


A simple and general debiased machine learning theorem with finite-sample guarantees

By V. CHERNOZHUKOV, W. K. NEWEY AND R. SINGH 

*Department of Economics, Massachusetts Institute of Technology,
50 Memorial Drive, Cambridge, Massachusetts 02142, U.S.A.*
vchern@mit.edu wnewey@mit.edu rahul.singh@mit.edu

SUMMARY

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 seek the confidence interval for a treatment effect estimated with a neural network. We present a non-asymptotic 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.

Some key words: Gaussian approximation; Ill-posed inverse; Non-asymptotic rate; Semiparametric inference.

1. INTRODUCTION

The goal of this paper is to provide a useful technical result for analysts who wish to find 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 such applications, confidence intervals are essential.

We give 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. This work makes conceptual and statistical contributions to 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.

Statistically, we present 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. 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 (Hasminskii & Ibragimov, 1979; Robinson, 1988; Bickel et al., 1993; Andrews, 1994; Newey, 1994; Robins & Rotnitzky, 1995; Ai & Chen, 2003). 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 (Newey, 1994). For the same reasons, it appears in debiased machine learning confidence intervals.

In asymptotic inference, the Riesz representer is inevitable. A growing body of work directly incorporates the Riesz representer into estimation, which amounts to debiasing known estimators. Doubly robust estimating equations serve this purpose (Robins & Rotnitzky, 1995). A geometric perspective emphasizes Neyman orthogonality: by debiasing, the learning problem for the functional becomes orthogonal to the learning problem for the nonparametric object (Chernozhukov et al., 2022c, 2018; Foster & Syrgkanis, 2020). An analytic perspective emphasizes the mixed bias property: by debiasing, the functional has bias equal to the product of certain learning rates (Chernozhukov et al., 2018; Rotnitzky et al., 2021). In the present 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 (van der Laan & Rubin, 2006; Luedtke & van der Laan, 2016; van der Laan & Rose, 2018; Qiu et al., 2021) or must have slowly increasing entropy (Belloni et al., 2013, 2014; Javanmard & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014). However, popular nonparametric settings in machine learning may not satisfy either property. A solution to this challenging issue is to combine debiasing with sample splitting (Klaassen, 1987). The targeted (Zheng & van der Laan, 2011) and debiased (Belloni et al., 2012; Chernozhukov et al., 2018, 2022c) 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 Singh (2021, § 4).

3. FRAMEWORK AND EXAMPLES

The general inference problem is to find a confidence interval for some scalar $\theta_0 \in \mathbb{R}$ where $\theta_0 = E\{m(W, \gamma_0)\}$, with $\gamma_0 \in \Gamma$ and $m : \mathcal{W} \times \mathbb{L}_2 \rightarrow \mathbb{R}$ being 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 $\gamma : \mathcal{W} \rightarrow \mathbb{R}$ that are square-integrable with respect to measure pr , and Γ is a linear subset of \mathbb{L}_2 known by the analyst, which may be \mathbb{L}_2 itself.

Here γ_0 may be the conditional expectation function $\gamma_0(w) = E(Y | W = w)$ or some other nonparametric quantity. For example, it could be the function defined as the solution to the ill-posed inverse problem $E(Y | W_2 = w_2) = E\{\gamma(W_1) | W_2 = w_2\}$ where $W_1, W_2 \subset W$. Such a function is called a nonparametric instrumental variable regression in econometrics (Newey & Powell, 2003). We study the exactly identified case, which amounts to assuming completeness when $\Gamma = \mathbb{L}_2$ (Chen & Santos, 2018). If $W_1 = W_2$, then nonparametric instrumental variable regression simplifies to nonparametric regression.

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

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

where ℓ_h is a Nadaraya–Watson weighting with bandwidth h and W_j is a scalar component of W . Though θ_0^{lim} is a nonparametric quantity, it can be approximated by the sequence (θ_0^h) . Each θ_0^h can be analysed like

θ_0 above as long as we keep track of how certain quantities depend on h . By this logic, finite-sample semi-parametric theory for θ_0^h translates to finite-sample nonparametric theory for θ_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 1 (Heterogeneous treatment effect estimated by neural network). Let Y be a health outcome. Let $W = (D, V, X)$ concatenate a binary treatment D , a 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

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

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 1.

The heterogeneous treatment effect is defined with respect to some interpretable, low-dimensional characteristic V such as age, race or gender (Abrevaya et al., 2015). The same functional without the localization ℓ_h is the classic average treatment effect. See Bibaut & van der Laan (2017) and Colangelo & Lee (2021) for other meaningful localizations of the average treatment effect.

Example 2 (Regression discontinuity design estimated by random forest). Let Y be an educational outcome. Let $W = (D, X)$ concatenate a 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 cut-off for a scholarship is the test score $D = 0$. The regression discontinuity design parameter is

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

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

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

Example 3 (Demand elasticity estimated by kernel instrumental variable regression). Let Y be the logarithm of the quantity demanded of some goods. Let $W = (D, X, Z)$ concatenate the 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 (Singh et al., 2019). The demand elasticity is

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

The [Supplementary Material](#) includes the additional example of heterogeneous average derivative estimated by lasso, which is useful when an analyst has access to data on household spending behaviour.

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

Assumption 1 (Linearity and mean-square continuity). The functional $\gamma \mapsto E\{m(W, \gamma)\}$ is linear, and 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 $\gamma \in \Gamma$.

This condition will be key in § 5, where we reduce the problem of inference for θ_0 to the problem of learning $(\gamma_0, \alpha_0^{\text{min}})$, where α_0^{min} is introduced below. It is a powerful condition satisfied by many functionals of interest, or at least satisfied by their approximating sequences. Although the local functional θ_0^{lim} does not satisfy Assumption 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 2 for conditions that characterize \bar{Q}_h in local functionals, including Examples 1 and 2.

The restriction that γ_0 be 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 $\gamma_0 \in \Gamma$. We quote a general Riesz representation theorem for restricted models.

LEMMA 1 (RIESZ REPRESENTATION, Chernozhukov et al., 2022a). *Suppose that Assumption 1 holds. Further suppose that γ_0 is in Γ . Then there exists a Riesz representer $\alpha_0 \in \mathbb{L}_2$ such that for all γ in Γ , $E\{m(W, \gamma)\} = E\{\alpha_0(W)\gamma(W)\}$. There exists a unique minimal Riesz representer $\alpha_0^{\min} \in \text{closure}(\Gamma)$ 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 1 to hold. Since $\bar{M} \leq \bar{Q}^{1/2}$, $\bar{Q} < \infty$ in Assumption 1 is a sufficient condition. Nonetheless, we assume $\bar{Q} < \infty$ because mean-square continuity plays a central role in the main results of § 5. In Examples 1 and 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 $\theta_0 \in \mathbb{R}$. For the case where $\gamma_0(w)$ is defined as a nonparametric regression in Γ or a 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 α_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 . 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 $E(Y | W = w)$ in Γ . The minimal Riesz representer α_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 $E(Y | W = w)$ onto Γ ; see Chernozhukov et al. (2022a, 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 (Severini & Tripathi, 2012; Ichimura & Newey, 2022). The relevant nuisance parameters are $(\gamma_0, \alpha_0^{\min})$, defined as unique solutions (γ, α) to

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

where η_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 1. Uniqueness is due to the assumption of exact identification, which amounts to completeness when $\Gamma = \mathbb{L}_2$. In Example 3, $w_1 = (d, x)$, $w_2 = (z, x)$ and $\eta_0(d, x) = -\partial_d \log f(d | x)$ where $f(d | x)$ 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 provided by this work is that for any mean-square-continuous functional, $n^{-1/2}$ Gaussian approximation is still possible if either γ_0 or α_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 $\theta_0 \in \mathbb{R}$, which is a mean-square-continuous functional of $\gamma_0 \in \Gamma$. Lemma 1 demonstrates that any such θ_0 has a unique minimal

representer $\alpha_0^{\min} \in \Gamma$. In this section, we describe a meta-algorithm for turning estimators $\hat{\gamma}$ of γ_0 and $\hat{\alpha}$ of α_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 $\hat{\gamma}$ converge 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 α_0^{\min} . As we will see in subsequent analysis, the general theory only requires that $\hat{\alpha}$ converge to α_0^{\min} in mean-square error. Recent literature provides $\hat{\alpha}$ estimators with fast rates inspired by the Dantzig selector (Chernozhukov et al., 2022a), lasso (Chernozhukov et al., 2022b; Avagyan & Vansteelandt, 2023; Smucler et al., 2019), adversarial neural networks (Chernozhukov et al., 2020; Kallus et al., 2021) and kernel ridge regression (Singh, 2021).

Algorithm 1. Debiased machine learning.

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

Step 1. For each fold ℓ , estimate $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ from observations in I_ℓ^c .

Step 2. Estimate θ_0 as $\hat{\theta} = n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} [m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_i)\{Y_i - \hat{\gamma}_\ell(W_i)\}]$.

Step 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 (Pfanzagl, 1982) amenable to the use of modern machine learning, and it has been termed debiased machine learning (Chernozhukov et al., 2018). It departs from targeted machine learning inference with a finite sample (van der Laan, 2017; Cai & van der Laan, 2020) 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 Chernozhukov et al. (2022b) for an algorithm that combines the two approaches.

5. VALIDITY OF THE CONFIDENCE INTERVAL

This section is presented at a high level of generality so that 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 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 $c^{\text{BE}} = 0.4748$ (Shevtsova, 2011). The result will be in terms of abstract mean-square rates.

DEFINITION 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 (Blundell et al., 2007; Singh et al., 2019; Dikkala et al., 2020). Our main result is a finite-sample Gaussian approximation.

THEOREM 1 (FINITE-SAMPLE GAUSSIAN APPROXIMATION). *Suppose that Assumption 1 holds, $E[\{Y - \gamma_0(W)\}^2 | W] \leq \bar{\sigma}^2$ and $\|\alpha_0^{\min}\|_\infty \leq \bar{\alpha}$. Then with probability $1 - \epsilon$,*

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

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 upon updating Δ to

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

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

Theorem 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. Here $\bar{\alpha}'$ is a bound on the chosen estimator $\hat{\alpha}_\ell$ that can be imposed by censoring extreme evaluations. Theorem 1 is a finite-sample refinement of the asymptotic black-box result in Chernozhukov et al. (2022c).

In the bound Δ , the expression $\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}$ allows a trade-off: 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 side-steps ill-posedness. Moreover, the analyst only needs $\mathcal{P}(\hat{\gamma}_\ell)$ to be fast enough to compensate for the ill-posedness encoded in $\mathcal{R}(\hat{\alpha}_\ell)$, or $\mathcal{P}(\hat{\alpha}_\ell)$ to be 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, Kallus et al. (2021) documented an asymptotic special case of this result for a specific global functional and specific nuisance estimators; see the [Supplementary Material](#).

By Theorem 1, the neighbourhood 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 is provided in the [Supplementary Material](#) and summarized here. It turns out that global functionals have a 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 2 (CHARACTERIZATION OF KEY QUANTITIES). *If the noise has finite variance, then $\bar{\sigma}^2 < \infty$. Suppose that the bounded-moment and heteroscedasticity conditions in the [Supplementary Material](#) 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 that the bounded-moment, heteroscedasticity, density and derivative conditions in the [Supplementary Material](#) 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 $\Delta_h \lesssim n^{1/2}h^{\nu+1/2}$, where ν is the order of differentiability defined in the [Supplementary Material](#).*

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 be bounded away from 0 and 1. For derivatives, a sufficient condition is that Γ should satisfy Sobolev conditions. For local functionals, we

handle $(\bar{Q}_h, \bar{\alpha}_h)$ on a case-by-case basis. See the [Supplementary Material](#) for interpretable and complete characterizations.

Observe that the finite-sample Gaussian approximation in Theorem 1 is in terms of the true asymptotic variance σ^2 . We now provide a guarantee for its estimator $\hat{\sigma}^2$.

THEOREM 3 (VARIANCE ESTIMATION). *Suppose that Assumption 1 holds, $E[\{Y - \gamma_0(W)\}^2 | W] \leq \bar{\sigma}^2$ 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*

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

Theorem 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 1 and 3 immediately imply simple, interpretable conditions for validity of the confidence interval. We conclude by summarizing these conditions.

COROLLARY 1 (CONFIDENCE INTERVAL). *Suppose that Assumption 1 holds, as well as the following regularity and learning-rate conditions as $n \rightarrow \infty$ and as $h \rightarrow 0$:*

- (i) $E[\{Y - \gamma_0(W)\}^2 | W] \leq \bar{\sigma}^2$, $\|\alpha_0^{\min}\|_\infty \leq \bar{\alpha}$, $\|\hat{\alpha}_\ell\|_\infty \leq \bar{\alpha}'$ and $\{(\kappa/\sigma)^3 + \zeta^2\} n^{-1/2} \rightarrow 0$;
- (ii) $(\bar{Q}^{1/2} + \bar{\alpha}/\sigma + \bar{\alpha}') \{\mathcal{R}(\hat{\gamma}_\ell)\}^{q/2} = o_p(1)$;
- (iii) $\bar{\sigma} \{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} = o_p(1)$;
- (iv) $[\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{nP(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \wedge \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{P}(\hat{\alpha}_\ell)\}^{1/2}] / \sigma = o_p(1)$.

Then the estimator $\hat{\theta}$ in Algorithm 1 is consistent and asymptotically Gaussian, and the confidence interval in Algorithm 1 includes θ_0 with probability approaching the nominal level. Formally,

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

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

SUPPLEMENTARY MATERIAL

The [Supplementary Material](#) includes simulation results, details of Theorem 2, further discussion, proofs and code.

REFERENCES

- ABREVAYA, J., HSU, Y.-C. & LIELI, R. P. (2015). Estimating conditional average treatment effects. *J. Bus. Econ. Statist.* **33**, 485–505.
- AI, C. & CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71**, 1795–843.
- ANDREWS, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* **62**, 43–72.
- AVAGYAN, V. & VANSTEELENDT, S. (2023). High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostatist. Epidemiol.* to appear, DOI: 10.1080/24709360.2021.1898730.
- BELLONI, A., CHEN, D., CHERNOZHUKOV, V. & HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–429.
- BELLONI, A., CHERNOZHUKOV, V. & HANSEN, C. (2013). Inference for high-dimensional sparse econometric models. In *Advances in Economics and Econometrics*. Cambridge: Cambridge University Press, pp. 245–95.
- BELLONI, A., CHERNOZHUKOV, V. & KATO, K. (2014). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102**, 77–94.
- BIBAUT, A. F. & VAN DER LAAN, M. J. (2017). Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters. *arXiv*: 1706.07408v2.
- BICKEL, P. J., KLAASSEN, C. A., BICKEL, P. J., RITOV, Y., KLAASSEN, J., WELLNER, J. A. & RITOV, Y. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, Maryland: Johns Hopkins University Press.
- BLUNDELL, R., CHEN, X. & KRISTENSEN, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75**, 1613–69.

- CAI, W. & VAN DER LAAN, M. (2020). Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (LASSO) estimator. *Int. J. Biostatist.* **16**. DOI: 10.1515/ijb-2017-0070.
- CHEN, X. & SANTOS, A. (2018). Overidentification in regular models. *Econometrica* **86**, 1771–817.
- CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. & ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Economet. J.* **21**, C1–68.
- CHERNOZHUKOV, V., ESCANCIANO, J. C., ICHIMURA, H., NEWEY, W. K. & ROBINS, J. M. (2022c). Locally robust semiparametric estimation. *Econometrica* **90**, 1501–35.
- CHERNOZHUKOV, V., NEWEY, W. & SINGH, R. (2022a). Debiased machine learning of global and local parameters using regularized Riesz representers. *Economet. J.* **25**, 576–601.
- CHERNOZHUKOV, V., NEWEY, W. K. & SINGH, R. (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica* **90**, 967–1027.
- CHERNOZHUKOV, V., NEWEY, W., SINGH, R. & SYRGKANIS, V. (2020). Adversarial estimation of Riesz representers. *arXiv*: 2101.00009.
- COLANGELO, K. & LEE, Y.-Y. (2021). Double debiased machine learning nonparametric inference with continuous treatments. *arXiv*: 2004.03036v6.
- DIKKALA, N., LEWIS, G., MACKAY, L. & SYRGKANIS, V. (2020). Minimax estimation of conditional moment models. *arXiv*: 2006.07201.
- FOSTER, D. J. & SYRGKANIS, V. (2020). Orthogonal statistical learning. *arXiv*: 1901.09036v3.
- HASMINSKII, R. Z. & IBRAGIMOV, I. A. (1979). On the nonparametric estimation of functionals. In *Proc. 2nd Prague Symp. on Asymptotic Statistics*. Amsterdam: North-Holland, pp. 41–51.
- ICHIMURA, H. & NEWEY, W. K. (2022). The influence function of semiparametric estimators. *Quant. Econ.* **13**, 29–61.
- JAVANMARD, A. & MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–909.
- KALLUS, N., MAO, X. & UEHARA, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv*: 2103.14029v2.
- KLAASSEN, C. A. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15**, 1548–62.
- LUEDTKE, A. R. & VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44**, 713–42.
- NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62**, 1349–82.
- NEWEY, W. K. & POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–78.
- PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory*, vol. 13 of *Lecture Notes in Statistics*. New York: Springer.
- QIU, H., LUEDTKE, A. & CARONE, M. (2021). Universal sieve-based strategies for efficient estimation using machine learning tools. *Bernoulli* **27**, 2300–36.
- ROBINS, J. M. & ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Assoc.* **90**, 122–9.
- ROBINSON, P. M. (1988). Root- N -consistent semiparametric regression. *Econometrica* **56**, 931–54.
- ROTNITZKY, A., SMUCLER, E. & ROBINS, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika* **108**, 231–8.
- SEVERINI, T. A. & TRIPATHI, G. (2012). Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *J. Economet.* **170**, 491–8.
- SHEVTSOVA, I. (2011). On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv*: 1111.6554.
- SINGH, R. (2021). Debiased kernel methods. *arXiv*: 2102.11076v2.
- SINGH, R., SAHANI, M. & GRETTON, A. (2019). Kernel instrumental variable regression. In *Proc. 33rd Int. Conf. Neural Information Processing Systems (NeurIPS 2019)*. New York: Curran Associates, pp. 4593–605.
- SMUCLER, E., ROTNITZKY, A. & ROBINS, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv*: 1904.03737v3.
- VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.
- VAN DER LAAN, M. (2017). Finite sample inference for targeted learning. *arXiv*: 1708.09502.
- VAN DER LAAN, M. J. & ROSE, S. (2018). *Targeted Learning in Data Science*. Cham, Switzerland: Springer.
- VAN DER LAAN, M. J. & RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostatist.* **2**, article no. 11. DOI: 10.2202/1557-4679.1043.
- ZHANG, C.-H. & ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B* **76**, 217–42.
- ZHENG, W. & VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*. New York: Springer, pp. 459–74.

[Received on 31 May 2021. Editorial decision on 19 May 2022]