic variance σ^2. We now provide a guarantee for its estimator $\hat{\sigma}^2$.

Theorem 5.3 (Variance estimation)**.** *Suppose Assumption 3.1 holds, $E[\{Y - \gamma_0(W)\}^2 \mid W] \leq \bar{\sigma}^2$, and $\|\hat{\alpha}_\ell\|_\infty \leq \bar{\alpha}'$. Then with probability $1 - \epsilon'$, $|\hat{\sigma}^2 - \sigma^2| \leq \Delta' + 2(\Delta')^{1/2}\{(\Delta'')^{1/2} + \sigma\} + \Delta''$, where*

$$\Delta' = 4(\hat{\theta} - \theta_0)^2 + \frac{24L}{\epsilon'}\left[\{\bar{Q} + (\bar{\alpha}')^2\}\mathcal{R}(\hat{\gamma}_\ell)^q + \bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell)\right], \quad \Delta'' = \left(\frac{2}{\epsilon'}\right)^{1/2}\zeta^2 n^{-1/2}.$$

Theorem 5.3 is a finite sample variance estimation guarantee. It degrades gracefully if the parameters $(\bar{Q}, \bar{\sigma}, \bar{\alpha}')$ diverge relative to n and the learning rates. Theorems 5.1 and 5.3 immediately imply simple, interpretable conditions for validity of the confidence interval. We conclude by summarizing these conditions.

Corollary 5.1 (Confidence interval)**.** *Suppose Assumption 3.1 holds as well as the following regularity and learning rate conditions, as $n \to \infty$ and as $h \to 0$:*

$$E[\{Y - \gamma_0(W)\}^2 \mid W] \leq \bar{\sigma}^2, \quad \|\alpha_0^{\min}\|_\infty \leq \bar{\alpha}, \quad \|\hat{\alpha}_\ell\|_\infty \leq \bar{\alpha}', \quad \left\{(\kappa/\sigma)^3 + \zeta^2\right\} n^{-1/2} \to 0;$$

1. $\left(\bar{Q}^{1/2} + \bar{\alpha}/\sigma + \bar{\alpha}'\right)\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} = o_p(1)$;

2. $\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} = o_p(1)$;

3. $[\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{P}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}]/\sigma = o_p(1)$.

Then the estimator $\hat{\theta}$ in Algorithm 4.1 is consistent and asymptotically Gaussian, and the confidence interval in Algorithm 4.1 includes θ_0 with probability approaching the nominal level. Formally,

$$\hat{\theta} = \theta_0 + o_p(1), \quad \sigma^{-1}n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0,1), \quad \text{pr}\left\{\theta_0 \text{ in } \left(\hat{\theta} \pm c_a\hat{\sigma}n^{-1/2}\right)\right\} \to 1 - a.$$

For local functionals, if $\Delta_h \to 0$, then the same result holds replacing $(\hat{\theta}, \theta_0)$ with $(\hat{\theta}^h, \theta_0^{\text{lim}})$.

Acknowledgments and disclosure of funding

The National Science Foundation provided partial financial support via grants 1559172 and 1757140. Rahul Singh thanks the Jerry Hausman Dissertation Fellowship.

References

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[2] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

[3] Donald WK Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, pages 43–72, 1994.

[4] Vahe Avagyan and Stijn Vansteelandt. High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostatistics & Epidemiology*, pages 1–18, 2021.

[5] Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

[6] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. In *Advances in Economics and Econometrics*, page 245–295, 2013.

[7] Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika*, 102(1):77–94, 2014.

[8] Aurelien F Bibaut and Mark J van der Laan. Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters. *arXiv:1706.07408*, 2017.

[9] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[10] Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669, 2007.

[11] Weixin Cai and Mark van der Laan. Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (LASSO) estimator. *The International Journal of Biostatistics*, 16(2), 2020.

[12] Xiaohong Chen and Andres Santos. Overidentification in regular models. *Econometrica*, 86(5):1771–1817, 2018.

[13] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[14] Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv:1608.00033, Econometrica (to appear)*, 2016.

[15] Victor Chernozhukov, Jerry A Hausman, and Whitney K Newey. Demand analysis with many prices. Technical report, National Bureau of Economic Research, 2019.

[16] Victor Chernozhukov, Whitney Newey, and Rahul Singh. Debiased machine learning of global and local parameters using regularized Riesz representers. *arXiv:1802.08667, Econometrics Journal (to appear)*, 2018.

[17] Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of Riesz representers. *arXiv:2101.00009*, 2020.

[18] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *arXiv:1809.05224, Econometrica, (to appear)*, 2018.

[19] Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv:2004.03036*, 2020.

[20] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *arXiv:2006.07201*, 2020.

[21] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv:1901.09036*, 2019.

[22] Rafail Z Hasminskii and Ildar A Ibragimov. On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, 1979.

[23] Hidehiko Ichimura and Whitney K Newey. The influence function of semiparametric estimators. *arXiv:1508.01378, Quantitative Economics (to appear)*, 2021.

[24] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[25] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv:2103.14029*, 2021.

[26] Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562, 1987.

[27] Alexander R Luedtke and Mark J van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, 44(2):713, 2016.

[28] Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382, 1994.

[29] Whitney K Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2):1–21, 1994.

[30] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[31] Johann Pfanzagl. Lecture notes in statistics. *Contributions to a General Asymptotic Statistical Theory*, 13, 1982.

[32] Hongxiang Qiu, Alex Luedtke, and Marco Carone. Universal sieve-based strategies for efficient estimation using machine learning tools. *Bernoulli*, 27(4):2300–2336, 2021.

[33] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[34] Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, pages 931–954, 1988.

[35] Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

[36] Thomas A Severini and Gautam Tripathi. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, 170(2):491–498, 2012.

[37] Irina Shevtsova. On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv:1111.6554*, 2011.

[38] Rahul Singh. Debiased kernel methods. *arXiv:2102.11076*, 2021.

[39] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.

[40] Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv:1904.03737*, 2019.

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

[42] Mark van der Laan. Finite sample inference for targeted learning. *arXiv:1708.09502*, 2017.

[43] Mark J van der Laan and Sherri Rose. *Targeted Learning in Data Science*. Springer, 2018.

[44] Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[45] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[46] Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer Science & Business Media, 2011.

A Simulations

We present simulations for Example 3.1: heterogeneous treatment effect estimated by neural network. In addition, we present results for heterogeneous treatment effect estimated by random forest and lasso.

Recall that the localized functional is

$$\text{CATE}(v) = \lim_{h \to 0} E[\ell_{h,v}(V)\{\gamma_0(1, V, X) - \gamma_0(0, V, X)\}],$$

where $\ell_{h,v}(V) = (h\omega)^{-1} K\left\{(V - v)/h\right\}$ and $\omega = E[h^{-1} K\left\{(V - v)/h\right\}]$. V is an interpretable, low dimensional characteristic such as age, race, or gender. We implement the heterogeneous treatment effect design of [1], where $\text{CATE}(v) = v(1 + 2v)^2(v - 1)^2$ and v is a value in the interval $(-0.5, 0.5)$. A single observations consists of the tuple (Y_i, D_i, V_i, X_i) for outcome, treatment, covariate of interest, and other covariates. In this design, Y_i, D_i, V_i are in $\mathbb{R}$ and X_i is in $\mathbb{R}^3$.

A single observation is generated as follows. Draw the latent variables (ϵ_{ij}) $(j = 1, ..., 4)$ independently and identically from the uniform distribution $\mathcal{U}(-0.5, 0.5)$. Then set the covariates (V_i, X_i) according to

$$V_i = \epsilon_{i1}, \quad X_{i1} = 1 + 2V_i + \epsilon_{i2}, \quad X_{i2} = 1 + 2V_i + \epsilon_{i3}, \quad X_{i3} = (V_i - 1)^2 + \epsilon_{i4}.$$

Under the assumption of selection on observables, treatment assignment is as good as random conditional on (V_i, X_i). Draw the treatment D_i from the Bernoulli distribution with parameter $\Lambda\{1/2(V_i + X_{i1} + X_{i2} + X_{i3})\}$ where Λ is the logistic link function. Finally, calculate outcome Y_i as 0 if $D_i = 0$ and $V_i X_{i1} X_{i2} X_{i3} + \nu_i$ if $D_i = 1$, where the response noise ν_i is independently drawn from the Gaussian distribution $\mathcal{N}(0, 1/16)$. A random sample consists of $n = 100$ such observations (Y_i, D_i, V_i, X_i) $(i = 1, ..., n)$.

We implement different variations of Algorithm 4.1 with $L = 5$ folds. Across variations, we use a lasso estimator $\hat{\alpha}$ for the minimal Riesz representer $\alpha_0^{\min}$ [18]. We consider different estimators $\hat{\gamma}$ for the nonparametric regression γ_0: neural network, random forest, and lasso. We consider both low dimensional and high dimensional variations. In the low dimensional variation, the estimators $(\hat{\alpha}_\ell, \hat{\gamma}_\ell)$ use (D_i, V_i, X_i) $(i \text{ in } I_\ell^c)$ as well as their interactions. In the high dimensional variation, the estimators $(\hat{\alpha}_\ell, \hat{\gamma}_\ell)$ use fourth order polynomials of (D_i, V_i, X_i) $(i \text{ in } I_\ell^c)$.

Some tuning choices are necessary. We follow the default hyperparameter settings to tune the lasso Riesz representer and lasso regression from [18]. We implement the neural network with a single hidden layer of eight neurons and the random forest with 1000 trees as in [13]. Finally, to tune the bandwidth, we use the heuristic $h = c_h \hat{\sigma}_v n^{-0.2}$ [19], where $\hat{\sigma}_v^2$ is the sample variance of (V_i) $(i = 1, ..., n)$. The bandwidth hyperparameter c_h is chosen by the analyst. We evaluate robustness of coverage with respect to hyperparameter values $c_h = 0.25, 0.50, 1.00$ below. Empirically, we find that $c_h = 0.25$ and $c_h = 0.50$ work well.

For each choice of nonparametric regression estimator $\hat{\gamma}$, whether neural network, random forest, or lasso, and for each choice of specification, whether low or high dimensional, we report a coverage table summarizing 500 simulations. The initial columns denote the grid value v, the corresponding heterogeneous treatment effect $\text{CATE}(v)$, and the bandwidth hyperparamter value c_h. The subsequent columns calculate the average point estimate and the average standard error across the 500 simulations for this choice of $\{v, \text{CATE}(v), c_h\}$. The final columns report what percentage of the 500 confidence intervals contain the true value $\text{CATE}(v)$ compared to the theoretical benchmarks of 80% and 95%, respectively.

For the low dimensional regime, Tables 1, 2, and 3 summarize results for neural network, random forest, and lasso, respectively. With bandwidth hyperparameter values $c_h = 0.25$ and $c_h = 0.50$, coverage is close to the nominal level across $\hat{\gamma}$ estimators and across grid values $v = -0.25, 0.00, 0.25$. Neural network and random forest have comparable performance. Lasso has higher bias and compensates with higher variance for the grid value $v = 0.25$.

Table 1: Low dimensional coverage simulation with neural network

v	CATE(v)	Tuning	Ave. Est.	Ave. S.E.	80% Cov.	95% Cov.
-0.25	-0.10	0.25	-0.10	0.05	83%	94%
-0.25	-0.10	0.50	-0.10	0.04	85%	95%
-0.25	-0.10	1.00	-0.08	0.03	71%	88%
0.00	0.00	0.25	0.00	0.04	78%	95%
0.00	0.00	0.50	0.00	0.03	78%	94%
0.00	0.00	1.00	0.02	0.02	62%	85%
0.25	0.32	0.25	0.31	0.12	85%	92%
0.25	0.32	0.50	0.30	0.09	85%	93%
0.25	0.32	1.00	0.28	0.06	76%	88%

Ave., average; Est., estimate; S.E., standard error; Cov., coverage. The largest standard error for the results in column 6 is 2%. The largest standard error for the results in column 7 is 2%.

Table 2: Low dimensional coverage simulation with random forest

v	CATE(v)	Tuning	Ave. Est.	Ave. S.E.	80% Cov.	95% Cov.
-0.25	-0.10	0.25	-0.10	0.05	86%	93%
-0.25	-0.10	0.50	-0.09	0.03	83%	94%
-0.25	-0.10	1.00	-0.08	0.02	60%	79%
0.00	0.00	0.25	0.00	0.02	70%	91%
0.00	0.00	0.50	0.01	0.02	72%	91%
0.00	0.00	1.00	0.02	0.02	48%	75%
0.25	0.32	0.25	0.30	0.12	83%	91%
0.25	0.32	0.50	0.29	0.08	82%	91%
0.25	0.32	1.00	0.28	0.06	71%	86%

Ave., average; Est., estimate; S.E., standard error; Cov., coverage. The largest standard error for the results in column 6 is 2%. The largest standard error for the results in column 7 is 2%.

Table 3: Low dimensional coverage simulation with lasso

v	CATE(v)	Tuning	Ave. Est.	Ave. S.E.	80% Cov.	95% Cov.
-0.25	-0.10	0.25	-0.08	0.08	81%	95%
-0.25	-0.10	0.50	-0.08	0.05	81%	95%
-0.25	-0.10	1.00	-0.06	0.04	63%	88%
0.00	0.00	0.25	0.00	0.06	79%	94%
0.00	0.00	0.50	0.01	0.04	83%	96%
0.00	0.00	1.00	0.02	0.03	73%	92%
0.25	0.32	0.25	0.30	0.11	86%	94%
0.25	0.32	0.50	0.29	0.08	85%	95%
0.25	0.32	1.00	0.28	0.06	71%	89%

Ave., average; Est., estimate; S.E., standard error; Cov., coverage. The largest standard error for the results in column 6 is 2%. The largest standard error for the results in column 7 is 1%.

For the high dimensional regime, Tables 4, 5, and 6 summarize results for neural network, random forest, and lasso, respectively. With bandwidth hyperparameter values $c_h = 0.25$ and $c_h = 0.50$, coverage is close to the nominal level across $\hat{\gamma}$ estimators and for grid values $v = -0.25$ and $v = 0.00$. Across $\hat{\gamma}$ estimators, the grid value $v = 0.25$ is more challenging. Compared to the low dimensional regime, each estimator in the high dimensional regime has higher bias and compensates with higher variance for the grid value $v = 0.25$.

Table 4: High dimensional coverage simulation with neural network

v	CATE(v)	Tuning	Ave. Est.	Ave. S.E.	80% Cov.	95% Cov.
-0.25	-0.10	0.25	-0.09	0.04	84%	91%
-0.25	-0.10	0.50	-0.09	0.03	78%	91%
-0.25	-0.10	1.00	-0.07	0.02	49%	74%
0.00	0.00	0.25	0.00	0.03	75%	95%
0.00	0.00	0.50	0.01	0.02	74%	91%
0.00	0.00	1.00	0.04	0.03	52%	75%
0.25	0.32	0.25	0.39	0.20	90%	97%
0.25	0.32	0.50	0.39	0.15	88%	97%
0.25	0.32	1.00	0.38	0.13	81%	95%

Ave., average; Est., estimate; S.E., standard error; Cov., coverage. The largest standard error for the results in column 6 is 2%. The largest standard error for the results in column 7 is 2%.

Table 5: High dimensional coverage simulation with random forest

v	CATE(v)	Tuning	Ave. Est.	Ave. S.E.	80% Cov.	95% Cov.
-0.25	-0.10	0.25	-0.09	0.05	81%	91%
-0.25	-0.10	0.50	-0.09	0.03	78%	91%
-0.25	-0.10	1.00	-0.07	0.02	53%	75%
0.00	0.00	0.25	0.00	0.02	76%	94%
0.00	0.00	0.50	0.01	0.02	74%	92%
0.00	0.00	1.00	0.04	0.03	44%	71%
0.25	0.32	0.25	0.37	0.18	91%	96%
0.25	0.32	0.50	0.40	0.16	88%	97%
0.25	0.32	1.00	0.39	0.14	81%	95%

Ave., average; Est., estimate; S.E., standard error; Cov., coverage. The largest standard error for the results in column 6 is 2%. The largest standard error for the results in column 7 is 2%.

Table 6: High dimensional coverage simulation with lasso

v	CATE(v)	Tuning	Ave. Est.	Ave. S.E.	80% Cov.	95% Cov.
-0.25	-0.10	0.25	-0.08	0.06	75%	90%
-0.25	-0.10	0.50	-0.07	0.04	74%	90%
-0.25	-0.10	1.00	-0.06	0.03	50%	74%
0.00	0.00	0.25	0.01	0.05	78%	96%
0.00	0.00	0.50	0.02	0.04	80%	97%
0.00	0.00	1.00	0.04	0.04	59%	84%
0.25	0.32	0.25	0.41	0.20	89%	96%
0.25	0.32	0.50	0.45	0.18	87%	97%
0.25	0.32	1.00	0.43	0.17	82%	95%

Ave., average; Est., estimate; S.E., standard error; Cov., coverage. The largest standard error for the results in column 6 is 2%. The largest standard error for the results in column 7 is 2%.

B Characterization of key parameters

B.1 Additional example

We present an additional example useful in commercial applications where an analyst has access to data on household spending behavior. See [15] for a motivating economic model and further interpretation.

Example B.1 (Heterogeneous average derivative estimated by lasso)**.** *Let Y be share of household expenditure on some good. Let $W = (D, V, X)$ concatenate log price of the good D, covariate of*

interest V such as household size, and other covariates X such as log prices of other goods and log total household expenditure. Let $\gamma_0(d, v, x) = E(Y \mid D = d, V = v, X = x)$ be a function estimated by lasso. The heterogeneous average derivative is

$$\text{DERIV}(v) = E\left\{\frac{\partial}{\partial d}\gamma_0(D, V, X) \mid V = v\right\} = \lim_{h\to 0} E\left\{\ell_h(V)\frac{\partial}{\partial d}\gamma_0(D, V, X)\right\},$$

where $\ell_h(V) = (h\omega)^{-1}K\{(V - v)/h\}$, $\omega = E[h^{-1}K\{(V - v)/h\}]$, and K is a bounded and symmetric kernel that integrates to one.

The heterogeneous average derivative is defined with respect to some interpretable, low dimensional characteristic V such as household size [1]. The same functional without the localization ℓ_h is the classic average derivative.

B.2 Riesz representers

We begin by deriving explicit expressions for Riesz representer α_0. Recall that the unique minimal Riesz representer $\alpha_0^{\min}$ is the projection of any valid Riesz representer α_0 onto Γ. The explicit expressions for α_0 help to articulate simple sufficient conditions for existence of $\alpha_0^{\min}$.

Lemma B.1. *In Examples 3.1, 3.2, 3.3, and B.1, the minimal representer $\alpha_0^{\min}(w)$ can be obtained by projecting the following Riesz representer $\alpha_0(w)$ onto Γ.*

1. *Example 3.1. Denote the propensity score $\pi_0(v, x) = \text{pr}(D = 1 \mid V = v, X = x)$. Then*

$$\alpha_0(d, v, x) = \ell_h(v)\left\{\frac{d}{\pi_0(v, x)} - \frac{1 - d}{1 - \pi_0(v, x)}\right\}.$$

   Hence the minimal Riesz representer $\alpha_0^{\min}$ exists if $\pi_0(v, x)$ is bounded away from zero and one.

2. *Example 3.2. Denote the Riesz representer for the first term by α_0^+ and the Riesz representer for the second term by α_0^-. Then*

$$\alpha_0^+(d, x) = \ell_h^+(d), \quad \alpha_0^-(d, x) = \ell_h^-(d).$$

   Hence the minimal Riesz representer $\alpha_0^{\min}$ exists.

3. *Example 3.3. Denote the density $f(d, x)$. If $f(d \mid x)$ vanishes for each d in the boundary of the support of D given $X = x$ almost everywhere then*

$$\eta_0(d, x) = -\partial_d \log f(d \mid x).$$

   Hence the minimal representer $\eta_0^{\min}$ exists if $-\partial_d \log f(d \mid x)$ is bounded above. Subsequently, $\alpha_0^{\min}$ is the solution α to

$$\eta_0^{\min}(d, x) = E\{\alpha(X, Z) \mid D = d, X = x\}.$$

4. *Example B.1. Denote the density $f(d, v, x)$. If $f(d \mid v, x)$ vanishes for each d in the boundary of the support of D given $(V, X) = (v, x)$ almost everywhere then*

$$\alpha_0(d, v, x) = -\ell_h(v)\partial_d \log f(d \mid v, x).$$

Next, we characterize key quantities $(\bar{Q}, \bar{\sigma}, \bar{\alpha}, \bar{\alpha}')$ and moments (σ, κ, ζ) that appear in Theorems 5.1 and 5.3. Recall that

$$E[\{Y - \gamma_0(W)\}^2 \mid W] \leq \bar{\sigma}^2, \quad \|\hat{\alpha}_\ell\|_\infty \leq \bar{\alpha}',$$

so $\bar{\sigma}$ is a constant if noise has bounded variance and $\bar{\alpha}'$ can be imposed by trimming the algorithm $\hat{\alpha}_\ell$. We therefore focus on $(\bar{Q}, \bar{\alpha})$ and (σ, κ, ζ).

We consider two cases: global functionals, with weighting ℓ_h that has a fixed bandwidth; and local functionals, with weighting ℓ_h that depends on some vanishing bandwidth $h \to 0$. To lighten notation, let $\|X\|_{\text{pr},q} = \{E(|X|^q)\}^{1/q}$.

B.3 Global functionals

Lemma B.2 (Oracle moments for global functionals)**.** *Suppose the weighting is bounded, namely $\ell_h \leq C < \infty$, and $\Gamma \subset \mathbb{L}_2$. Suppose $\alpha_0^{\min}$ exists. Further suppose there exist $(c, \tilde{c}, \bar{c})$ bounded away from zero and above such that the following conditions hold.*

1. *Control of representer moments. For $q = 3, 4$*
$$\|\alpha_0^{\min}\|_{\text{pr},q} \leq c \left(\|\alpha_0^{\min}\|_{\text{pr},2}^2 \vee 1 \right).$$
2. *Bounded moments. For $q = 2, 3, 4$,*
$$U_1 = m(W, \gamma_0) - E\{m(W, \gamma_0)\}, \quad \|U_1\|_{\text{pr},q} \leq \bar{c}.$$
3. *Bounded heteroscedasticity. For $q = 2, 3, 4$,*
$$U_2 = Y - \gamma_0(W), \quad \tilde{c} \leq \|U_2 \mid W\|_{\text{pr},q} \leq \bar{c}.$$

Then
$$\tilde{c}\bar{M} \leq \sigma \leq \bar{c}\sqrt{1 + \bar{M}^2}, \quad \kappa, \zeta \leq \bar{c}\{1 + c(\bar{M}^2 \vee 1)\}.$$
In summary,
$$\frac{\kappa}{\sigma} \lesssim \sigma \asymp \bar{M} < \infty, \quad \kappa, \zeta \lesssim \bar{M}^2 < \infty.$$

Clearly Assumption 3.1 depends on the functional of interest. Towards a characterization of $\bar{Q}$ and q in Examples 3.3 and B.1, we prove the following technical lemma.

Lemma B.3 (A weak reverse Poincare inequality)**.** *Assume that $f(d \mid x)$ vanishes for each d in the boundary of the support of D given $X = x$ almost everywhere. Next assume the following restrictions on $\Gamma \subset \mathbb{L}_2$.*

1. *Each γ in Γ is twice continuously differentiable.*
2. *For each γ in Γ, $\|k_\gamma\|_{\text{pr},2} < \infty$ where*
$$k_\gamma(d, x) = \{-\partial_d \log f(d \mid x)\}\{\partial_d \gamma(d, x)\} - \partial_d^2 \gamma(d, x).$$

Then
$$E[\{\partial_d \gamma(D, X)\}^2] \leq \|k_\gamma\|_{\text{pr},2} [E\{\gamma(D, X)^2\}]^{1/2}.$$
Furthermore, $\sup_{\gamma \in \Gamma} \|k_\gamma\|_{\text{pr},2} < \infty$ if either of the following conditions hold.

1. *$\|\partial_d \log f(D \mid X)\|_{\text{pr},2} < \infty$ and for all γ in Γ, $\|\partial_d \gamma\|_\infty < \infty$ and $\|\partial_d^2 \gamma\|_{\text{pr},2} < \infty$;*
2. *$-\partial_d \log f(d \mid x)$ is bounded above and for all γ in Γ, $\|\partial_d \gamma\|_{\text{pr},2}$ and $\|\partial_d^2 \gamma\|_{\text{pr},2} < \infty$.*

With this technical lemma, we return to the characterization of $\bar{Q}$ and q across examples.

Lemma B.4 (Mean square continuity for global functionals)**.** *Suppose the weighting is bounded, namely $\ell_h \leq C < \infty$, and $\Gamma \subset \mathbb{L}_2$. The following conditions are sufficient for $\bar{Q} < \infty$ with $q = 1$ in Examples 3.1 and 3.2, and for $\bar{Q} < \infty$ with $q = 1/2$ in Examples 3.3 and B.1.*

1. *Example 3.1. $\pi_0(v, x)$ is bounded away from zero and one.*
2. *Example 3.2. The bandwidth is fixed.*
3. *Example 3.3. $f(d \mid x)$ vanishes for each d in the boundary of the support of D given $X = x$ almost everywhere. $-\partial_d \log f(d \mid x)$ is a bounded above, and Γ consists of functions γ that are twice continuously differentiable in the first argument and that satisfy $E[\{\partial_d \gamma(D, X)\}^2] < \infty$ and $E[\{\partial_d^2 \gamma(D, X)\}^2] < \infty$.*
4. *Example B.1. $f(d \mid v, x)$ vanishes for each d in the boundary of the support of D given $(V, X) = (v, x)$ almost everywhere. $-\partial_d \log f(d \mid v, x)$ is a bounded above, and Γ consists of functions γ that are twice continuously differentiable in the first argument and that satisfy $E[\{\partial_d \gamma(D, V, X)\}^2] < \infty$ and $E[\{\partial_d^2 \gamma(D, V, X)\}^2] < \infty$.*

Therefore a Sobolev type property with respect to the first argument is a sufficient condition in Examples 3.3 and B.1.

Next we examine the assumption of $\|\alpha_0^{\min}\|_\infty \leq \bar{\alpha}$, which depends on the functional of interest.

Lemma B.5 (Bounded Riesz representer for global functionals)**.** *The following conditions are sufficient for $\bar{\alpha} < \infty$ in Examples 3.1, 3.2, 3.3, and B.1 with $\ell_h \leq C < \infty$ and $\Gamma = \mathbb{L}_2$.*

1. *Example 3.1. $\pi_0(v, x)$ is bounded away from zero and one.*
2. *Example 3.2. The bandwidth is fixed.*
3. *Example 3.3. $\alpha_0^{\min}$ that solves $-\partial_d \log f(d \mid x) = E\{\alpha(X, Z) \mid D = d, X = x\}$ is bounded above.*
4. *Example B.1. $-\partial_d \log f(d \mid v, x)$ is bounded above.*

B.4 Local functionals

Given a local functional

$$\theta_0^h = E\{m_h(W, \gamma_0)\} = E\{\ell_h(V) m(W, \gamma_0)\},$$

we will now refer to the Riesz representer of m_h by α_0^h and the Riesz representer of m by α_0 for this subsection. The latter objects correspond to the global setting where the weighting is bounded. To lighten notation, we write $\ell = \ell_h$.

Lemma B.6 (Oracle moments for local functionals)**.** *Suppose $\alpha_0^{\min}$ exists and $\Gamma = \mathbb{L}_2$. Further suppose there exist $(\tilde{\alpha}, \check{\alpha}, \tilde{c}, \bar{c}, \tilde{f}, \bar{f}, \bar{f}', h_0)$ bounded away from zero and above such that the following conditions hold.*

1. *Control of representer absolute value:*

$$0 < \tilde{\alpha} \leq \alpha_0(w) \leq \check{\alpha}.$$

2. *Valid neighborhood. There exists $N_{h_0}(v) = (v' : |v' - v| \leq h) \subset \mathcal{V}$.*
3. *Bounded moments. For all $h \leq h_0$ and for $q = 2, 3, 4$,*

$$U_1 = m_h(W, \gamma_0) - E\{m_h(W, \gamma_0)\}, \quad \|U_1\|_{\text{pr},q} \leq \bar{c} \|\ell\|_{\text{pr},q}.$$

4. *Bounded heteroscedasticity. For $q = 2, 3, 4$,*

$$U_2 = Y - \gamma_0(W), \quad \tilde{c} \leq \|U_2 \mid W\|_{\text{pr},q} \leq \bar{c}.$$

5. *Bounded density. The density f_V obeys, for all v' in $N_{h_0}(v)$,*

$$0 < \tilde{f} \leq f_V(v') \leq \bar{f}, \quad |\partial f_V(v')| \leq \bar{f}'.$$

Then finite sample bounds in the proof hold. In summary,

$$\frac{\kappa_h}{\sigma_h} \lesssim h^{-1/6} \lesssim \sigma_h \asymp \bar{M}_h \asymp h^{-1/2} \to \infty, \quad \kappa_h \lesssim h^{-2/3} \to \infty, \quad \zeta_h \lesssim h^{-3/4} \to \infty.$$

The conditions of Lemma B.5 suffice for $0 < \tilde{\alpha} \leq \alpha_0(w) \leq \check{\alpha}$ in Examples 3.1 and 3.2.

As before, Assumption 3.1 depends on the functional of interest.

Lemma B.7 (Mean square continuity for local functionals)**.** *Suppose $\alpha_0^{\min}$ exists and $\Gamma \subset \mathbb{L}_2$. Further suppose there exist $(\tilde{f}, \bar{f}, \bar{f}', h_0)$ bounded away from zero and above such that the following conditions hold.*

1. *Valid neighborhood. There exists $N_{h_0}(v) = (v' : |v' - v| \leq h) \subset \mathcal{V}$.*
2. *Bounded density. The density f_V obeys, for all v' in $N_{h_0}(v)$,*

$$0 < \tilde{f} \leq f_V(v') \leq \bar{f}, \quad |\partial f_V(v')| \leq \bar{f}'.$$

3. *The conditions of Lemma B.4 hold.*

Then the finite sample bound in the proof holds for Examples 3.1 and 3.2. In summary,

$$\bar{Q}_h \lesssim h^{-2} \to \infty.$$

As before, the assumption of $\|\alpha_0^{\min,h}\|_\infty \leq \bar{\alpha}$ depends on the functional of interest.

Lemma B.8 (Bounded Riesz representer for local functionals)**.** *Suppose $\alpha_0^{\min}$ exists and $\Gamma \subset \mathbb{L}_2$. Further suppose there exist $(\tilde{f}, \bar{f}, \bar{f}', h_0, \bar{K})$ bounded away from zero and above such that the following conditions hold.*

1. *Valid neighborhood. There exists $N_{h_0}(v) = (v' : |v' - v| \leq h) \subset \mathcal{V}$.*
2. *Bounded density. The density f_V obeys, for all v' in $N_{h_0}(v)$,*

$$0 < \tilde{f} \leq f_V(v') \leq \bar{f}, \quad |\partial f_V(v')| \leq \bar{f}'.$$

3. *Bounded kernel. $|K(u)| \leq \bar{K}$.*
4. *The conditions of Lemma B.5 hold, allowing bandwidth to vanish.*

Then the finite sample bound in the proof holds. In summary,

$$\bar{\alpha}_h \lesssim h^{-1} \to \infty.$$

The main results are in terms of abstract mean square rates $\mathcal{R}(\hat{\alpha}_\ell^h)$ and $\mathcal{P}(\hat{\alpha}_\ell^h)$ for the local Riesz representer $\alpha_0^{\min,h}$ of the functional m_h. A growing literature proposes machine learning estimators $\hat{\alpha}$ with rates $\mathcal{R}(\hat{\alpha}_\ell)$ and $\mathcal{P}(\hat{\alpha}_\ell)$ for the global Riesz representer $\alpha_0^{\min}$ of the functional m.

A natural choice of estimator $\hat{\alpha}^h$ for $\alpha_0^{\min,h}$ is the localization ℓ_h times an estimator $\hat{\alpha}$ for $\alpha_0^{\min}$. We prove that this choice allows an analyst to translate global Riesz representer rates into local Riesz representer rates under mild regularity conditions. In Supplement 1, we confirm that this choice performs well in simulations.

Lemma B.9 (Translating global rates to local rates)**.** *Suppose the conditions of Lemma B.8 hold with $\Gamma = \mathbb{L}_2$. Then*

$$\mathcal{R}(\hat{\alpha}_\ell^h) \lesssim h^{-2}\mathcal{R}(\hat{\alpha}_\ell), \quad \mathcal{P}(\hat{\alpha}_\ell^h) \lesssim h^{-2}\mathcal{P}(\hat{\alpha}_\ell).$$

B.5 Approximation error

Finally, we characterize the finite sample approximation error $\Delta_h = n^{1/2}\sigma^{-1}|\theta_0^h - \theta_0^{\lim}|$ where

$$\theta_0^{\lim} = \lim_{h \to 0} \theta_0^h, \quad \theta_0^h = E\{m_h(W, \gamma_0)\} = E\{\ell_h(V)m(W, \gamma_0)\}.$$

Δ_h is bias from using a semiparametric sequence to approximate a nonparametric quantity.

We define $m(v) = E[m(W, \gamma_0) \mid V = v]$ to lighten notation.

Lemma B.10 (Approximation error from localization [16])**.** *Suppose there exist constants $(h_0, K, \mathsf{v}, \bar{g}_\mathsf{v}, \bar{f}_\mathsf{v}, \tilde{f}, \bar{g})$ bounded away from zero and above such that the following conditions hold.*

1. *Valid neighborhood. There exists $N_{h_0}(v) = (v' : |v' - v| \leq h) \subset \mathcal{V}$.*
2. *Differentiability. On $N_{h_0}(v)$, $m(v')$ and $f_V(v')$ are differentiable to the integer order* sm.
3. *Bounded derivatives. Let $\mathsf{v} = \mathsf{sm} \wedge \mathsf{o}$ where o is the order of the kernel K. Let ∂_d^v denote the v order derivative $\partial^\mathsf{v}/(\partial d)^\mathsf{v}$. Assume*

$$\sup_{v' \in N_{h_0}(v)} \|\partial_v^\mathsf{v}(m(v')f_V(v'))\|_{op} \leq \bar{g}_\mathsf{v}, \quad \sup_{v' \in N_{h_0}(v)} \|\partial_v^\mathsf{v} f_V(v')\|_{op} \leq \bar{f}_\mathsf{v}, \quad \inf_{v' \in N_{h_0}(v)} f_V(v') \geq \tilde{f}.$$

4. *Bounded conditional formula. $m(v)f_V(v) \leq \bar{g}$.*

Then there exist constants (C, h_1) depending only on $(h_0, K, \mathsf{v}, \bar{g}_\mathsf{v}, \bar{f}_\mathsf{v}, \tilde{f}, \bar{g})$ such that for all h_1 in (h, h_0), $|\theta_0^h - \theta_0^{\lim}| \leq Ch^\mathsf{v}$. In summary,

$$\Delta_h \lesssim n^{1/2} h^{\mathsf{v}+1/2}.$$

To summarize the characterizations in this Supplement, we provide a corollary for local functionals. Let $\mathcal{R}(\hat{\alpha}_\ell)$ and $\mathcal{P}(\hat{\alpha}_\ell)$ be defined as in Lemma B.9.

Corollary B.1 (Confidence interval for local functionals)**.** *Suppose the conditions of Corollary 5.1 and Lemmas B.6, B.7, B.8, B.9, and B.10 hold. As $n \to \infty$ and $h \to 0$, suppose the regularity condition on moments $n^{-1/2}h^{-3/2} \to 0$ as well as the following learning rate conditions:*

1. $\left(h^{-1} + \bar{\alpha}'\right) \{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} = o_p(1)$;
2. $\bar{\sigma} h^{-1} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} = o_p(1)$;
3. $h^{-1/2}[\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{P}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}] = o_p(1)$.

Finally suppose the approximation error condition $\Delta_h \to 0$. Then the estimator $\hat{\theta}^h$ in Algorithm 4.1 is consistent and asymptotically Gaussian, and the confidence interval in Algorithm 4.1 includes $\theta_0^{\lim}$ with probability approaching the nominal level. Formally,

$$\hat{\theta}^h = \theta_0^{\lim} + o_p(1), \quad \sigma_h^{-1} n^{1/2} (\hat{\theta}^h - \theta_0^{\lim}) \rightsquigarrow \mathcal{N}(0,1), \quad \text{pr}\left\{\theta_0^{\lim} \text{ in } \left(\hat{\theta}^h \pm c_a \hat{\sigma} n^{-1/2}\right)\right\} \to 1 - a.$$

C Discussion

In independent work, [25, Theorem 9] present an asymptotic Gaussian approximation result for a particular global functional: average treatment effect identified by negative controls. This functional fits within our framework because it is a mean square continuous functional of a nonparametric instrumental variable regression. To verify mean square continuity, see Example 3.1 in Lemma B.4.

In their analysis, the authors write the sufficient condition

$$\min(\tau_{\gamma,n}, \tau_{\alpha,n}) \iota_{\gamma,n} \iota_{\alpha,n} = o(n^{-1/2}), \quad \tau_{\gamma,n} = \sup_{\gamma \in \mathcal{G}_n} \frac{\{\mathcal{R}(\gamma)\}^{1/2}}{\{\mathcal{P}(\gamma)\}^{1/2}}, \quad \tau_{\alpha,n} = \sup_{\alpha \in \mathcal{A}_n} \frac{\{\mathcal{R}(\alpha)\}^{1/2}}{\{\mathcal{P}(\alpha)\}^{1/2}}$$

where $(\tau_{\gamma,n}, \tau_{\alpha,n})$ are ratio measures of ill posedness and $(\iota_{\gamma,n}, \iota_{\alpha,n})$ are critical radii for the sequence of function classes $(\mathcal{G}_n, \mathcal{A}_n)$ used in adversarial estimation procedures for $(\hat{\gamma}, \hat{\alpha})$. In particular $(\iota_{\gamma,n}, \iota_{\alpha,n})$ appear in the authors' bounds for $\{\mathcal{P}(\hat{\gamma})\}^{1/2}$ and $\{\mathcal{P}(\hat{\alpha})\}^{1/2}$, respectively.

For comparison, our analogous condition in Theorem 5.1 is that

$$\min\left[\{\mathcal{P}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}, \{\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}\right] = o(\sigma n^{-1/2}).$$

By contrast, our result is (i) for the entire class of mean square continuous functionals, (ii) for black box estimators $(\hat{\gamma}, \hat{\alpha})$, and (iii) finite sample, so it also handles local functionals in which σ diverges. Critical radius arguments are one way to prove mean square rates for certain machine learning estimators, but not the only way. This distinction is important, since many existing statistical learning theory rates, whether for neural networks as in Example 3.1, random forest as in Example 3.2, kernel ridge regression as in Example 3.3, or lasso as in Example B.1, are not in terms of a critical radius.

D Proof of main result

D.1 Gateaux differentiation

Recall the notation

$$\psi_0(w) = \psi(w, \theta_0, \gamma_0, \alpha_0^{\min}), \quad \psi(w, \theta, \gamma, \alpha) = m(w, \gamma) + \alpha(w)\{y - \gamma(w)\} - \theta,$$

where $\gamma \mapsto m(w, \gamma)$ is linear. For readability, we introduce the following notation for Gateaux differentiation.

Definition D.1 (Gateaux derivative)**.** *Let $u(w), v(w)$ be functions and let τ, ζ in $\mathbb{R}$ be scalars. The Gateaux derivative of $\psi(w, \theta, \gamma, \alpha)$ with respect to its argument γ in the direction u is*

$$\{\partial_\gamma \psi(w, \theta, \gamma, \alpha)\}(u) = \frac{\partial}{\partial \tau} \psi(w, \theta, \gamma + \tau u, \alpha)\bigg|_{\tau=0}.$$

The cross derivative of $\psi(w, \theta, \gamma, \alpha)$ with respect to its arguments (γ, α) in the directions (u, v) is

$$\{\partial^2_{\gamma,\alpha} \psi(w, \theta, \gamma, \alpha)\}(u, v) = \frac{\partial^2}{\partial \tau \partial \zeta} \psi(w, \theta, \gamma + \tau u, \alpha + \zeta v)\bigg|_{\tau=0,\zeta=0}.$$

Proposition D.1 (Calculation of derivatives)**.**

$$\begin{aligned}
\{\partial_\gamma \psi(w, \theta, \gamma, \alpha)\}(u) &= m(w, u) - \alpha(w)u(w); \\
\{\partial_\alpha \psi(w, \theta, \gamma, \alpha)\}(v) &= v(w)\{y - \gamma(w)\}; \\
\{\partial^2_{\gamma,\alpha} \psi(w, \theta, \gamma, \alpha)\}(u, v) &= -v(w)u(w).
\end{aligned}$$

Proof. For the first result, write

$$\psi(w, \theta, \gamma + \tau u, \alpha) = m(w, \gamma) + \tau m(w, u) + \alpha(w)\{y - \gamma(w) - \tau u(w)\} - \theta.$$

For the second result, write

$$\psi(w, \theta, \gamma, \alpha + \zeta v) = m(w, \gamma) + \alpha(w)\{y - \gamma(w)\} + \zeta v(w)\{y - \gamma(w)\} - \theta.$$

For the final result, write

$$\begin{aligned}
&\psi(w, \theta, \gamma + \tau u, \alpha + \zeta v) \\
&= m(w, \gamma) + \tau m(w, u) + \alpha(w)\{y - \gamma(w) - \tau u(w)\} + \zeta v(w)\{y - \gamma(w) - \tau u(w)\} - \theta.
\end{aligned}$$

Finally, take scalar derivatives with respect to (τ, ζ). □

By using the doubly robust moment function, we have the following helpful property.

Proposition D.2 (Mean zero derivatives)**.** *For any (u, v),*

$$E\{\partial_\gamma \psi_0(W)\}(u) = 0, \quad E\{\partial_\alpha \psi_0(W)\}(v) = 0.$$

Proof. We appeal to Proposition D.1. For the first result, write

$$E\{\partial_\gamma \psi_0(W)\}(u) = E\{m(W, u) - \alpha_0^{\min}(W)u(W)\}.$$

In the case of nonparametric regression or projection, appeal to the definition of minimal Riesz representer $\alpha_0^{\min}$. In the case of nonparametric instrumental variable regression,

$$\begin{aligned}
E\{m(W_1, u) - \alpha_0^{\min}(W_2)u(W_1)\} &= E[\{\eta_0^{\min}(W_1) - \alpha_0^{\min}(W_2)\}u(W_1)] \\
&= E([\eta_0^{\min}(W_1) - E\{\alpha_0^{\min}(W_2) \mid W_1\}]u(W_1)) \\
&= 0.
\end{aligned}$$

For the second result, write

$$E\{\partial_\alpha \psi_0(W)\}(v) = E[v(W)\{Y - \gamma_0(W)\}].$$

In the case of nonparametric regression, $\gamma_0(w) = E(Y \mid W = w)$ and we appeal to law of iterated expectations. In the case of nonparametric projection, the desired result holds by orthogonality of the projection residual. In the case of nonparametric instrumental variable regression,

$$E[v(W_2)\{Y - \gamma_0(W_1)\}] = E(v(W_2)[E(Y \mid W_2) - E\{\gamma_0(W_1) \mid W_2\}]) = 0.$$

□

D.2 Taylor expansion

Train $(\hat{\gamma}_\ell, \hat{\alpha}_\ell)$ on observations in I_ℓ^c. Let $n_\ell = |I_\ell| = n/L$ be the number of observations in I_ℓ. Denote by $E_\ell(\cdot) = n_\ell^{-1} \sum_{i \in I_\ell}(\cdot)$ the average over observations in I_ℓ. Denote by $E_n(\cdot) = n^{-1} \sum_{i=1}^n (\cdot)$ the average over all observations in the sample.

Definition D.2 (Foldwise target and oracle)**.**

$$\hat{\theta}_\ell = E_\ell[m(W, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W)\{Y - \hat{\gamma}_\ell(W)\}];$$
$$\bar{\theta}_\ell = E_\ell[m(W, \gamma_0) + \alpha_0^{\min}(W)\{Y - \gamma_0(W)\}].$$

Proposition D.3 (Taylor expansion)**.** *Let $u = \hat{\gamma}_\ell - \gamma_0$ and $v = \hat{\alpha}_\ell - \alpha_0^{\min}$. Then $n_\ell^{1/2}(\hat{\theta}_\ell - \bar{\theta}_\ell) = \sum_{j=1}^3 \Delta_{j\ell}$ where*

$$\Delta_{1\ell} = n_\ell^{1/2} E_\ell\{m(W, u) - \alpha_0^{\min}(W)u(W)\};$$
$$\Delta_{2\ell} = n_\ell^{1/2} E_\ell[v(W)\{Y - \gamma_0(W)\}];$$
$$\Delta_{3\ell} = \frac{n_\ell^{1/2}}{2} E_\ell\{-u(W)v(W)\}.$$

Proof. An exact Taylor expansion gives

$$\psi(w, \theta_0, \hat{\gamma}_\ell, \hat{\alpha}_\ell) - \psi_0(w) = \{\partial_\gamma \psi_0(w)\}(u) + \{\partial_\alpha \psi_0(w)\}(v) + \frac{1}{2}\{\partial^2_{\gamma,\alpha}\psi_0(w)\}(u, v).$$

Averaging over observations in I_ℓ

$$\begin{aligned}\hat{\theta}_\ell - \bar{\theta}_\ell &= E_\ell\{\psi(W, \theta_0, \hat{\gamma}_\ell, \hat{\alpha}_\ell)\} - E_\ell\{\psi_0(W)\} \\ &= E_\ell\{\partial_\gamma \psi_0(W)\}(u) + E_\ell\{\partial_\alpha \psi_0(W)\}(v) + \frac{1}{2} E_\ell\{\partial^2_{\gamma,\alpha}\psi_0(W)\}(u, v).\end{aligned}$$

Finally appeal to Proposition D.1. □

D.3 Residuals

Proposition D.4 (Residuals)**.** *Suppose Assumption 3.1 holds and*

$$E[\{Y - \gamma_0(W)\}^2 \mid W] \leq \bar{\sigma}^2, \quad \|\alpha_0^{\min}\|_\infty \leq \bar{\alpha}.$$

Then with probability $1 - \epsilon/L$,

$$|\Delta_{1\ell}| \leq t_1 = \left(\frac{6L}{\epsilon}\right)^{1/2} (\bar{Q} + \bar{\alpha}^2)^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2};$$
$$|\Delta_{2\ell}| \leq t_2 = \left(\frac{3L}{\epsilon}\right)^{1/2} \bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2};$$
$$|\Delta_{3\ell}| \leq t_3 = \frac{3L^{1/2}}{2\epsilon}\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.$$

Proof. We proceed in steps.

1. Markov inequality implies

$$\text{pr}(|\Delta_{1\ell}| > t_1) \leq \frac{E(\Delta_{1\ell}^2)}{t_1^2};$$
$$\text{pr}(|\Delta_{2\ell}| > t_2) \leq \frac{E(\Delta_{2\ell}^2)}{t_2^2};$$
$$\text{pr}(|\Delta_{3\ell}| > t_3) \leq \frac{E(|\Delta_{3\ell}|)}{t_3}.$$

2. Law of iterated expectations implies

$$E(\Delta_{1\ell}^2) = E\{E(\Delta_{1\ell}^2 \mid I_\ell^c)\};$$
$$E(\Delta_{2\ell}^2) = E\{E(\Delta_{2\ell}^2 \mid I_\ell^c)\};$$
$$E(|\Delta_{3\ell}|) = E\{E(|\Delta_{3\ell}| \mid I_\ell^c)\}.$$

3. Bounding conditional moments.

Conditional on I_ℓ^c, (u, v) are nonrandom. Moreover, observations within fold I_ℓ are independent and identically distributed. Hence by Proposition D.2 and assumption of finite $(\bar{Q}, \bar{\alpha})$

$$\begin{aligned}
E(\Delta_{1\ell}^2 \mid I_\ell^c) &= E\left([n_\ell^{1/2} E_\ell\{m(W,u) - \alpha_0^{\min}(W)u(W)\}]^2 \mid I_\ell^c\right) \\
&= E\left[\frac{n_\ell}{n_\ell^2} \sum_{i,j \in I_\ell} \{m(W_i,u) - \alpha_0^{\min}(W_i)u(W_i)\}\{m(W_j,u) - \alpha_0^{\min}(W_j)u(W_j)\} \mid I_\ell^c\right] \\
&= \frac{n_\ell}{n_\ell^2} \sum_{i,j \in I_\ell} E\left[\{m(W_i,u) - \alpha_0^{\min}(W_i)u(W_i)\}\{m(W_j,u) - \alpha_0^{\min}(W_j)u(W_j)\} \mid I_\ell^c\right] \\
&= \frac{n_\ell}{n_\ell^2} \sum_{i \in I_\ell} E\left[\{m(W_i,u) - \alpha_0^{\min}(W_i)u(W_i)\}^2 \mid I_\ell^c\right] \\
&= E[\{m(W,u) - \alpha_0^{\min}(W)u(W)\}^2 \mid I_\ell^c] \\
&\leq 2E\{m(W,u)^2 \mid I_\ell^c\} + 2E[\{\alpha_0^{\min}(W)u(W)\}^2 \mid I_\ell^c] \\
&\leq 2(\bar{Q} + \bar{\alpha}^2)\mathcal{R}(\hat{\gamma}_\ell)^q.
\end{aligned}$$

Similarly by Proposition D.2 and assumption of finite $\bar{\sigma}$

$$\begin{aligned}
E(\Delta_{2\ell}^2 \mid I_\ell^c) &= E\left\{(n_\ell^{1/2} E_\ell[v(W)\{Y - \gamma_0(W)\}])^2 \mid I_\ell^c\right\} \\
&= E\left[\frac{n_\ell}{n_\ell^2} \sum_{i,j \in I_\ell} v(W_i)\{Y_i - \gamma_0(W_i)\}v(W_j)\{Y_j - \gamma_0(W_j)\} \mid I_\ell^c\right] \\
&= \frac{n_\ell}{n_\ell^2} \sum_{i,j \in I_\ell} E\left[v(W_i)\{Y_i - \gamma_0(W_i)\}v(W_j)\{Y_j - \gamma_0(W_j)\} \mid I_\ell^c\right] \\
&= \frac{n_\ell}{n_\ell^2} \sum_{i \in I_\ell} E\left[v(W_i)^2\{Y_i - \gamma_0(W_i)\}^2 \mid I_\ell^c\right] \\
&= E[v(W)^2\{Y - \gamma_0(W)\}^2 \mid I_\ell^c] \\
&= E\left(v(W)^2 E[\{Y - \gamma_0(W)\}^2 \mid W, I_\ell^c] \mid I_\ell^c\right) \\
&\leq \bar{\sigma}^2 \mathcal{R}(\hat{\alpha}_\ell).
\end{aligned}$$

Finally by Cauchy Schwarz inequality

$$\begin{aligned}
E(|\Delta_{3\ell}| \mid I_\ell^c) &= \frac{n_\ell^{1/2}}{2} E\{|-u(W)v(W)| \mid I_\ell^c\} \\
&\leq \frac{n_\ell^{1/2}}{2} [E\{u(W)^2 \mid I_\ell^c\}]^{1/2} [E\{v(W)^2 \mid I_\ell^c\}]^{1/2} \\
&= \frac{n_\ell^{1/2}}{2} \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.
\end{aligned}$$

4. Collecting results gives

$$\begin{aligned}
\text{pr}(|\Delta_{1\ell}| > t_1) &\leq \frac{2(\bar{Q} + \bar{\alpha}^2)\mathcal{R}(\hat{\gamma}_\ell)^q}{t_1^2} = \frac{\epsilon}{3L}; \\
\text{pr}(|\Delta_{2\ell}| > t_2) &\leq \frac{\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell)}{t_2^2} = \frac{\epsilon}{3L}; \\
\text{pr}(|\Delta_{3\ell}| > t_3) &\leq \frac{n_\ell^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}}{2t_3} = \frac{\epsilon}{3L}.
\end{aligned}$$

Therefore with probability $1 - \epsilon/L$, the following inequalities hold:

$$\begin{aligned}
|\Delta_{1\ell}| &\leq t_1 = \left(\frac{6L}{\epsilon}\right)^{1/2} (\bar{Q} + \bar{\alpha}^2)^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2}; \\
|\Delta_{2\ell}| &\leq t_2 = \left(\frac{3L}{\epsilon}\right)^{1/2} \bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}; \\
|\Delta_{3\ell}| &\leq t_3 = \frac{3L}{2\epsilon} n_\ell^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.
\end{aligned}$$

Finally recall $n_\ell = n/L$.

□

Proposition D.5 (Residuals: Alternative path)**.** *Suppose Assumption 3.1 holds and*

$$E[\{Y - \gamma_0(W)\}^2 \mid W] \leq \bar{\sigma}^2, \quad \|\alpha_0^{\min}\|_\infty \leq \bar{\alpha}, \quad \|\hat{\alpha}_\ell\|_\infty \leq \bar{\alpha}'.$$

Then with probability $1 - \epsilon/L$*,*

$$\begin{aligned}
|\Delta_{1\ell}| &\leq t_1 = \left(\frac{6L}{\epsilon}\right)^{1/2} (\bar{Q} + \bar{\alpha}^2)^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2}; \\
|\Delta_{2\ell}| &\leq t_2 = \left(\frac{3L}{\epsilon}\right)^{1/2} \bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}; \\
|\Delta_{3\ell}| &\leq t_3 = \left(\frac{3L}{4\epsilon}\right)^{1/2} (\bar{\alpha} + \bar{\alpha}')\{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} \\
&\qquad + (4L)^{-1/2}[\{n\mathcal{P}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}].
\end{aligned}$$

Proof. See Proposition D.4 for (t_1, t_2). We focus on the alternative bound t_3.

1. Decomposition.

   Write

$$2\Delta_{3\ell} = n_\ell^{1/2} E_\ell\{-u(W)v(W)\} = \Delta_{3'\ell} + \Delta_{3''\ell},$$

   where

$$\begin{aligned}
\Delta_{3'\ell} &= n_\ell^{1/2} E_\ell[-u(W)v(W) + E\{u(W)v(W) \mid I_\ell^c\}]; \\
\Delta_{3''\ell} &= n_\ell^{1/2} E\{-u(W)v(W) \mid I_\ell^c\}.
\end{aligned}$$

2. Former term.

   By Markov inequality

$$\text{pr}(|\Delta_{3'\ell}| > t) \leq \frac{E(\Delta_{3'\ell}^2)}{t^2}.$$

   Law of iterated expectations implies

$$E(\Delta_{3'\ell}^2) = E\{E(\Delta_{3'\ell}^2 \mid I_\ell^c)\}.$$

We bound the conditional moment. Conditional on I_ℓ^c, (u, v) are nonrandom. Moreover, observations within fold I_ℓ are independent and identically distributed. Since each summand in $\Delta_{3'\ell}$ has conditional mean zero by construction, and since $(\bar{\alpha}, \bar{\alpha}')$ are finite by hypothesis,

$$
\begin{aligned}
&E(\Delta_{3'\ell}^2 \mid I_\ell^c) \\
&= E\left\{ \left(n_\ell^{1/2} E_\ell[-u(W)v(W) + E\{u(W)v(W) \mid I_\ell^c\}] \right)^2 \mid I_\ell^c \right\} \\
&= E\left(\frac{n_\ell}{n_\ell^2} \sum_{i,j \in I_\ell} [-u(W_i)v(W_i) + E\{u(W_i)v(W_i) \mid I_\ell^c\}][-u(W_j)v(W_j) + E\{u(W_j)v(W_j) \mid I_\ell^c\}] \mid I_\ell^c \right) \\
&= \frac{n_\ell}{n_\ell^2} \sum_{i,j \in I_\ell} E\left([-u(W_i)v(W_i) + E\{u(W_i)v(W_i) \mid I_\ell^c\}][-u(W_j)v(W_j) + E\{u(W_j)v(W_j) \mid I_\ell^c\}] \mid I_\ell^c \right) \\
&= \frac{n_\ell}{n_\ell^2} \sum_{i \in I_\ell} E\left([-u(W_i)v(W_i) + E\{u(W_i)v(W_i) \mid I_\ell^c\}]^2 \mid I_\ell^c \right) \\
&= E([u(W)v(W) - E\{u(W)v(W) \mid I_\ell^c\}]^2 \mid I_\ell^c) \\
&\leq E\{u(W)^2 v(W)^2 \mid I_\ell^c\} \\
&\leq (\bar{\alpha} + \bar{\alpha}')^2 \mathcal{R}(\hat{\gamma}_\ell).
\end{aligned}
$$

Collecting results gives

$$
\text{pr}(|\Delta_{3'\ell}| > t) \leq \frac{(\bar{\alpha} + \bar{\alpha}')^2 \mathcal{R}(\hat{\gamma}_\ell)}{t^2} = \frac{\epsilon}{3L}.
$$

Therefore with probability $1 - \epsilon/(3L)$,

$$
|\Delta_{3'\ell}| \leq t = \left(\frac{3L}{\epsilon} \right)^{1/2} (\bar{\alpha} + \bar{\alpha}') \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2}.
$$

3. Latter term.

Specializing to nonparametric instrumental variable regression,

$$
\begin{aligned}
E\{-u(W)v(W) \mid I_\ell^c\} &= E[E\{-u(W_1) \mid W_2, I_\ell^c\} v(W_2) \mid I_\ell^c] \\
&\leq \{E([E\{u(W_1) \mid W_2, I_\ell^c\}]^2 \mid I_\ell^c)\}^{1/2} [E\{v(W_2)^2 \mid I_\ell^c\}]^{1/2} \\
&= \{\mathcal{P}(\hat{\gamma}_\ell)\}^{1/2} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.
\end{aligned}
$$

Hence

$$
\Delta_{3''\ell} \leq n_\ell^{1/2} \{\mathcal{P}(\hat{\gamma}_\ell)\}^{1/2} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} = L^{-1/2} \{n \mathcal{P}(\hat{\gamma}_\ell) \mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.
$$

Likewise

$$
\begin{aligned}
E\{-u(W)v(W) \mid I_\ell^c\} &= E[-u(W_1) E\{v(W_2) \mid W_1, I_\ell^c\} \mid I_\ell^c] \\
&\leq [E\{u(W_1)^2 \mid I_\ell^c\}]^{1/2} \{E([E\{v(W_2) \mid W_1, I_\ell^c\}]^2 \mid I_\ell^c)\}^{1/2} \\
&= \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} \{\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}.
\end{aligned}
$$

Hence

$$
\Delta_{3''\ell} \leq n_\ell^{1/2} \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} \{\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2} = L^{-1/2} \{n \mathcal{R}(\hat{\gamma}_\ell) \mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}.
$$

4. Combining terms.

With probability $1 - \epsilon/(3L)$,

$$
\begin{aligned}
|\Delta_3| \leq t_3 = &\left(\frac{3L}{4\epsilon} \right)^{1/2} (\bar{\alpha} + \bar{\alpha}') \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} \\
&+ (4L)^{-1/2} [\{n \mathcal{P}(\hat{\gamma}_\ell) \mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n \mathcal{R}(\hat{\gamma}_\ell) \mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}].
\end{aligned}
$$

□

D.4 Main argument

Definition D.3 (Overall target and oracle)**.**

$$\hat{\theta} = \frac{1}{L}\sum_{\ell=1}^{L}\hat{\theta}_\ell, \quad \bar{\theta} = \frac{1}{L}\sum_{\ell=1}^{L}\bar{\theta}_\ell.$$

Proposition D.6 (Oracle approximation)**.** *Suppose the conditions of Proposition D.4 hold. Then with probability* $1-\epsilon$

$$\frac{n^{1/2}}{\sigma}|\hat{\theta}-\bar{\theta}| \leq \Delta = \frac{3L}{\epsilon\sigma}\left[(\bar{Q}^{1/2}+\bar{\alpha})\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} + \bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} + \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}\right].$$

Proof. We proceed in steps.

1. Decomposition.

   By Proposition D.3, write

   $$\begin{aligned} n^{1/2}(\hat{\theta}-\bar{\theta}) &= \frac{n^{1/2}}{n_\ell^{1/2}}\frac{1}{L}\sum_{\ell=1}^{L} n_\ell^{1/2}(\hat{\theta}_\ell - \bar{\theta}_\ell) \\ &= L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{3}\Delta_{j\ell}. \end{aligned}$$

2. Union bound.

   Define the events

   $$\mathcal{E}_\ell = \{\text{for all } j \ (j=1,2,3),\ |\Delta_{j\ell}| \leq t_j\}, \quad \mathcal{E} = \cap_{\ell=1}^{L}\mathcal{E}_\ell, \quad \mathcal{E}^c = \cup_{\ell=1}^{L}\mathcal{E}_\ell^c.$$

   Hence by the union bound and Proposition D.4,

   $$\text{pr}(\mathcal{E}^c) \leq \sum_{\ell=1}^{L}\text{pr}(\mathcal{E}_\ell^c) \leq L\frac{\epsilon}{L} = \epsilon.$$

3. Collecting results.

   Therefore with probability $1-\epsilon$,

   $$\begin{aligned} n^{1/2}|\hat{\theta}-\bar{\theta}| &\leq L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{3}|\Delta_{jk}| \\ &\leq L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{3}t_j \\ &= L^{1/2}\sum_{j=1}^{3}t_j. \end{aligned}$$

   Finally, we simplify (t_j). For $a, b > 0$, $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$. Moreover, $3 > 6^{1/2} > 3^{1/2} > 3/2$. Finally, for $\epsilon \leq 1$, $\epsilon^{-1/2} \leq \epsilon^{-1}$. In summary

   $$\begin{aligned} t_1 &= \left(\frac{6L}{\epsilon}\right)^{1/2}(\bar{Q}+\bar{\alpha}^2)^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} \leq \frac{3L^{1/2}}{\epsilon}(\bar{Q}^{1/2}+\bar{\alpha})\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2}; \\ t_2 &= \left(\frac{3L}{\epsilon}\right)^{1/2}\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \leq \frac{3L^{1/2}}{\epsilon}\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}; \\ t_3 &= \frac{3L^{1/2}}{2\epsilon}\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \leq \frac{3L^{1/2}}{\epsilon}\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}. \end{aligned}$$

□

Proposition D.7 (Oracle approximation: Alternative path)**.** *Suppose the conditions of Proposition D.5 hold. Then with probability* $1-\epsilon$

$$\frac{n^{1/2}}{\sigma}|\hat{\theta}-\bar{\theta}| \leq \Delta = \frac{4L}{\epsilon^{1/2}\sigma}\left[(\bar{Q}^{1/2}+\bar{\alpha}+\bar{\alpha}')\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2}+\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}\right] + \frac{1}{2L^{1/2}\sigma}[\{n\mathcal{P}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}].$$

Proof. As in Proposition D.6, Propositions D.3 and D.5 imply that with probability $1-\epsilon$

$$n^{1/2}|\hat{\theta}-\bar{\theta}| \leq L^{1/2}\sum_{j=1}^{3} t_j.$$

Finally, we simplify (t_j). For $a, b > 0$, $(a+b)^{1/2} \leq a^{1/2}+b^{1/2}$. In summary,

$$t_1 = \left(\frac{6L}{\epsilon}\right)^{1/2}(\bar{Q}+\bar{\alpha}^2)^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} \leq \left(\frac{6L}{\epsilon}\right)^{1/2}(\bar{Q}^{1/2}+\bar{\alpha})\{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2};$$

$$t_2 = \left(\frac{3L}{\epsilon}\right)^{1/2}\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2};$$

$$t_3 = \left(\frac{3L}{4\epsilon}\right)^{1/2}(\bar{\alpha}+\bar{\alpha}')\{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} + (4L)^{-1/2}[\{n\mathcal{P}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}].$$

Finally note $6^{1/2}+(3/4)^{1/2} \leq 4$ when combining terms from t_1 and t_3. □

Lemma D.1 (Berry Esseen Theorem [37])**.** *Suppose* (Z_i) $(i=1,...,n)$ *are independent and identically distributed random variables with* $E(Z_i)=0$*,* $E(Z_i^2)=\sigma^2$*, and* $E(|Z_i|^3)=\xi^3$*. Then*

$$\sup_{z\in\mathbb{R}}\left|\text{pr}\left\{\frac{n^{1/2}}{\sigma}E_n(Z_i)\leq z\right\}-\Phi(z)\right| \leq c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}},$$

where $c^{BE}=0.4748$ *and* $\Phi(z)$ *is the standard Gaussian cumulative distribution function.*

of Theorem 5.1. Fix z in $\mathbb{R}$. First, we show that

$$\text{pr}\left\{\frac{n^{1/2}}{\sigma}(\hat{\theta}-\theta_0)\leq z\right\}-\Phi(z) \leq c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}}+\frac{\Delta}{(2\pi)^{1/2}}+\epsilon,$$

where Δ is defined in Propositions D.6 and D.7. We proceed in steps.

1. High probability bound.

   By Propositions D.6 and D.7, with probability $1-\epsilon$,

   $$\frac{n^{1/2}}{\sigma}(\bar{\theta}-\hat{\theta}) \leq \frac{n^{1/2}}{\sigma}|\hat{\theta}-\bar{\theta}| \leq \Delta.$$

   Observe that

   $$\begin{aligned}\text{pr}\left\{\frac{n^{1/2}}{\sigma}(\hat{\theta}-\theta_0)\leq z\right\} &= \text{pr}\left\{\frac{n^{1/2}}{\sigma}(\bar{\theta}-\theta_0)\leq z+\frac{n^{1/2}}{\sigma}(\bar{\theta}-\hat{\theta})\right\}\\ &\leq \text{pr}\left\{\frac{n^{1/2}}{\sigma}(\bar{\theta}-\theta_0)\leq z+\Delta\right\}+\epsilon.\end{aligned}$$

2. Mean value theorem.

   Let $\phi(z)$ be the standard Gaussian probability density function. There exists some z' such that

   $$\Phi(z+\Delta)-\Phi(z)=\phi(z')\Delta \leq \frac{\Delta}{\sqrt{2\pi}}.$$

3. Berry Esseen theorem.

   Observe that

$$\bar{\theta} - \theta_0 = E_n[m(W, \gamma_0) + \alpha_0^{\min}(W)\{Y - \gamma_0(W)\}] - \theta_0 = E_n[\psi_0(W)].$$

   Therefore taking $Z_i = \psi_0(W_i)$ in Lemma D.1,

$$\sup_{z''} \left| \text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z'' \right\} - \Phi(z'') \right| \leq c^{BE} \left(\frac{\xi}{\sigma} \right)^3 n^{-\frac{1}{2}}.$$

   Hence by the high probability bound and mean value theorem steps above, taking $z'' = z + \Delta$

$$\begin{aligned}
&\text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\hat{\theta} - \theta_0) \leq z \right\} - \Phi(z) \\
&\leq \text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z + \Delta \right\} - \Phi(z) + \epsilon \\
&= \text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z + \Delta \right\} - \Phi(z + \Delta) + \Phi(z + \Delta) - \Phi(z) + \epsilon \\
&\leq c^{BE} \left(\frac{\xi}{\sigma} \right)^3 n^{-\frac{1}{2}} + \frac{\Delta}{\sqrt{2\pi}} + \epsilon.
\end{aligned}$$

Next, we show that

$$\Phi(z) - \text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\hat{\theta} - \theta_0) \leq z \right\} \leq c^{BE} \left(\frac{\xi}{\sigma} \right)^3 n^{-\frac{1}{2}} + \frac{\Delta}{(2\pi)^{1/2}} + \epsilon.$$

where Δ is defined in Propositions D.6 and D.7. We proceed in steps.

1. High probability bound.

   By Propositions D.6 and D.7, with probability $1 - \epsilon$,

$$\frac{n^{1/2}}{\sigma}(\hat{\theta} - \bar{\theta}) \leq \frac{n^{1/2}}{\sigma}|\hat{\theta} - \bar{\theta}| \leq \Delta,$$

   hence

$$z - \Delta \leq z - \frac{n^{1/2}}{\sigma}(\hat{\theta} - \bar{\theta}).$$

   Observe that

$$\begin{aligned}
\text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z - \Delta \right\} &\leq \text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z - \frac{n^{1/2}}{\sigma}(\hat{\theta} - \bar{\theta}) \right\} + \epsilon \\
&= \text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\hat{\theta} - \theta_0) \leq z \right\} + \epsilon.
\end{aligned}$$

2. Mean value theorem.

   There exists some z' such that

$$\Phi(z) - \Phi(z - \Delta) = \phi(z')\Delta \leq \frac{\Delta}{\sqrt{2\pi}}.$$

3. Berry Esseen theorem.

   As argued above,

$$\sup_{z''} \left| \text{pr}\left\{ \frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z'' \right\} - \Phi(z'') \right| \leq c^{BE} \left(\frac{\xi}{\sigma} \right)^3 n^{-\frac{1}{2}}.$$

Hence by the mean value theorem and high probability bound steps above, taking $z'' = z - \Delta$

$$\begin{aligned}
&\Phi(z) - \text{pr}\left\{\frac{n^{1/2}}{\sigma}(\hat{\theta} - \theta_0) \le z\right\} \\
&\le \Phi(z) - \text{pr}\left\{\frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \le z - \Delta\right\} + \epsilon \\
&= \Phi(z) - \Phi(z - \Delta) + \Phi(z - \Delta) - \text{pr}\left\{\frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \le z - \Delta\right\} + \epsilon \\
&\le \frac{\Delta}{\sqrt{2\pi}} + c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}} + \epsilon.
\end{aligned}$$

□

D.5 Variance estimation

Recall that $E_\ell(\cdot) = n_\ell^{-1}\sum_{i\in I_\ell}(\cdot)$ means the average over observations in I_ℓ and $E_n(\cdot) = n^{-1}\sum_{i=1}^n(\cdot)$ means the average over all observations in the sample.

Definition D.4 (Shorter notation)**.** *For i in I_ℓ, define*

$$\begin{aligned}
\psi_0(W_i) &= \psi(W_i, \theta_0, \gamma_0, \alpha_0^{\min}); \\
\hat{\psi}(W_i) &= \psi(W_i, \hat{\theta}, \hat{\gamma}_\ell, \hat{\alpha}_\ell).
\end{aligned}$$

Proposition D.8 (Foldwise second moment)**.**

$$E_\ell[\{\hat{\psi}(W) - \psi_0(W)\}^2] \le 4\left\{(\hat{\theta} - \theta_0)^2 + \sum_{j=4}^{6}\Delta_{j\ell}\right\},$$

where

$$\begin{aligned}
\Delta_{4\ell} &= E_\ell\{m(W,u)^2\}; \\
\Delta_{5\ell} &= E_\ell[\{\hat{\alpha}_\ell(W)u(W)\}^2]; \\
\Delta_{6\ell} &= E_\ell[v(W)^2\{Y - \gamma_0(W)\}^2].
\end{aligned}$$

Proof. Write

$$\begin{aligned}
\hat{\psi}(W_i) - \psi_0(W_i) &= m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_i)\{Y_i - \hat{\gamma}_\ell(W_i)\} - \hat{\theta} \\
&\quad - \left[m(W_i, \gamma_0) + \alpha_0^{\min}(W_i)\{Y_i - \gamma_0(W_i)\} - \theta_0\right] \\
&\quad \pm \hat{\alpha}_\ell\{Y - \gamma_0(W_i)\} \\
&= (\theta_0 - \hat{\theta}) + m(W_i, u) - \hat{\alpha}_\ell(W_i)u(W_i) + v(W_i)\{Y - \gamma_0(W_i)\}.
\end{aligned}$$

Hence

$$\{\hat{\psi}(W_i) - \psi_0(W_i)\}^2 \le 4\left[(\theta_0 - \hat{\theta})^2 + m(W_i,u)^2 + \{\hat{\alpha}_\ell(W_i)u(W_i)\}^2 + v(W_i)^2\{Y - \gamma_0(W_i)\}^2\right].$$

Finally take $E_\ell(\cdot)$ of both sides. □

Proposition D.9 (Residuals)**.** *Suppose Assumption 3.1 holds and*

$$E[\{Y - \gamma_0(W)\} \mid W]^2 \le \bar{\sigma}^2, \quad \|\hat{\alpha}_\ell\|_\infty \le \bar{\alpha}'.$$

Then with probability $1 - \epsilon'/(2L)$,

$$\begin{aligned}
\Delta_{4\ell} &\le t_4 = \frac{6L}{\epsilon'}\bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^q; \\
\Delta_{5\ell} &\le t_5 = \frac{6L}{\epsilon'}(\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell); \\
\Delta_{6\ell} &\le t_6 = \frac{6L}{\epsilon'}\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell).
\end{aligned}$$

Proof. We proceed in steps analogous to Proposition D.4.

1. Markov inequality implies

$$\begin{aligned}\text{pr}(|\Delta_{4\ell}| > t_4) &\leq \frac{E(|\Delta_{4\ell}|)}{t_4};\\ \text{pr}(|\Delta_{5\ell}| > t_5) &\leq \frac{E(|\Delta_{5\ell}|)}{t_5};\\ \text{pr}(|\Delta_{6\ell}| > t_6) &\leq \frac{E(|\Delta_{6\ell}|)}{t_6}.\end{aligned}$$

2. Law of iterated expectations implies

$$\begin{aligned}E(|\Delta_{4\ell}|) &= E\{E(|\Delta_{4\ell}| \mid I_\ell^c)\};\\ E(|\Delta_{5\ell}|) &= E\{E(|\Delta_{5\ell}| \mid I_\ell^c)\};\\ E(|\Delta_{6\ell}|) &= E\{E(|\Delta_{6\ell}| \mid I_\ell^c)\}.\end{aligned}$$

3. Bounding conditional moments.

   Conditional on I_ℓ^c, (u, v) are nonrandom. Moreover, observations within fold I_ℓ are independent and identically distributed. Hence by assumption of finite $(\bar{Q}, \bar{\sigma}, \bar{\alpha}')$

$$E(|\Delta_{4\ell}| \mid I_\ell^c) = E(\Delta_{4\ell} \mid I_\ell^c) = E[\{m(W, u)\}^2 \mid I_\ell^c] \leq \bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^q.$$

   Similarly

$$E(|\Delta_{5\ell}| \mid I_\ell^c) = E(\Delta_{5\ell} \mid I_\ell^c) = E[\{\hat{\alpha}_\ell(W)u(W)\}^2 \mid I_\ell^c] \leq (\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell).$$

   Finally

$$\begin{aligned}E(|\Delta_{6\ell}| \mid I_\ell^c) &= E(\Delta_{6\ell} \mid I_\ell^c)\\ &= E[v(W)^2\{Y - \gamma_0(W)\}^2 \mid I_\ell^c]\\ &= E\{v(W)^2 E[\{Y - \gamma_0(W)\}^2 \mid W, I_\ell^c] \mid I_\ell^c\}\\ &\leq \bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell).\end{aligned}$$

4. Collecting results gives

$$\begin{aligned}\text{pr}(|\Delta_{4\ell}| > t_4) &\leq \frac{\bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^q}{t_4} = \frac{\epsilon'}{6L}\\ \text{pr}(|\Delta_{5\ell}| > t_5) &\leq \frac{(\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell)}{t_5} = \frac{\epsilon'}{6L}\\ \text{pr}(|\Delta_{6\ell}| > t_6) &\leq \frac{\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell)}{t_6} = \frac{\epsilon'}{6L}\end{aligned}$$

   Therefore with probability $1 - \epsilon'/(2L)$, the following inequalities hold:

$$\begin{aligned}|\Delta_{4\ell}| &\leq t_4 = \frac{6L}{\epsilon'}\bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^q;\\ |\Delta_{5\ell}| &\leq t_5 = \frac{6L}{\epsilon'}(\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell);\\ |\Delta_{6\ell}| &\leq t_6 = \frac{6L}{\epsilon'}\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell).\end{aligned}$$

□

Proposition D.10 (Oracle approximation)**.** *Suppose the conditions of Proposition D.9 hold. Then with probability* $1 - \epsilon'/2$

$$E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2] \leq \Delta' = 4(\hat{\theta} - \theta_0)^2 + \frac{24L}{\epsilon'}\left[\{\bar{Q} + (\bar{\alpha}')^2\}\mathcal{R}(\hat{\gamma}_\ell)^q + \bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell)\right].$$

Proof. We proceed in steps analogous to Proposition D.6.

1. Decomposition.

   By Proposition D.8

$$E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2] = \frac{1}{L}\sum_{\ell=1}^{L} E_\ell[\{\hat{\psi}(W) - \psi_0(W)\}^2]$$
$$\leq 4(\hat{\theta} - \theta_0)^2 + \frac{4}{L}\sum_{\ell=1}^{L}\sum_{j=4}^{6}\Delta_{j\ell}.$$

2. Union bound.

   Define the events

$$\mathcal{E}'_\ell = \{\text{for all } j \ (j = 4, 5, 6),\ |\Delta_{j\ell}| \leq t_j\}, \quad \mathcal{E}' = \cap_{\ell=1}^{L}\mathcal{E}'_\ell, \quad (\mathcal{E}')^c = \cup_{\ell=1}^{L}(\mathcal{E}'_\ell)^c.$$

   Hence by the union bound and Proposition D.9,

$$\text{pr}\{(\mathcal{E}')^c\} \leq \sum_{\ell=1}^{L}\text{pr}\{(\mathcal{E}'_\ell)^c\} \leq L\frac{\epsilon'}{2L} = \frac{\epsilon'}{2}.$$

3. Collecting results.

   Therefore with probability $1 - \epsilon'/2$,

$$E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2] \leq 4(\hat{\theta} - \theta_0)^2 + \frac{4}{L}\sum_{\ell=1}^{L}\sum_{j=4}^{6}|\Delta_{j\ell}|$$
$$\leq 4(\hat{\theta} - \theta_0)^2 + \frac{4}{L}\sum_{\ell=1}^{L}\sum_{j=4}^{6}t_j$$
$$= 4(\hat{\theta} - \theta_0)^2 + 4\sum_{j=4}^{6}t_j.$$

□

Proposition D.11 (Markov inequality)**.** *Recall* $\sigma^2 = E\{\psi_0(W)^2\}$ *and* $\zeta^4 = E\{\psi_0(W)^4\}$*. Suppose* $\zeta < \infty$*. Then with probability* $1 - \epsilon'/2$

$$|E_n\{\psi_0(W)^2\} - \sigma^2| \leq \Delta'' = \left(\frac{2}{\epsilon'}\right)^{1/2}\frac{\zeta^2}{n^{1/2}}.$$

Proof. Let

$$A = \psi_0(W)^2, \quad \bar{A} = E_n(A).$$

Observe that

$$E(\bar{A}) = E(A) = E\{\psi_0(W)^2\} = \sigma^2, \quad \text{var}(\bar{A}) = \frac{\text{var}(A)}{n} \leq \frac{E(A^2)}{n} = \frac{E\{\psi_0(W)^4\}}{n} = \frac{\zeta^4}{n}.$$

By Markov inequality

$$\text{pr}[|E_n\{\psi_0(W)^2\} - \sigma^2| > t] = \text{pr}\{|\bar{A} - E(\bar{A})| > t\} \leq \frac{\text{var}(\bar{A})}{t^2} \leq \frac{\zeta^4}{nt^2}.$$

Solving,

$$\frac{\zeta^4}{nt^2} = \frac{\epsilon'}{2} \iff t = \left(\frac{2}{\epsilon'}\right)^{1/2}\frac{\zeta^2}{n^{1/2}}.$$

□

of Theorem 5.3. We proceed in steps.

1. Decomposition of variance estimator.

   Write
   $$\begin{aligned}\hat{\sigma}^2 &= E_n\{\hat{\psi}(W)^2\} \\ &= E_n[\{\hat{\psi}(W) - \psi_0(W) + \psi_0(W)\}^2] \\ &= E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2] + 2E_n[\{\hat{\psi}(W) - \psi_0(W)\}\psi_0(W)] + E_n\{\psi_0(W)^2\}.\end{aligned}$$
   Hence
   $$\hat{\sigma}^2 - E_n\{\psi_0(W)^2\} = E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2] + 2E_n[\{\hat{\psi}(W) - \psi_0(W)\}\psi_0(W)].$$

2. Decomposition of difference.

   Next write
   $$\hat{\sigma}^2 - \sigma^2 = [\hat{\sigma}^2 - E_n\{\psi_0(W)^2\}] + [E_n\{\psi_0(W)^2\} - \sigma^2].$$
   Focusing on the former term
   $$\hat{\sigma}^2 - E_n\{\psi_0(W)^2\} = E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2] + 2E_n[\{\hat{\psi}(W) - \psi_0(W)\}\psi_0(W)].$$
   Moreover
   $$\begin{aligned}E_n[\{\hat{\psi}(W) - \psi_0(W)\}\psi_0(W)] &\leq \left(E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2]\right)^{1/2} \left[E_n\{\psi_0(W)^2\}\right]^{1/2} \\ &\leq \left(E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2]\right)^{1/2} \left[|E_n\{\psi_0(W)^2\} - \sigma^2| + \sigma^2\right]^{1/2} \\ &\leq \left(E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2]\right)^{1/2} \left(\left[|E_n\{\psi_0(W)^2\} - \sigma^2|\right]^{1/2} + \sigma\right).\end{aligned}$$

3. High probability events.

   From the previous step, we see that to control $|\hat{\sigma}^2 - \sigma^2|$, it is sufficient to control two expressions: $E_n[\{\hat{\psi}(W) - \psi_0(W)\}^2]$ and $|E_n\{\psi_0(W)\}^2 - \sigma^2|$. These are controlled in Propositions D.10 and D.11, respectively. Therefore with probability $1 - \epsilon'$,
   $$|\hat{\sigma}^2 - \sigma^2| \leq \Delta' + 2(\Delta')^{1/2}\{(\Delta'')^{1/2} + \sigma\} + \Delta''.$$

□

D.6 Corollary

of Corollary 5.1. Immediately from $\Delta = o_p(1)$ in Theorem 5.1,
$$\hat{\theta} = \theta_0 + o_p(1), \quad \sigma^{-1}n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0,1), \quad \text{pr}\left\{\theta_0 \text{ in } \left(\hat{\theta} \pm \frac{\sigma}{n^{1/2}}\right)\right\} \to 1 - a.$$
For the final result, it is sufficient that $\hat{\sigma}^2 = \sigma^2 + o_p(1)$, which follows from $\Delta' = o_p(1)$ and $\Delta'' = o_p(1)$ in Theorem 5.3. □

E Proofs of lemmas

E.1 Riesz representers

of Lemma 3.1. As the operator norm of $\gamma \mapsto E\{m(W,\gamma)\}$,
$$\bar{M} = \inf[c \geq 0 : |E\{m(W,\gamma)\}| \leq c \text{ for all } \gamma \in \Gamma \text{ such that } \|\gamma\|_{\text{pr},2} = 1].$$
By Jensen's inequality and Assumption 3.1,
$$|E\{m(W,\gamma)\}| \leq [E\{m(W,\gamma)^2\}]^{1/2} \leq \left(\bar{Q}[E\{\gamma(W)^2\}]^q\right)^{1/2} = \bar{Q}^{1/2}\|\gamma\|_{\text{pr},2}^q.$$
Taking the supremum of both sides over γ in Γ such that $\|\gamma\|_{\text{pr},2} = 1$, we conclude that $\bar{M} \leq \bar{Q}^{1/2} < \infty$. The rest of the claim is shown in [16, Lemma 2.1]. □

of Lemma B.1. For Examples 3.1 and 3.2, the result is immediate from standard propensity score and regression arguments. For Example 3.3. the result follows from [23, Proposition 3 and Example 5]. For Example B.1, the result follows from integration by parts. □

E.2 Global functionals

of Lemma B.2. We extend [16, Lemma 3.3]. Write

$$\psi_0(W) = U_1 + \alpha_0^{\min}(W)U_2.$$

By law of iterated expectations,

$$\begin{aligned}\sigma^2 &= E(U_1^2) + 2E\{U_1\alpha_0^{\min}(W)U_2\} + E(\{\alpha_0^{\min}(W)U_2\}^2) \\ &= E\{E(U_1^2 \mid W)\} + 2E\{U_1\alpha_0^{\min}(W)E(U_2 \mid W)\} + E\{\alpha_0^{\min}(W)^2E(U_2^2 \mid W)\}.\end{aligned}$$

$E(U_2 \mid W) = 0$ by definition of U_2. Note that

$$0 \leq E\{E(U_1^2 \mid W)\} \leq \bar{c}^2,$$

and

$$\tilde{c}^2\bar{M}^2 \leq E\{\alpha_0^{\min}(W)^2E(U_2^2 \mid W)\} \leq \bar{c}^2\bar{M}^2.$$

In summary

$$\tilde{c}^2\bar{M}^2 \leq \sigma^2 \leq \bar{c}^2(1 + \bar{M}^2).$$

By triangle inequality,

$$\begin{aligned}\|\psi_0\|_{\text{pr},q} &\leq \|U_1\|_{\text{pr},q} + \|\alpha_0^{\min}(W)U_2\|_{\text{pr},q} \\ &= \|U_1\|_{\text{pr},q} + [E\{\alpha_0^{\min}(W)^qE(U_2^q \mid W)\}]^{1/q} \\ &\leq \bar{c} + \bar{c}\|\alpha_0^{\min}\|_{\text{pr},q} \\ &\leq \bar{c}(1 + c(\bar{M}^2 \vee 1)).\end{aligned}$$

□

of Lemma B.3. To begin, observe that

$$\begin{aligned}\partial_d\{f(d \mid x)\partial_d\gamma(d,x)\} &= \{\partial_d f(d \mid x)\}\{\partial_d\gamma(d,x)\} + f(d \mid x)\{\partial_d^2\gamma(d,x)\} \\ &= [\{\partial_d \log f(d \mid x)\}\{\partial_d\gamma(d,x)\} + \partial_d^2\gamma(d,x)]f(d \mid x) \\ &= -k_\gamma(d,x)f(d \mid x).\end{aligned}$$

Using integration by parts and the boundary condition together with this result,

$$\begin{aligned}E[\{\partial_d\gamma(D,X)\}^2] &= \int \{\partial_d\gamma(d,x)\}^2 f(d \mid x)f(x)\mathrm{d}d\mathrm{d}x \\ &= \int \{\partial_d\gamma(d,x)\}\{f(d \mid x)\partial_d\gamma(d,x)\}f(x)\mathrm{d}d\mathrm{d}x \\ &= -\int \gamma(d,x)\partial_d\{f(d \mid x)\partial_d\gamma(d,x)\}f(x)\mathrm{d}d\mathrm{d}x \\ &= \int \gamma(d,x)k_\gamma(d,x)f(d \mid x)f(x)\mathrm{d}d\mathrm{d}x \\ &= E\{\gamma(D,X)k_\gamma(D,X)\} \\ &\leq \|\gamma\|_{\text{pr},2}\|k_\gamma\|_{\text{pr},2},\end{aligned}$$

where the inequality is Cauchy Schwarz. The final results immediately follow from the definition of k_γ and triangle inequality. □

of Lemma B.4. For Example 3.1, write

$$E[\{\ell_h(V)\gamma(1,V,X) - \ell_h(V)\gamma(0,V,X)\}^2] \leq 2E\{\ell_h(V)^2\gamma(1,V,X)^2\} + 2E\{\ell_h(V)^2\gamma(0,V,X)^2\}.$$

Invoking the bounded weighting and propensity score assumptions,

$$\begin{aligned}E\{\ell_h(V)^2\gamma(1,V,X)^2\} &= E\left\{\frac{D}{\pi_0(V,X)}\ell_h(V)^2\gamma(D,V,X)^2\right\} \leq CE\left\{\gamma(D,V,X)^2\right\}; \\ E\{\ell_h(V)^2\gamma(0,V,X)^2\} &= E\left\{\frac{1-D}{1-\pi_0(V,X)}\ell_h(V)^2\gamma(D,V,X)^2\right\} \leq CE\left\{\gamma(D,V,X)^2\right\}.\end{aligned}$$

For Example 3.2, write

$E[\{\ell_h^+(D)\gamma_0(D,X) - \ell_h^-(D)\gamma_0(D,X)\}^2] \leq 2E[\{\ell_h^+(D)\gamma_0(D,X)\}^2] + 2E[\{\ell_h^-(D)\gamma_0(D,X)\}^2]$.

Invoking the bounded weighting assumption,

$$E[\{\ell_h^+(D)\gamma_0(D,X)\}^2] \leq CE\{\gamma_0(D,X)\}^2;$$
$$E[\{\ell_h^-(D)\gamma_0(D,X)\}^2] \leq CE\{\gamma_0(D,X)\}^2.$$

For Examples 3.3 and B.1, appeal to Lemma B.3. □

of Lemma B.5. The result is immediate from Lemma B.1. □

E.3 Local functionals

of Lemma B.6. We extend [16, Lemma 3.4]. We proceed in steps.

1. Moment bounds.

   As in the of Lemma B.2,

   $$\sigma^2 = E\{E(U_1^2 \mid W)\} + E\{\alpha_0^{\min,h}(W)^2 E(U_2^2 \mid W)\}.$$

   Note that

   $$0 \leq E\{E(U_1^2 \mid W)\} \leq \bar{c}^2 \|\ell\|_{\text{pr},2}^2,$$

   and

   $$\tilde{c}^2 \|\alpha_0^{\min,h}\|_{\text{pr},2}^2 \leq E\{\alpha_0^{\min,h}(W)^2 E(U_2^2 \mid W)\} \leq \bar{c}^2 \|\alpha_0^{\min,h}\|_{\text{pr},2}^2.$$

   In summary

   $$\tilde{c}^2 \|\alpha_0^{\min,h}\|_{\text{pr},2}^2 \leq \sigma^2 \leq \bar{c}^2 (\|\ell\|_{\text{pr},2}^2 + \|\alpha_0^{\min,h}\|_{\text{pr},2}^2).$$

   As in the proof of Lemma B.2,

   $$\|\psi_0\|_{\text{pr},q} \leq \|U_1\|_{\text{pr},q} + [E\{\alpha_0^{\min,h}(W)^q E(U_2^q \mid W)\}]^{1/q} \leq \bar{c}(\|\ell\|_{\text{pr,q}} + \|\alpha_0^{\min,h}\|_{\text{pr,q}}).$$

   Next we characterize $\|\alpha_0^{\min,h}\|_{\text{pr,q}}$ in terms of $\|\ell\|_{\text{pr,q}}$. Since $\alpha_0^{\min,h}(w) = \ell_h(w_j)\alpha_0^{\min}$,

   $$\tilde{\alpha}\|\ell\|_{\text{pr},q} \leq \|\alpha_0^{\min,h}\|_{\text{pr},q} \leq \check{\alpha}\|\ell\|_{\text{pr},q}, \quad \|\alpha_0^{\min,h}\|_{\text{pr},2} = \bar{M}.$$

   In summary,

   $$\tilde{c}\tilde{\alpha}\|\ell\|_{\text{pr},2} \leq \sigma \leq \bar{c}\sqrt{1+\check{\alpha}^2}\|\ell\|_{\text{pr},2}, \quad \tilde{\alpha}\|\ell\|_{\text{pr},2} \leq \bar{M} \leq \check{\alpha}\|\ell\|_{\text{pr},2}, \quad \|\psi_0\|_{\text{pr},q} \leq \bar{c}(1+\check{\alpha})\|\ell\|_{\text{pr},q}.$$

2. Taylor expansion.

   Consider the change of variables $u = (v' - v)/h$ so that $\mathsf{d}u = h^{-1}\mathsf{d}v'$. Hence

   $$\begin{aligned}
   \|\ell\|_{\text{pr},q}^q \omega^q &= \|\ell\omega\|_{\text{pr},q}^q \\
   &= \left\| h^{-1} K\left(\frac{v - v'}{h}\right) \right\|_{\text{pr},q}^q \\
   &= \int h^{-q} \left| K\left(\frac{v' - v}{h}\right) \right|^q f_V(v')\mathsf{d}v' \\
   &= \int h^{-(q-1)} |K(u)|^q f_V(v - uh)\mathsf{d}u.
   \end{aligned}$$

   It follows that

   $$h^{-(q-1)/q} \tilde{f}^{1/q} \left(\int |K|^q \right)^{1/q} \leq \|\ell\|_{\text{pr},q}\omega \leq h^{-(q-1)/q} \bar{f}^{1/q} \left(\int |K|^q \right)^{1/q}.$$

   Further, we have that

   $$\omega = \int h^{-1} K\left(\frac{v' - v}{h}\right) f_V(v')\mathsf{d}v' = \int K(u) f_V(v - uh)\mathsf{d}u.$$

Note that

$$\int K(u) f_V(v - 0u) \mathsf{d}u = \int K(u) f_V(v) \mathsf{d}u = f_V(v).$$

Using the Taylor expansion in h around $h = 0$ and the Holder inequality, there exist some $\tilde{h}$ in $[0, h]$ such that

$$|\omega - f_V(v)| = \left| h \int K(u) \partial_v f_V(v - u\tilde{h}) u \mathsf{d}u \right| \leq h\bar{f}' \int |u||K(u)|du.$$

Hence there exists some h_1 in (h, h_0) depending only on $(K, \bar{f}', \tilde{f}, \bar{f})$ such that

$$\tilde{f}/2 \leq \omega \leq 2\bar{f}.$$

In summary,

$$h^{-(q-1)/q} \tilde{f}^{1/q} \left(\int |K|^q \right)^{1/q} \frac{1}{2\bar{f}} \leq \|\ell\|_{\text{pr},q} \leq h^{-(q-1)/q} \bar{f}^{1/q} \left(\int |K|^q \right)^{1/q} \frac{2}{\tilde{f}}.$$

3. Collecting results.

In summary, for all $h < h_1$

$$\tilde{c}\tilde{\alpha}\|\ell\|_{\text{pr},2} \leq \sigma \leq \bar{c}\sqrt{1 + \check{\alpha}^2}\|\ell\|_{\text{pr},2}, \quad \tilde{\alpha}\|\ell\|_{\text{pr},2} \leq \bar{M} \leq \check{\alpha}\|\ell\|_{\text{pr},2}, \quad \|\psi_0\|_{\text{pr},q} \leq \bar{c}(1+\check{\alpha})\|\ell\|_{\text{pr},q},$$

where

$$h^{-(q-1)/q} \tilde{f}^{1/q} \left(\int |K|^q \right)^{1/q} \frac{1}{2\bar{f}} \leq \|\ell\|_{\text{pr},q} \leq h^{-(q-1)/q} \bar{f}^{1/q} \left(\int |K|^q \right)^{1/q} \frac{2}{\tilde{f}},$$

so

$$\sigma \asymp \bar{M} \asymp \|\ell\|_{\text{pr},2}, \quad \|\psi_0\|_{\text{pr,q}} \lesssim \|\ell\|_{\text{pr},q}, \quad \|\ell\|_{\text{pr},q} \asymp h^{-(q-1)/q}.$$

□

of Lemma B.7. We prove the result for Example 3.1. The result for Example 3.2 is similar.

Without loss of generality, let $\bar{Q}_h$ be the smallest finite constant for which Assumption 3.1 holds, i.e.

$$\bar{Q}_h = \inf[c \geq 0 : E\{m_h(W, \gamma)^2\} \leq c\|\gamma\|_{\text{pr},2}^2 \text{ for all } \gamma \text{ in } \Gamma].$$

To begin, write

$$\begin{aligned} E\{m_h(W, \gamma)^2\} &= E[\ell_h(V)^2\{\gamma(1, V, X) - \gamma(0, V, X)\}^2] \\ &\leq 2E\{\ell_h(V)^2\gamma(1, V, X)^2\} + 2E\{\ell_h(V)^2\gamma(0, V, X)^2\}. \end{aligned}$$

Since $\pi_0(v, x)$ is bounded away from zero and one,

$$E\{\ell_h(V)^2\gamma(1, V, X)^2\} = E\left\{ \ell_h(V)^2 \frac{D}{\pi_0(V, X)} \gamma(D, V, X)^2 \right\} \leq CE\left\{ \ell_h(V)^2\gamma(D, V, X)^2 \right\}.$$

Likewise for $E\{\ell_h(V)^2\gamma(0, V, X)^2\}$. In summary,

$$E\{m_h(W, \gamma)^2\} \leq 2CE\left\{ \ell_h(V)^2\gamma(D, V, X)^2 \right\}.$$

Viewing the latter expression as an inner product in $\mathbb{L}_2$, it is maximized by alignment, i.e. taking $\gamma(D, V, X)^2 = \ell_h(V)^2$. Therefore

$$\frac{E\{m_h(W, \gamma)^2\}}{E\{\gamma(W)^2\}} \leq \frac{2CE\left\{ \ell_h(V)^4 \right\}}{E\left\{ \ell_h(V)^2 \right\}} = 2C \frac{\|\ell\|_{\text{pr},4}^4}{\|\ell\|_{\text{pr},2}^2}.$$

Appealing to $\|\ell\|_{\text{pr},q} \asymp h^{-(q-1)/q}$ from the proof of Lemma B.6,

$$\frac{E\{m_h(W, \gamma)^2\}}{E\{\gamma(W)^2\}} \lesssim \frac{h^{-3}}{h^{-1}} = h^{-2}.$$

□

of Lemma B.8. Write

$$\|\alpha_0^{\min,h}\|_\infty \le \check{\alpha}\|\ell\|_\infty.$$

By the proof of Lemma B.6,

$$\|\ell\|_\infty = \left\| \frac{1}{h\omega} K\left(\frac{v'-v}{h}\right)\right\|_\infty \le \bar{K}\frac{1}{h\omega} \le \bar{K}\frac{2}{h\tilde{f}}.$$

Therefore

$$\|\alpha_0^{\min,h}\|_\infty \le \check{\alpha}\bar{K}\frac{2}{h\tilde{f}} \lesssim h^{-1}.$$

□

of Lemma B.9. Write

$$\begin{aligned}
\mathcal{R}(\hat{\alpha}_\ell^h) &= E[\{\hat{\alpha}_\ell^h(W) - \alpha_0^{\min,h}(W)\}^2 \mid I_\ell^c] \\
&= E[\{\ell_h(W_i)\hat{\alpha}_\ell(W) - \ell_h(W_i)\alpha_0^{\min}(W)\}^2 \mid I_\ell^c] \\
&\le \|\ell_h\|_\infty^2 E[\{\hat{\alpha}_\ell(W) - \alpha_0^{\min}(W)\}^2 \mid I_\ell^c] \\
&= \|\ell_h\|_\infty^2 \mathcal{R}(\hat{\alpha}_\ell).
\end{aligned}$$

Finally recall from the proof of Lemma B.8 that $\|\ell_h\|_\infty \lesssim h^{-1}$. An identical argument holds for $\mathcal{P}(\hat{\alpha}_\ell^h)$. □

E.4 Approximation error

of Lemma B.10. For completeness, we quote the proof of [16, Lemma 3.6]. Define the quantities

$$\begin{aligned}
\vartheta_1(h) &= \int m(v')h^{-1}K\left(\frac{v-v'}{h}\right) f_V(v')\mathsf{d}v' = \int m(v-hu)K(u)f_V(v-hu)\mathsf{d}u; \\
\vartheta_2(h) &= \int h^{-1}K\left(\frac{v-v'}{h}\right) f_V(v')\mathsf{d}v' = \int K(u)f_V(v-uh)\mathsf{d}u.
\end{aligned}$$

By $\int K = 1$,

$$\vartheta_1(0) = m(v)f_V(v), \quad \vartheta_2(0) = f_V(v).$$

Hence

$$\theta_0^h = \frac{\vartheta_1(h)}{\vartheta_2(h)}, \quad \theta_0^{\lim} = \frac{\vartheta_1(0)}{\vartheta_2(0)} = m(v).$$

The standard argument to control the bias of the higher order kernels employs the Taylor expansion of order v in h around $h = 0$; see e.g. [29, Lemma B2]. Such an argument implies there exists some constant A_v that depends only on v such that

$$|\vartheta_1(h) - \vartheta_1(0)| \le A_\mathsf{v} h^\mathsf{v} \bar{g}_\mathsf{v} \int |u|^\mathsf{v}|K(u)|du,$$

$$|\vartheta_2(h) - \vartheta_2(0)| \le A_\mathsf{v} h^\mathsf{v} \bar{f}_\mathsf{v} \int |u|^\mathsf{v}|K(u)|du.$$

Then using the relation

$$\begin{aligned}
&\frac{\vartheta_1(h)}{\vartheta_2(h)} - \frac{\vartheta_1(0)}{\vartheta_2(0)} \\
&= \vartheta_2^{-1}(0)\{\vartheta_1(h) - \vartheta_1(0)\} + \vartheta_1(0)\{\vartheta_2^{-1}(h) - \vartheta_2^{-1}(0)\} + \{\vartheta_1(h) - \vartheta_1(0)\}\{\vartheta_2^{-1}(h) - \vartheta_2^{-1}(0)\},
\end{aligned}$$

we deduce that for all $h < h_1 \le h_0$,

$$|\theta_0^h - \theta_0^{\lim}| \le \left|\frac{\vartheta_1(h)}{\vartheta_2(h)} - \frac{\vartheta_1(0)}{\vartheta_2(0)}\right| \le Ch^\mathsf{v},$$

where C and h_1 depend only on $(K, \mathsf{v}, \bar{g}_\mathsf{v}, \bar{f}_\mathsf{v}, \tilde{f})$. □

of Corollary B.1. By Lemma B.6, write the regularity condition on moments as

$$\left\{(\kappa/\sigma)^3 + \zeta^2\right\} n^{-1/2} \lesssim \left\{\left(h^{-1/6}\right)^3 + (h^{-3/4})^2\right\} n^{-1/2} \lesssim h^{-3/2} n^{-1/2}.$$

By Lemmas B.6, B.7, and B.8, write the first learning rate condition as

$$\left(\bar{Q}^{1/2} + \bar{\alpha}/\sigma + \bar{\alpha}'\right) \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} \lesssim \left(h^{-1} + h^{-1}/h^{-1/2} + \bar{\alpha}'\right) \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2} \lesssim \left(h^{-1} + \bar{\alpha}'\right) \{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2}.$$

By Lemma B.9, write the second learning rate condition as

$$\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell^h)\}^{1/2} \lesssim \bar{\sigma} h^{-1} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.$$

By Lemmas B.6 and B.9, write the initial term in the third learning rate condition as

$$\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell^h)\}^{1/2}/\sigma \lesssim \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} h^{-1}/h^{-1/2} = h^{-1/2}\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.$$

Likewise for the other terms. The approximation error condition is immediate from Lemma B.10. □