# AUTOMATIC DEBIASED MACHINE LEARNING VIA RIESZ REGRESSION

BY VICTOR CHERNOZHUKOV[1,a], WHITNEY K. NEWEY[1,b]
VÍCTOR QUINTAS-MARTÍNEZ[1,c] AND VASILIS SYRGKANIS[2,d]

[1]*Department of Economics, MIT,* [a]*vchern@mit.edu;* [b]*wnewey@mit.edu;* [c]*vquintas@mit.edu*

[2]*Department of Management Science and Engineering, Stanford University,* [d]*vsyrgk@stanford.edu*

A variety of interesting parameters may depend on high dimensional regressions. Machine learning can be used to estimate such parameters. However estimators based on machine learners can be severely biased by regularization and/or model selection. Debiased machine learning uses Neyman orthogonal estimating equations to reduce such biases. Debiased machine learning generally requires estimation of unknown Riesz representers. A primary innovation of this paper is to provide Riesz regression estimators of Riesz representers that depend on the parameter of interest, rather than explicit formulae, and that can employ any machine learner, including neural nets and random forests. End-to-end algorithms emerge where the researcher chooses the parameter of interest and the machine learner and the debiasing follows automatically. Another innovation here is debiased machine learners of parameters depending on generalized regressions, including high-dimensional generalized linear models. An empirical example of automatic debiased machine learning using neural nets is given. We find in Monte Carlo examples that automatic debiasing sometimes performs better than debiasing via inverse propensity scores and never worse. Finite sample mean square error bounds for Riesz regression estimators and asymptotic theory are also given.

**1. Introduction.** Many parameters of interest depend on regressions. Examples include treatment effects, regression decompositions, and policy effects. Often, a regression may be high dimensional, depending on many variables. For example there may be many covariates for treatment effects. Machine learning methods such as neural nets, random forests, and Lasso can be used to estimate parameters of interest that depend on high dimensional regressions.

A general problem with estimating parameters of interest using machine learning is that machine learners are biased by regularization and/or model selection. This bias may pass through when the learner is plugged into a formula for a parameter of interest and make the parameter estimator highly biased. This problem can be avoided by using Neyman orthogonal estimating equations where machine learners have zero first-order effect. Cross-fitting, a form of sample splitting, can also help.

The orthogonal estimating equations for regressions depend on a Riesz representer $\alpha_0$ that must be estimated. The primary innovation of this paper is to provide an automatic estimator of $\alpha_0$ that uses only the definition of the parameter of interest and the regression but does not require knowing a formula for $\alpha_0$. We give an objective function with expectation that is minimized at $\alpha_0$ that depends only the parameter of interest. We refer to minimization of this objective function as a Riesz regression, being equivalent to minimizing the expected squared deviation from $\alpha_0$. Neural nets, random forests, and other methods can be used for

*MSC2020 subject classifications:* Primary 62D20, 62P20; secondary 62G20, 62J02.

*Keywords and phrases:* Debiased Machine Learning, Generalized Linear Models, Riesz Representers, Neural Nets.

this Riesz regression. Using the Riesz regression estimator in the bias correction completes an algorithm that 1) specifies the parameter of interest; 2) specifies a learner of the unknown regression; and 3) uses the Riesz regression estimator of $\alpha_0$ determined by steps 1) and 2).

A second innovation of this paper is to construct and derive properties of estimators that depend on generalized regressions, which minimize an expected loss over some linear set of functions. These generalized regressions include conditional means, least squares projections, functions that minimize quasi-likelihoods, and quantile regressions. Debiasing for generalized regressions depends on a weighted version of the Riesz representer. We give a weighted Riesz regression that only uses the parameter of interest and the generalized regression for bias correction.

A third contribution of this paper is finite sample mean square error bounds for Reisz regressions. These bounds are obtained using the critical radius of functions of $\alpha$ on which the objective function depends and approximation error bounds for the unknown $\alpha_0$. A fourth contribution is convergence rates for neural net Riesz regressions. These are based on known results on critical radius and approximation error for neural nets and the finite sample bounds given here.

In work that followed up on the first version of this paper (Chernozhukov et al. (2022a)) we found that using Riesz regressions to debias neural net and random forest estimators of the average treatment effect was much more accurate than state of the art methods based on inverse propensity score weighting, in a Monte Carlo study. Both the automatic neural net and random forest debiasing also led to accurate confidence intervals in those experiments.

Automatic debiasing for Lasso and reproducing kernel Hilbert space regressions was previously given by Chernozhukov, Newey and Singh (2022) and Singh, Xu and Gretton (2022) respectively. The estimator of $\alpha_0$ given here goes beyond these to provide automatically debiasing for generalized regressions based on neural nets, random forests, and other machine learners. These innovations allow researchers to use any of a wide variety of automatically debiased machine learners learners to estimate parameters of interest that depend on generalized regressions. For example, automatic debiased machine learning with neural nets could be especially useful for parameters that depend on high dimensional, nonlinear generalized regressions.

This paper builds on recent work on Neyman orthogonal scores and debiased machine learning. We use model free orthogonal estimating equations like those of Chernozhukov et al. (2022b) that are the sum of an identifying moment function and a bias adjustment (influence function) term for generalized regression from Ichimura and Newey (2022). Those papers did not give the Riesz regression. Finite sample mean square error bounds for a general learner of $\alpha_0$ are obtained by applying results of Foster and Syrgkanis (2019) that characterize error bounds in terms of critical radius and approximation. The rate of convergence for neural net Riesz regression use critical radius and approximation rate results given in Farrell, Liang and Misra (2021a). Additional neural net rate conditions could be obtained using Yarotsky (2018). The learner of $\alpha_0$ differs from those of Farrell, Liang and Misra (2021a,b) in using the Riesz regression rather than a known form for $\alpha_0$.

We also build upon ideas in classical semi- and nonparametric learning theory with low dimensional regressions using traditional smoothing methods (Van Der Vaart (1991); Bickel et al. (1993); Newey (1994); Robins and Rotnitzky (1995); Van der Vaart (2000)), that do not apply to machine learners. The orthogonal estimating equations given in Chernozhukov et al. (2022b) and used here build on previous work on nonparametric orthogonal moment functions by Levit (1975); Hasminskii and Ibragimov (1978); Bickel and Ritov (1988); Newey, Hsieh and Robins (2004). Targeted maximum likelihood (Van Der Laan and Rubin (2006)) based on machine learners has been considered by Van der Laan and Rose (2011) and large sample theory given by Luedtke and van der Laan (2016).

In section 2 we give the Reisz regression and an automatic debiased machine learning algorithm for parameters that depend linearly on a nonparametric regression, including examples. Section 2 also gives finite sample mean square error bounds for the general and neural net Riesz regression and asymptotic inference results for parameters that are linear functionals of a nonparametric regression. Section 3.2 extends the estimation methods and theory to nonlinear functionals of generalized regressions. In section 4 we illustrate the usefulness of our methods with an empirical application. Section 5 presents the results of our simulation exercises.

## 2. Average Linear Effects for a Conditional Mean.

To highlight the innovation provided by the Riesz regression, we first consider average linear effects that depend on a conditional mean. In section 3 we consider the general setting of nonlinear functions of generalized regressions.

### 2.1. *Parameters of Interest.*

We consider data that consists of i.i.d. observations $W_1, \ldots,$ $W_n$, each having CDF $F_0$. A data observation $W$ includes an outcome variable $Y$ and regressors $X$. In this section, we focus on parameters that depend on the conditional mean of $Y$ given $X$. We will denote a possible such regression function by $\gamma$, with $\gamma_0(x) = \mathrm{E}[Y \mid X = x]$ being the true regression function.

The parameter of interest $\theta_0$ has the form

$$\theta_0 = \mathrm{E}[m(W, \gamma_0)], \tag{2.1}$$

where $m(w, \gamma)$ is a functional that depends on a data observation $w$ and a possible regression function $\gamma$. For now, we assume that the expected functional $\gamma \mapsto \mathrm{E}[m(W, \gamma)]$ is linear and continuous in $\gamma$, meaning that there is a constant $C > 0$ with $|\mathrm{E}[m(W, \gamma)]|^2 \leq C\,\mathrm{E}[\gamma(X)^2]$ for all $\gamma$ with $\mathrm{E}[\gamma(X)^2] < \infty$. Under this assumption, there exists a function $v_m$ with $\mathrm{E}[v_m(X)^2] < \infty$ such that

$$\mathrm{E}[m(W, \gamma)] = \mathrm{E}[v_m(X)\gamma(X)] \quad \text{for all } \gamma \text{ with } \mathrm{E}[\gamma(X)^2] < \infty. \tag{2.2}$$

The existence of this $v_m$ follows from the Riesz representation theorem, and it is equivalent to the semiparametric variance bound for $\theta_0$ being finite (see Newey (1994); Hirshberg and Wager (2021); Chernozhukov, Newey and Singh (2019)). For these reasons $v_m$ is often referred to as the Riesz representer. In this Section, where the parameter of interest depends on a nonparametric regression, the Riesz representer $v_m = \alpha_0$ needs to be estimated for debiased machine learning. In section 3, where $\gamma$ may be a generalized regression, $\alpha_0$ will be a weighted version of the Riesz regression.

There are a variety of important, empirically relevant parameters of interest that have this form. We illustrate with some familiar examples to help highlight and motivate the Riesz regression:

EXAMPLE 1 (Average Treatment Effect ). Suppose that $X = (D, Z)$ where $D$ is a binary treatment indicator, and $Z$ are covariates. The parameter of interest is $\theta_0$ in equation (2.1) with

$$m(W, \gamma) = \gamma(1, Z) - \gamma(0, Z).$$

If $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$, where potential outcomes $(Y(1), Y(0))$ are conditionally independent of treatment $D$ given covariates $Z$, then this object is the Average Treatment Effect or ATE (Rosenbaum and Rubin (1983)). In this example the Riesz representer is

$$\alpha_0(X) = \frac{D}{\pi_0(Z)} - \frac{1 - D}{1 - \pi_0(Z)},$$

where $\pi_0(z) = \Pr(D = 1 \mid Z = z)$ is the propensity score. Here $\alpha_0(X)$ is the difference of the Horvitz and Thompson (1952) weights for treated and untreated and $\mathrm{E}[\alpha_0(X)^2] < \infty$ if and only if $\mathrm{E}[\pi_0(Z)^{-1}(1-\pi_0(Z))^{-1}] < \infty$.

EXAMPLE 2 (Average Marginal Effect). Suppose again that $X = (D, Z)$ where $D$ is now a continuous treatment or policy variable, and $Z$ are covariates. The parameter of interest is $\theta_0$ in equation (2.1) with

$$m(W, \gamma) = \partial_d \gamma(X),$$

where we denote $\partial_d g(x) = \partial g(x)/\partial d$ for any function $g$. This object can be interpreted as an average treatment effect for continuous treatment $D$, see Imbens and Newey (2009). Here the Riesz representer is

$$\alpha_0(X) = -\partial_d \ln f_0(X),$$

where $f_0(X)$ is the (true) joint probability density function (pdf) of $X$. Here equation (2.2) follows by integrating by parts, and then multiplying and dividing by $f_0(x)$.

EXAMPLE 3 (Average Policy Effect). Suppose that $\gamma_0$ does not vary with the distribution of $X$. Then, the average effect of a conterfactual shift in the distribution of regressors, from a known distribution with pdf $g_0$ to another known distribution with pdf $g_1$ is the $\theta_0$ of equation (2.1) with

$$m(W, \gamma) = \int \gamma(x)(g_1(x) - g_0(x))dx = \mathrm{E}\left[\frac{g_1(X) - g_0(X)}{f_0(X)}\gamma(X)\right],$$

where $f_0(x)$ is the (true) pdf of $X$ in the data. Here the Riesz representer is

$$\alpha_0(X) = \frac{g_1(X) - g_0(X)}{f_0(X)},$$

with equation (2.2) following from the second equality in the expression for $m(W, \gamma)$.

2.2. *Estimation* . We will base estimation of $\theta_0$ on a Neyman orthogonal estimating equation, i.e. score, where first step estimation has zero first order effects, that is also doubly robust in having expectation zero if either $\gamma = \gamma_0$ or $\alpha = \alpha_0$. This score is

$$\psi(w, \gamma, \alpha, \theta) = m(w, \gamma) - \theta + \alpha(x)(y - \gamma(x)), \tag{2.3}$$

as in Chernozhukov et al. (2022b), where taking expectations gives, for any $\alpha, \gamma$,

$$\begin{aligned} \mathrm{E}[\psi(W, \gamma, \alpha, \theta_0)] &= \mathrm{E}[m(W, \gamma)] - \theta_0 + \mathrm{E}[\alpha(X)(Y - \gamma(X))] \\ &= \mathrm{E}[m(W, \gamma - \gamma_0)] - \mathrm{E}[\alpha(X)(\gamma(X) - \gamma_0(X))] \\ &= -\mathrm{E}[(\alpha(X) - \alpha_0(X))(\gamma(X) - \gamma_0(X))]. \qquad (2.4) \end{aligned}$$

Here we see that at the true parameter value $\theta_0$, the expectation of the score $\psi(W, \gamma, \alpha, \theta_0)$ differs from zero only to second order and equals zero if either $\gamma = \gamma_0$ or $\alpha = \alpha_0$. Thus the score is Neyman orthogonal and doubly robust in that it has zero expectation if either $\gamma = \gamma_0$ or $\alpha = \alpha_0$.

Estimation of $\alpha_0$ is essential to construction of a debiased machine learner of the parameter of interest. The primary innovation of this paper is to give an extremum characterization of

$\alpha_0$ and use this to estimate $\alpha_0$. This extremum characterization is given by

$$\begin{aligned}
\alpha_0 &= \arg\min_{\alpha} \mathrm{E}[(\alpha_0(X) - \alpha(X))^2] \\
&= \arg\min_{\alpha} \mathrm{E}[\alpha_0(X)^2 - 2\alpha_0(X)\alpha(X) + \alpha(X)^2] \\
&= \arg\min_{\alpha}\{-2\,\mathrm{E}[v_m(X)\alpha(X)] + \mathrm{E}[\alpha(X)^2]\} \\
&= \arg\min_{\alpha} \mathrm{E}[-2m(W,\alpha) + \alpha(X)^2], \qquad (2.5)
\end{aligned}$$

where the third equality holds because $\mathrm{E}[\alpha_0(X)^2]$ does not depend on $\alpha$ and $\alpha_0 = v_m$, and the fourth equality follows from equation (2.2) with $\gamma = \alpha$. This characterization can be used to estimate $\alpha_0$ by replacing the expectation with the sample average and minimizing over some set of possible $\alpha$ functions.

We call a resulting estimator of $\alpha_0$ a Riesz regression, motivated by minimization of the least squares objective function in equation (2.5). This estimator is *automatic* in dependng only on the function $m(w,\gamma)$ that determines the parameter of interest and in not requiring the form of $\alpha_0$. In particular, this method does not depend on plugging in non-parametric estimates of components of $\alpha_0$. This feature is useful when $\alpha_0$ does not have a simple form. For causal parameters such as those of Examples 1-3, the Riesz regression avoids inverting a learner of a conditional probability or a pdf. Instead, the Riesz regression learns $\alpha_0$ directly.

Our estimation strategy for the parameter of interest is to combine the Reisz regression estimator of $\alpha_0$ and estimation of $\gamma_0$ in the Neyman orthogonal score with the use of cross-fitting to reduce overfitting bias.[1] The outline of our estimation strategy is as follows:

1. Partition the set of data indices $1,\ldots,n$ into $L$ disjoint subsets of about equal size $I_\ell$, $\ell = 1,\ldots,L$;
2. For each data fold $\ell = 1,\ldots,L$:
   a) Estimate $\hat{\gamma}_\ell \in \mathcal{G}_n$ as a non-parametric regression of $Y$ on $X$ over some class of functions $\mathcal{G}_n$ using observations *not* in $I_\ell$.
   b) Estimate the debiasing function $\hat{\alpha}_\ell$ using observations *not* in $I_\ell$ by minimizing a sample version of the objective function in equation (2.5) over a set of functions, as in

   $$\hat{\alpha}_\ell = \arg\min_{\alpha \in \mathcal{A}_n} \left[ \sum_{i \notin I_\ell} \left\{ -2m(W_i,\alpha) + \alpha(X_i)^2 \right\} + \Lambda_r(\alpha) \right]$$

   Where $\Lambda_r(\alpha)$ is a penalty term and $r$ is a scalar determining the magnitude of penalization.
3. Estimate the parameter of interest using the cross-fitted regression and debiasing function in the moment function of equation (2.3) to obtain

   $$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)(Y_i - \hat{\gamma}_\ell(X_i))\}$$

4. Estimate the standard error of $\hat{\theta}$ as $\sqrt{\hat{V}/n}$, where:

   $$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)(Y_i - \hat{\gamma}_\ell(X_i)) - \hat{\theta}\right\}^2$$

[1] See Newey and Robins (2018) for more on the advantages of cross-fitting.

Our estimation strategy is very general, allowing for any choice of learner $\hat{\gamma}_\ell$ and any Riesz regression $\hat{\alpha}_\ell$ encoded in the class of functions $\mathcal{A}_n$. Special kinds of Riesz regressions have been given in previous literature. These include linear combinations of a dictionary of functions $(b_1(x), \dots, b_p(x))'$, and $p$ large, with an $L_1$ penalty in the loss function (Chernozhukov, Newey and Singh (2022)), or functions embedded in a reproducing kernel Hilbert space (Singh (2021)). Chernozhukov et al. (2020) allowed for the estimation of $\alpha_0$ in arbitrary function spaces, but proposed a computationally harder minimax loss formulation. A primary innovation of this paper is to provide the Riesz regression for automatic estimation of $\alpha_0$ and corresponding asymptotic theory.

As an example, below we will give primitive conditions for a neural net Riesz regression. A general neural net takes the form

$$x \xmapsto{f_1} H^{(1)} \xmapsto{f_2} \cdots \xmapsto{f_m} H^{(m)}$$

where $H^{(l)} = \{H_k^{(l)}\}_{k=1}^{K_l}$ are called neurons, $x$ is the original finite-dimensional input, and the function $f_l$ maps one layer of neurons to the next as in

$$f_l : v \xmapsto{f_l} \{H_k^{(l)}(v)\}_{k=1}^{K_l} := (1, \{\sigma(v'\beta_{k,l})\}_{k=2}^{K_l}),$$

where each $\beta_{k,l}$ is a $K_{l-1}$ vector of parameters and $\sigma(u)$ is a nonlinear activation function. We will focus on the case where $\sigma(u)$ is the RELU function $\sigma(u) = \max\{0, u\}$. An important special case is a multilayer perceptron (MLP) network where the number of neurons $K_l = K$ is the same for each layer, for which results were recently given by Farrell, Liang and Misra (2021a). Sparse versions of this specification, where many of the elements of the coefficient vectors $\beta_{k,l}$ may be zero, have also been considered recently by Schmidt-Hieber (2020). Yarotsky (2018) gave other neural net specifications with good approximation properties.

In the setting of Example 1, a neural net Riesz regression would be constructed as $\hat{\alpha}_\ell(d, z) = \alpha(d, z; \hat{\beta}_\ell)$, where:

$$\hat{\beta}_\ell = \arg\min_\beta \left[ \sum_{i \notin I_\ell} \left\{ -2[\alpha(1, Z_i; \beta) - \alpha(0, Z_i; \beta)] + \alpha(D_i, Z_i; \beta)^2 \right\} + \Lambda_r(\beta) \right]$$

for some penalty function $\Lambda_r(\beta)$ (e.g., L1, L2, or the elastic net). Because $D$ is binary, a convenient neural net architecture in this case could be a bi-headed MLP, $\alpha(d, z; \beta, \delta_0, \delta_1) = dg(z;\beta)'\delta_1 + (1-d)g(z;\beta)'\delta_0$, where $g(z;\beta)$ is an MLP. An even more flexible specification would be to have $\alpha(d, z; \beta_0, \beta_1) = dg(z, \beta_0) + (1-d)g(z, \beta_1)$, i.e. an MLP for the case $d = 1$ and another MLP for the case $d = 0$.

For Example 2, a neural net Riesz regression is $\hat{\alpha}_\ell(d, z) = \alpha(d, z; \hat{\beta}_\ell)$, where:

$$\hat{\beta}_\ell = \arg\min_\beta \left[ \sum_{i \notin I_\ell} \left\{ -2[\partial_d \alpha(D_i, Z_i; \beta)] + \alpha(D_i, Z_i; \beta)^2 \right\} + \Lambda_r(\beta) \right].$$

In particular, notice that the loss function involves taking a derivative of the neural net with respect to one of the inputs. A convenient parametrization of the neural net in this case is a locally linear function $\alpha(d, z; \phi, \beta) = \phi(d, z)'g(z; \beta)$, where $\phi(d, z)$ is a dictionary of known, differentiable basis functions, and $g(z; \beta)$ is a neural net (e.g. an MLP). In that case, $\partial_d \alpha(d, z; \beta) = [\partial_d \phi(d, x)]'g(z; \beta)$. This approach was used in Chernozhukov et al. (2022a) to construct a random forest estimator of $\alpha_0$, exhibiting good performance in Monte Carlo simulations.

2.3. *Large Sample Inference for Linear Effects of Regression.* In this Section, we give mean square convergence rates for learners $\hat{\alpha}_\ell$ and $\sqrt{n}$-consistency and asymptotic normality results for the learner $\hat{\theta}$ of the object of interest and its asymptotic variance estimator $\hat{V}$. We first derive convergence rates for $\hat{\alpha}_\ell$.

2.3.1. *Convergence Rates for $\hat{\alpha}_\ell$ .* In this subsection we suppress the $\ell$ subscript for notational convenience. We consider the problem of estimating

$$\alpha_0 = \arg\min_{\alpha} \mathrm{E}[-2m(W,\alpha) + \alpha(X)^2],$$

where we have used the extremum characterization of $\alpha_0$ in equation (2.5). For any random variable $a(W)$ let $\|a\| = \sqrt{\mathrm{E}[a(W)^2]}$ and $\|a\|_\infty = \sup_{w\in\mathcal{W}} |a(w)|$ . For simplicity of exposition we will only consider the case where the estimate is defined over a growing sieve space $\mathcal{A}_n$ and no regularization is used, i.e. $\Lambda_r(\alpha) = 0$:

$$\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}_n} \sum_{i=1}^n \left\{-2m(W_i,\alpha) + \alpha(X_i)^2\right\}, \tag{2.6}$$

Our estimation rate can easily be extended to regularized estimation with appropriate regularization weight. We assume that $m(W,\alpha)$ is mean square continuous in the following sense:

ASSUMPTION 1. For some $M > 0$ it is the case that $\mathrm{E}[m(W,\alpha)^2] \le M\|\alpha\|^2$.

Define:

$$\begin{aligned} \mathrm{star}(\mathcal{A}_n - \alpha_0) &= \{x \to \xi\,(\alpha(x) - \alpha_0(x)) : \alpha \in \mathcal{A}_n,\ \xi \in [0,1]\} \\ \mathrm{star}(m\circ\mathcal{A}_n - m\circ\alpha_0) &= \{w \to \xi\,(m(W,\alpha) - m(W,\alpha_0)) : \alpha \in \mathcal{A}_n,\ \xi \in [0,1]\} \end{aligned}$$

ASSUMPTION 2. $\|f\|_\infty \le 1$ for all $f \in \mathrm{star}(\mathcal{A}_n - \alpha_0)$ and $f \in \mathrm{star}(m\circ\mathcal{A}_n - m\circ\alpha_0)$.

Define:

$$\alpha^* = \arg\min_{\alpha\in\mathcal{A}_n} \mathrm{E}[-2m(W,\alpha) + \alpha(X)^2]$$

to be the best approximation of $\alpha_0$ by an element of $\mathcal{A}_n$.

THEOREM 2.1. *Let $\delta_n$ be an upper bound on the critical radius of* $\mathrm{star}(\mathcal{A}_n - \alpha_0)$ *and* $\mathrm{star}(m\circ\mathcal{A}_n - m\circ\alpha_0)$. *If Assumptions 1 and 2 are satisfied then it follows that with probability $1-\zeta$, for some universal constant $C$,*

$$\|\hat{\alpha} - \alpha_0\|^2 \le C\left(M\delta_n^2 + \|\alpha^* - \alpha_0\|^2 + \frac{M\ln(1/\zeta)}{n}\right).$$

See e.g. Foster and Syrgkanis (2019) for the definition of the critical radius used in the statement this result. To use Theorem 2.1 to obtain a mean square convergence rate for $\hat{\alpha}$ it is important to know the critical radius and the rate at which $\|\alpha_* - \alpha_0\|$ shrinks as the approximating set $\mathcal{A}_n$ becomes richer. For example, Farrell, Liang and Misra (2021a) have recently obtained such results for deep, ReLU neural nets. We can apply their results to obtain a mean square rate for such a learner of $\alpha_0$ when $x$ is a $d$ dimensional input for the multilayer perceptron (MLP) network with $m$ layers and width $K$.

The convergence rate depends on the smoothness of the function $\alpha_0(x)$, as specified in the following result. Specifically we assume that the support of $X$ is contained in a Cartesian product $\mathcal{X}$ of compact intervals and $\alpha_0(X)$ can be extended to a function that is continuously differentiable on $\mathcal{X}$ and has $\beta$ continuous derivatives.

COROLLARY 2.2. *If (i) the support of $X$ is contained in a Cartesian product of compact intervals and $\alpha_0(X)$ can be extended to a function that is continuously differentiable with $\beta$ continuous derivatives; (ii) $\mathcal{A}_n$ is an MLP network with $d$ inputs, width $K$, and depth $m$ with $K \to \infty$ and $m \to \infty$; (iii) $m \circ \mathcal{A}_n$ is representable as such a network; then there is $C > 0$ such that, for any $\varepsilon > 0$,*

$$\|\hat{\alpha} - \alpha_0\|^2 = O_p(K^2 m^2 \ln(K^2 m)\ln(n)/n + [Km\sqrt{\ln(K^2 m)}]^{-2(\beta/d)+\varepsilon}).$$

When $\alpha_0$ is smooth enough, in that $\beta$ is large enough, the upper bound on $\|\hat{\alpha} - \alpha_0\|$ in Corollary 2.2 gives a mean square convergence rate that can be close to, but less than $n^{-1/2}$. Such rate can be obtained by choosing the width $K$ and depth $m$ to approximately balance the two terms in Corollary 2.2, $K \asymp n^{\frac{d}{2(\beta+d)}} \ln^2(n)$, $m \asymp \ln(n)$, as in Farrell, Liang and Misra (2021a), in which case

$$\|\hat{\alpha} - \alpha_0\|^2 = O_p\left(n^{-\frac{\beta}{\beta+d}} \ln^8(n)\right).$$

Faster rates could be obtained using the neural nets of Yarotsky (2018) or the sparse neural nets of Schmidt-Hieber (2020). We focus on Corollary 2.2 for the MLP neural net because it is a widely used architecture in practice, and because the rates obtained are fast enough for the estimators of the parameter of interest to be asymptotically normal.

2.3.2. *Large Sample Inference for $\theta_0$.* We use additional regularity conditions to show asymptotic normality of $\hat{\theta}$ and consistency of the asymptotic variance estimator $\hat{V}$. We will first give a general result for $\hat{\theta}$ that applies to any $\hat{\alpha}_\ell$ and does not rely on Theorem 2.1 for a convergence rate for $\hat{\alpha}_\ell$. Similarly, any regression learner $\hat{\gamma}_\ell$ can be used here as long as its mean-square convergence rate is fast enough, as formalized below. Such convergence rate results are available for shallow (Chen and White (1999)) and deep (Yarotsky (2018); Schmidt-Hieber (2020); Farrell, Liang and Misra (2021a)) neural nets, random forests (Syrgkanis and Zampetakis (2020)), LASSO (Bickel, Ritov and Tsybakov (2009)), boosting (Luo, Spindler and Kück (2022)) and other high-dimensional methods.

The following assumption imposes a few additional regularity conditions. Let $\sigma_0^2(X) = \mathrm{E}[(Y - \gamma_0(X))^2 \mid X]$ denote the conditional variance of $Y$ given $X$.

ASSUMPTION 3. $\alpha_0(X)$ and $\sigma_0^2(X)$ are bounded and $\mathrm{E}[m(W, \gamma_0)^2] < \infty$.

Next, we require mean square consistency of $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$, that the product of their mean-square convergence rates is smaller than $n^{-1/2}$, and a boundedness condition for $\hat{\alpha}_\ell$.

ASSUMPTION 4. (i) $\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0$ and $\|\hat{\alpha}_\ell - \alpha_0\| \xrightarrow{p} 0$; (ii) $\sqrt{n}\,\|\hat{\gamma}_\ell - \gamma_0\|\,\|\hat{\alpha}_\ell - \alpha_0\| \xrightarrow{p} 0$; (iii) $\hat{\alpha}_\ell(X)$ is bounded.

Part (i) implies that both $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$ are consistent in mean square. Part (ii) captures an important tradeoff between the rates of convergence for $\hat{\gamma}_\ell$ and $\hat{\alpha}_\ell$. In settings where the regression can be estimated at a relatively fast rate of convergence, the learner for the debiasing function can converge more slowly, and vice versa, as long as the product of their mean-square convergence rates vanishes faster than $n^{-1/2}$. The results we have obtained for the neural net learner $\hat{\alpha}_\ell$ can be used to verify these conditions and we do so in Corollary 2.4 to follow. The mean square convergence of $\hat{\gamma}_\ell$ is a primitive condition for this paper and allows use of a wide variety of $\hat{\gamma}_\ell$ in the construction of the estimator.

We have the following large sample inference result under these conditions.

THEOREM 2.3. *If Assumptions 1, 3 and 4 are satisfied, then*

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,V) \quad \textit{and} \quad \hat{V} \xrightarrow{p} V,$$

*where $\hat{V}$ is the variance estimator defined in subsection 2.2 and $V = \mathrm{E}[\{m(W,\gamma_0) - \theta_0 + \alpha_0(X)(Y - \gamma_0(X))\}^2]$.*

Next we use Theorem 2.1 and Corollary 2.2 to formulate regularity conditions when $\hat{\alpha}_\ell$ is the neural net learner of $\alpha_0$ in section 2.2. Let

$$\epsilon_{\alpha n}^2 = K^2 m^2 \ln(K^2 m) \ln(n)/n + [Km\sqrt{\ln(K^2 m)}]^{-2(\beta/d)+\varepsilon}.$$

This $\epsilon_{\alpha n}^2$ is taken from the upper bound for $\|\hat{\alpha}_\ell - \alpha_0\|^2$ in Corollary 2.2 and so characterizes the mean square convergence rate of the automatic neural net learner $\hat{\alpha}_\ell$.

COROLLARY 2.4. *Suppose that Assumptions 1, 2, 3 and the hypotheses of Corollary 2.2 hold. Moreover, suppose that $\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0$ , $\epsilon_{an} \to 0$ and $\sqrt{n}\,\|\hat{\gamma}_\ell - \gamma_0\|\,\epsilon_{an} \xrightarrow{p} 0$. Then, for the neural net learner $\hat{\alpha}_\ell$ of Corollary 2.2, we have*

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{p} N(0,V) \quad \textit{and} \quad \hat{V} \xrightarrow{p} V.$$

## 3. Average Effects for Generalized Regressions.

3.1. *Linear Effects.* In this section we extend the results to parameters that depend on functions $\gamma_0$ other than the conditional mean, that we refer to as *generalized regressions*. Suppose that $\gamma_0$ is defined as the solution to a general $M$-estimation problem:

$$\gamma_0 := \arg\min_{\gamma\in\Gamma} \mathrm{E}[\ell(W,\gamma)], \tag{3.1}$$

where $\Gamma$ is a closed (in mean square) linear subspace of $L^2(X)$. For example, when $\ell(W,\gamma) = \frac{1}{2}(Y-\gamma(X))^2$ is the square loss and $\Gamma = L^2(X)$, then $\gamma_0(X) = \mathrm{E}[Y \mid X]$ and we recover the case of regression.

By the first order condition of the minimization problem (3.1), $\gamma_0$ satisfies

$$\mathrm{E}[\rho(W,\gamma_0)b(X)] = 0 \quad \text{for all } b \in \Gamma. \tag{3.2}$$

for some functional $\rho(W,\gamma)$, typically a generalized notion of the (negative) derivative of the loss function $\ell(W,\gamma)$. In the case of regression, we can take $\rho(W,\gamma) = Y - \gamma(X)$ to be the non-parametric residual. For other statistical problems, we will refer to the function $\rho(W,\gamma)$ as a *generalized residual*. The results of this section will apply to any $\gamma_0$ that is identified by an orthogonality condition as in (3.2), even beyond $M$-estimation problems.

This setting covers many interesting features of the conditional distribution of $Y$ given $X$. First, suppose that $\Gamma = L^2(X)$, so that the functional form of $\gamma_0$ is unrestricted. For example, when $\rho(W,\gamma) = \tau - 1(Y < \gamma(X))$ for $0 < \tau < 1$, then $\gamma_0(x)$ is the $\tau$-th conditional quantile of $Y$ given $X = x$. When $\rho(W,\gamma) = \lambda(\gamma(X))[Y - \mu(\gamma(X))]$ for a link function $\mu(a)$ and another function $\lambda(a)$, this corresponds to the first order conditions of a generalized linear model (Nelder and Wedderburn, 1972). For binary $Y \in \{0,1\}$, $\mu(a)$ the standard logistic CDF, and $\lambda(a) \equiv 1$, for instance, this set up corresponds to a (non-parametric) logistic regression, where $\gamma_0(X) = \mu^{-1}(\Pr(Y=1 \mid X)) = \ln(\Pr(Y=1 \mid X)/\Pr(Y=0 \mid X))$ corresponds to the log-odds.

The set $\Gamma$ could be used to encode parametric or semi-parametric restrictions on $\gamma_0$. One example is $X = (X_1, X_2, ...)$ and $\Gamma$ the mean square closure of finite linear combinations of $X$. This corresponds to a high (infinite) dimensional, approximately sparse $\Gamma$, where the orthogonality condition is equivalent to $\mathrm{E}[X_j\rho(W,\gamma_0)] = 0$ for all $j$. In a case considered also by Hirshberg and Wager (2021) and Farrell, Liang and Misra (2021a), $\Gamma$ is the mean square closure of $\{a(X_1) + X_2'b(X_1)\}$ where $a(X_1)$ is a scalar function and $b(X_1)$ a vector of functions, each having unrestricted functional form. We could also take $\Gamma$ to be the mean-square closure of additive functions $a_1(X_1) + a_2(X_2)$, where $X_1$ and $X_2$ are distinct components of $X$. In some cases, the resulting $\gamma_0$ will have a projection interpretation: for instance, when $\ell(W,\gamma)$ is the square loss, then $\gamma_0 = \arg\min_{\gamma\in\Gamma} \mathrm{E}[(\mathrm{E}[Y \mid X] - \gamma(X))^2]$ is the best approximation to $\mathrm{E}[Y \mid X]$ in $\Gamma$ in the mean-square sense.

For now, we also continue to assume that the parameter of interest has the form $\theta_0 = \mathrm{E}[m(W,\gamma_0)]$, where $\gamma \mapsto m(W,\gamma)$ is linear and $\mathrm{E}[m(W,\gamma)]$ is mean square continuous on $\Gamma$. We will relax the linearity assumption in section 3.2. We will extend the results of the previous section by modifying (2.3). Define, for any $\gamma,\alpha \in \Gamma$, a score:

$$\psi(w,\gamma,\alpha,\theta) = m(w,\gamma) - \theta + \alpha(x)\rho(w,\gamma), \tag{3.3}$$

where we have replaced $y - \gamma(x)$ with the generalized residual $\rho(W,\gamma)$.

This score satisfies $\mathrm{E}[\psi(W,\gamma_0,\alpha,\theta_0)] = \mathrm{E}[\alpha(X)\rho(W,\gamma_0)] = 0$ for any $\alpha \in \Gamma$ by (3.2), and hence it is Neyman-orthogonal with respect to $\alpha$. To get Neyman orthogonality with respect to $\gamma$ we need to find a function $\alpha_0$ that satisfies

$$\frac{\partial}{\partial r}\mathrm{E}[\psi(W,\gamma_0 + r\delta,\alpha_0)]\bigg|_{r=0} = \mathrm{E}[\{v_m(X) + \alpha_0(X)\,v_\rho(X)\}\delta(X)] = 0 \quad \forall\delta\in\Gamma, \tag{3.4}$$

where $v_m(X)$ is the Riesz representer in (2.2) and, for a scalar $a$,

$$v_\rho(X) := \frac{\partial}{\partial a}\mathrm{E}[\rho(W,\gamma_0 + a) \mid X]\bigg|_{a=0},$$

that we assume exists. We further assume that we can normalize the sign of $\rho(W,\gamma)$ so that $v_\rho(X) \le 0$, as will hold when $\mathrm{E}[\rho(W,\gamma_0 + a) \mid X]$ is monotonically decreasing in $a$. For example, when $\rho(W,\gamma) = Y - \gamma(X)$ as in Section 2 we have $v_\rho(X) = -1$. Also, when $\rho(W,\gamma) = p - 1(Y < \gamma(X))$ then $v_\rho(X) = -f_{Y|X}(\gamma_0(X) \mid X)$, the negative of the conditional pdf of $Y$ given $X$ evaluated at $y = \gamma_0(X)$.[2] The Neyman-orthogonality condition above includes $v_\rho(X)$, which was previously equal to $-1$. Here $v_\rho(X)$ is needed to account for the effect of $\gamma$ on the residual $\rho(W,\gamma)$.

REMARK 3.1. The orthogonal score will also be doubly robust, in the sense that $\mathrm{E}[\psi(W,\theta_0,\gamma,\alpha_0)] = 0$ for all $\gamma\in\Gamma$, if and only if $\mathrm{E}[\alpha_0(X)\rho(W,\gamma)]$ is affine in $\gamma$. This follows from $\mathrm{E}[m(W,\gamma)]$ being linear in $\gamma$ and from Chernozhukov et al. (2022b). There are many interesting cases where double robustness does not hold, such as conditional quantiles or generalized linear models. Even if the score in equation (3.3) is not doubly robust, it will still be orthogonal, enabling $\sqrt{n}$-consistent estimation and asymptotically normal inference on $\theta_0$ when $\gamma_0$ and $\alpha_0$ are estimated by machine learning.

A key innovation of our work is to note that Equation (3.4) can be viewed as the first order condition to the following optimization problem:

$$\begin{aligned}
\alpha_0 &= \arg\min_{\alpha\in\Gamma}\mathrm{E}[-2v_m(X)\alpha(X) - v_\rho(X)\alpha(X)^2] \\
&= \arg\min_{\alpha\in\Gamma}\left\{-2\,\mathrm{E}[v_m(X)\alpha(X)] - \mathrm{E}[v_\rho(X)\alpha(X)^2]\right\} \\
&= \arg\min_{\alpha\in\Gamma}\mathrm{E}[-2m(W,\alpha) - v_\rho(X)\alpha(X)^2], \qquad (3.5)
\end{aligned}$$

[2] Note that since $\gamma_0$ corresponds to the $p$ quantile of $Y \mid X$, if we denote with $\gamma_0(p,X)$ the $p$-th conditional quantile and with $\partial_p\gamma_0(p,X)$ its derivative with respect to $p$, then we have $v_\rho(X) = -(\partial_p\gamma_0(p,X))^{-1}$.

where the second equality follows by linearity of expectations, and the third equality follows by equation (2.2). Thus, $\alpha_0$ minimizes the expectation of an objective function that depends on $\alpha$ only through the functional of interest $m(W,\alpha)$ and $\alpha(X)$. As with equation (2.5), minimizing this objective function does not require any knowledge of the form of $\alpha_0$.

When $v_\rho(X) \neq -1$ the $\alpha_0$ will not be the Riesz representer $v_m(X)$. Instead, $\alpha_0$ can be interpreted as minimizing weighted least squares criterion that depends on the Riesz representer. As shown in Ichimura and Newey (2022),

$$\alpha_0 = \arg\min_{\alpha\in\Gamma} \mathrm{E}\left[-v_\rho(X)\left(-\frac{v_m(X)}{v_\rho(X)} - \alpha(X)\right)^2\right]. \tag{3.6}$$

Thus $\alpha_0(X)$ minimizes a weighted least square criterion with weight $-v_\rho(X)$ and the variable being predicted given by $-v_m(X)/v_\rho(X)$. For this reason we refer to $\alpha_0(X)$ as a weighted Riesz regression.

Though the objective functions of equations (3.5) and (3.6) differ only by a constant, only equation (3.5) possesses the desirable properties that we set out to accomplish of depending solely on known functions of $\alpha$. The objective in equation (3.5) was not given in Ichimura and Newey (2022).

In some cases there will be a function $\bar{v}_\rho(W)$ such that $\mathrm{E}[\bar{v}_\rho(W) \mid X] = v_\rho(X)$. By iterated expectations, the objective function is not affected by replacing $v_\rho(X)$ with $\bar{v}_\rho(W)$, because

$$\mathrm{E}[-2m(W,\alpha) - v_\rho(X)\alpha(X)^2] = \mathrm{E}[-2m(W,\alpha) - \bar{v}_\rho(W)\alpha(X)^2].$$

In practice, it may be easier to minimize the objective function that depends on $\bar{v}_\rho(W)$ to avoid having to estimate $v_\rho(X) = \mathrm{E}[\bar{v}_\rho(W) \mid X]$. For this reason, we focus on a sample objective function that depends on an estimator $\hat{v}_\rho(W)$ of $\bar{v}_\rho(W)$ that is allowed to take $W$ as input, instead of just $X$.

To obtain an estimate of $\alpha_0$, we replace the sample criterion in step (2b) of the algorithm in subsection 2.2 with:

$$\hat{\alpha}_\ell = \arg\min_{\alpha\in\mathcal{A}_n} \sum_{i\notin I_\ell} \left\{-2m(W_i,\alpha) - \hat{v}_\rho(W_i)\alpha(X_i)^2\right\} + \Lambda_r(\alpha), \tag{3.7}$$

for $\mathcal{A}_n \subset \Gamma$, where $\hat{v}_\rho(W)$ is an estimator of $v_\rho(X)$, in the sense that:

$$\|\hat{v}_\rho - v_\rho\|_X^2 := \mathrm{E}\left[(\mathrm{E}[\hat{v}_\rho(W) \mid X] - v_\rho(X))^2\right] = o_p(1). \tag{3.8}$$

We refer to this $\hat{\alpha}_\ell$ as a weighted Riesz regression estimator. When $v_\rho(x)$ is known, we can use $\hat{v}_\rho(W) = v_\rho(X)$, and the expectation of this objective function is (3.5), plus a penalty. Steps (3) and (4) are modified accordingly to:

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)\rho(W_i,\hat{\gamma}_\ell)\}$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left\{m(W_i,\hat{\gamma}_\ell) - \hat{\theta} + \hat{\alpha}_\ell(X_i)\rho(W_i,\hat{\gamma}_\ell)\right\}^2.$$

EXAMPLE 4 (Inverse Propensity Score Weighting). The propensity score is useful for recovering counterfacutual distributions from observational data by weighting using the inverse propensity score (Horvitz and Thompson (1952)). The superior performance of the automatic debiased machine learner in Chernozhukov et al. (2022a), which is based on estimating the inverse of the propensity score directly, suggests the potential usefulness of this approach more generally. In this example we consider estimators of counterfactual averages based on estimators of the inverse propensity score.

To describe the estimators let $D$ be a treatment indicator, $Y$ be an outcome variable, with counterfactual value $Y(1)$ satisfying $Y(1)D = YD$, $Z$ be covariates, and $\gamma_0(Z) = 1/\Pr(D = 1 \mid Z)$. be the inverse propensity score. When $D$ and $Y(1)$ are independent conditional on $Z$ and $\gamma_0(Z)$ is finite with probability one, the mean $\theta_0$ of $Y(1)$ is given by

$$\theta_0 = \mathrm{E}[m(W, \gamma_0)],\ m(W, \gamma) = DY\gamma(Z).$$

Also, the inverse propensity score satisfies

$$\mathrm{E}[\rho(W, \gamma_0) \mid Z] = 0,\ \rho(W, \gamma) = 1 - D\gamma(Z).$$

This conditional moment restriction can be interpreted as balancing for all possible functions of the covariates. This means that $\gamma_0(Z)$ is a generalized regression where $\Gamma$ is all functions of $Z$ with finite second moment and the residual is $\rho(W, \gamma) = 1 - D\gamma(Z)$. Furthermore, the conditional moment restriction corresponds to the first order condition for

$$\gamma_0 = \arg\min_{\gamma} \mathrm{E}[-2\gamma(Z) + D\gamma(Z)^2].$$

Thus $\gamma_0$ can be estimated by minimizing the sample average of $-2\gamma(Z) + D\gamma(Z)^2$ over some set $\Gamma_n$ of functions of $Z$, as in

$$\hat{\gamma}_\ell = \arg\min_{\gamma \in \Gamma_n} \left[ \sum_{i \notin I_\ell} \{-2\gamma(Z_i) + D_i\gamma(Z_i)^2\} + \Lambda_{r_\gamma}(\gamma) \right],$$

Also here $v_\rho(Z) = -\mathrm{E}[D \mid Z]$, so that we can take $v_\rho(W) = -D$, and obtain $\hat{\alpha}_\ell$ as

$$\begin{aligned} \hat{\alpha}_\ell &= \arg\min_{\alpha \in \mathcal{A}_n} [\sum_{i \notin I_\ell} \{-2D_iY_i\alpha(Z_i) + D_i\alpha(Z_i)^2\} + \Lambda_r(\alpha)] \\ &= \arg\min_{\alpha \in \mathcal{A}_n} [\sum_{i \notin I_\ell} D_i\{Y_i - \alpha(Z_i)\}^2 + \Lambda_r(\alpha)], \end{aligned}$$

where the last equality follows by adding $D_iY_i^2$ inside the brackets, which does not affect the minimizer, and completing the square. Here we see that $\hat{\alpha}$ is a least squares learner of $\mathrm{E}[Y \mid D = 1, Z]$. The resulting estimator of the parameter of interest is

$$\begin{aligned} \hat{\theta} &= \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{D_iY_i\hat{\gamma}_\ell(Z_i) + \hat{\alpha}_\ell(Z_i)[1 - D_i\hat{\gamma}_\ell(Z_i)]\} \\ &= \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{\hat{\alpha}_\ell(Z_i) + D_i\hat{\gamma}_\ell(Z_i)[Y_i - \hat{\alpha}_\ell(Z_i)]\}. \end{aligned}$$

Here $\hat{\theta}$ has the classic doubly robust form Robins and Rotnitzky (1995) of an average regression plus a bias correction term, with the key feature that the estimator $\hat{\gamma}_\ell(Z_i)$ of the inverse of the propensity score appears in place of the inverse of a propensity score estimator.

Below we give regularity conditions and a theorem to extend the results of subsection 2.3.1 to the case where $v_\rho(x)$ is unknown and needs to be estimated. For simplicity of exposition we will only consider the case where the estimator is defined over a growing sieve space $\mathcal{A}_n$ and no regularization is used, i.e. $\Lambda_r(\alpha) = 0$:

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}_n} \sum_{i=1}^{n} \left\{-2m(W_i, \alpha) - \hat{v}_\rho(W_i)\alpha(X_i)^2\right\}, \tag{3.9}$$

Our estimation rate can easily be extended to regularized estimation with appropriate regularization weight. Let $\mathcal{V}_n$ denote the function space in which the estimator $\hat{v}_\rho$ is restricted to lie in. Let $\alpha_*$ be any function in $\mathcal{A}_n$ (e.g. we will typically consider $\alpha_* = \inf_{\alpha \in \mathcal{A}_n} \|\alpha_* - \alpha_0\|$, but $\alpha_*$ can in fact be any function that is not chosen based on the samples). Define:

$$\begin{aligned}
\mathrm{star}(\sqrt{\mathcal{V}_n} \cdot (\mathcal{A}_n - \alpha_*)) &= \{w \to \xi \sqrt{|v(w)|}\,(\alpha(x) - \alpha_*(x)) : \alpha \in \mathcal{A}_n, v \in \mathcal{V}_n,\ \xi \in [0,1]\}\\
\mathrm{star}(m \circ \mathcal{A}_n - m \circ \alpha_*) &= \{w \to \xi\,(m(W,\alpha) - m(W,\alpha_*)) : \alpha \in \mathcal{A}_n,\ \xi \in [0,1]\}
\end{aligned}$$

ASSUMPTION 5. $\|f\|_\infty \leq 1$ for all $f \in \mathrm{star}(\sqrt{\mathcal{V}_n} \cdot (\mathcal{A}_n - \alpha_*))$ and $f \in \mathrm{star}(m \circ \mathcal{A}_n - m \circ \alpha_*)$.

We remark that the uniform upper bound of $1$ can be replaced by any constant upper bound $b$ and the rate that we achieve will be identical, up to an extra multiplicative factor $b$, via a standard re-scaling argument (i.e. applying our result to re-scaled version of the original problem and then scaling back the guarantee).

ASSUMPTION 6. The function $v_\rho$ and its estimate $\hat{v}_\rho$ satisfy that $|\hat{v}_\rho(W)|, |v_\rho(X)| \leq C$, almost surely, and that for any $\alpha \in \mathcal{A}_n$, the true function $v_\rho$ satisfies:

$$-\mathrm{E}[v_\rho(X)(\alpha(X) - \alpha_*(X))^2] \geq \lambda\,\mathrm{E}[(\alpha(X) - \alpha_*(X))^2], \tag{3.10}$$

for some constants $\lambda, C > 0$. For notational convenience, $\lambda \leq 1$.

THEOREM 3.2. *Let $\delta_n$ be an upper bound on the critical radius of* $\mathrm{star}(\sqrt{\mathcal{V}_n} \cdot (\mathcal{A}_n - \alpha_*))$ *and* $\mathrm{star}(m \circ \mathcal{A}_n - m \circ \alpha_*)$*. If Assumptions 1, 5 and 6 are satisfied then it follows that with probability* $1 - \zeta$*, for some universal constant C,*

$$\|\hat{\alpha} - \alpha_0\|^2 \leq C\left(\frac{M}{\lambda^2}\delta_n^2 + \frac{1}{\lambda}\|\alpha_* - \alpha_0\|^2 + \frac{1}{\lambda^2}\|\hat{v}_\rho - v_\rho\|_X^2 + \frac{M\ln(1/\zeta)}{\lambda n}\right).$$

Moreover, we note that if a separate sample was used to estimate $\hat{v}_\rho$ and not the same as the one that was used for $\hat{\alpha}$, then we can weaken the theorem to only require $\delta_n$ to upper bound the critical radius of $\mathrm{star}(\mathcal{A}_n - \alpha_*)$ and not $\mathrm{star}(\sqrt{\mathcal{V}_n} \cdot (\mathcal{A}_n - \alpha_*))$. Note that a sufficient condition for Equation (3.10) is that $|v_\rho(X)| \geq \lambda$, almost surely. However, for most function spaces $\mathcal{A}_n$, this condition can be satisfied by more benign assumptions. For this it is crucial that we only invoked the property at the difference of two functions that both lie in $\mathcal{A}_n$ and not for instance for $\alpha - \alpha_0$ (since $\alpha_0$ can potentially lie outside of the space). For instance, if the functions in $\mathcal{A}_n$ are are high-dimensional linear functions $\phi(X)'\beta$, then Equation (3.10) is satisfied, with $\lambda = \mu/C$, if:

$$\mathrm{E}[|v_\rho(X)|\phi(X)\phi(X)'] \succeq \mu I, \qquad \mathrm{E}[\phi(X)\phi(X)'] \preceq CI,$$

since then, if we let $\alpha = \phi(\cdot)'\beta$ and $\alpha_* = \phi(\cdot)'\beta_*$ and $\nu = \beta - \beta_*$, then:

$$\begin{aligned}
\|\alpha - \alpha_*\|^2 &= \nu'\,\mathrm{E}[\phi(X)\phi(X)']\nu \leq C\|\nu\|_2^2\\
&\leq \frac{C}{\mu}\nu'\,\mathrm{E}[|v_\rho(X)|\phi(X)\phi(X)']\nu = \frac{C}{\mu}\,\mathrm{E}[|v_\rho(X)|(\alpha(X) - \alpha_*(X))^2]
\end{aligned}$$

The following assumption provides regularity conditions on the residual $\rho(w,\gamma)$ and the functional of interest $m(w,\gamma)$ for the generalized regression case.

ASSUMPTION 7. (i) $\alpha_0(X)$ and $\mathrm{E}[\rho(W,\gamma_0)^2 \mid X]$ are bounded and $\mathrm{E}[m(W,\gamma_0)^2] < \infty$; (ii) $\hat{\alpha}_\ell(X)$ is bounded; (iii) $\mathrm{E}[\{\rho(W,\gamma) - \rho(W,\gamma_0)\}^2] \to 0$ if $\|\gamma - \gamma_0\| \to 0$; (iv) there is $C > 0$ such that for all $\|\gamma - \gamma_0\|$ small enough, $\mathrm{E}[\{\bar{\rho}(W,\gamma) - \bar{\rho}(W,\gamma_0)\}^2] \leq C\|\gamma - \gamma_0\|^2$, where $\bar{\rho}(X,\gamma) = \mathrm{E}[\rho(W,\gamma) \mid X]$.

The next condition allows for $\rho(W,\gamma)$ to be nonlinear in $\gamma$.

ASSUMPTION 8. Either $\rho(W,\gamma)$ is affine in $\gamma$ or $n^{1/4}\left\|\hat{\gamma}_\ell-\gamma_0\right\|\xrightarrow{p}0$ and there are $C,\varepsilon>0$ such that

$$|\mathrm{E}[m(W,\gamma)-\theta_0+\alpha_0(X)\rho(W,\gamma)]|\leq C\left\|\gamma-\gamma_0\right\|^2.$$

whenever $\left\|\gamma-\gamma_0\right\|^2\leq\varepsilon$.

This assumption imposes the usual faster than $n^{-1/4}$ convergence rate for $\hat{\gamma}_\ell$ when $\rho(w,\gamma)$ is nonlinear in $\gamma$ but does not require that rate when $\rho(W,\gamma)$ is linear in $\gamma$.

We have the following large sample inference result under these conditions.

THEOREM 3.3. *If Assumptions 1, 4, 7 and 8 are satisfied, then*

$$\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{p}N(0,V)\quad\textit{and}\quad\hat{V}\xrightarrow{p}V.$$

*where* $V=\mathrm{E}[\{m(W,\gamma_0)-\theta_0+\alpha_0(X)\rho(W,\gamma_0)\}^2]$.

3.2. *Nonlinear Effects of Multiple Regressions.* Some important objects of interest are expectations of nonlinear functionals of multiple regressions. In this Section we give Auto-DML for such effects. Such effects have the form $\theta_0=\mathrm{E}[m(W,\gamma_0)]$ where $m(w,\gamma)$ is nonlinear in a possible value $\gamma$ of multiple generalized regressions $(\gamma_1(X_1),...,\gamma_J(X_J))'$ with regressors $X_j$, residual $\rho_j(W,\gamma_j)$, and $\Gamma_j$ specific to each regression $\gamma_j(X_j)$. The corresponding orthogonal score like is like that of subsection 3.1 except that the bias correction is a sum of $J$ terms with the $j^{th}$ term being the bias correction for the learner of $\gamma_j$. Similarly to Newey (1994), pg. 1357, the orthogonal score is

$$\psi(w,\gamma,\alpha,\theta)=m(W,\gamma)-\theta+\sum_{j=1}^{J}\alpha_j(X_j)\rho_j(W,\gamma_j),\ \gamma_j,\alpha_j\in\Gamma_j. \tag{3.11}$$

Each of the terms in the bias correction can be estimated by the product of a learner $\hat{\alpha}_{j\ell}(X_j)$ and the residual $\rho_j(W,\hat{\gamma}_{j\ell})$, but now the learner $\hat{\alpha}_{j\ell}(X_j)$ differs from the one given in section 3.1 in the way needed to account correctly for nonlinearity of $m(W,\gamma)$ in $\gamma$. The difference is that in the objective function for $\hat{\alpha}_{j\ell}(X_j)$ the functional of interest $m(w,\alpha)$ is replaced by an estimated Gateaux derivative with respect to the $j^{th}$ component of $\gamma$. Let

$$\hat{D}_j(W,\alpha_j)=\left.\frac{d}{d\tau}m(W,\hat{\gamma}_\ell+\tau e_j\alpha_j)\right|_{\tau=0}$$

be such a Gateaux derivative estimator, where $e_j$ denotes the $j$-th column of the identity matrix. This derivative will often be straightforward to calculate as an analytic derivative with respect to the scalar $\tau$. When $m(w,\gamma)$ is linear in a single $\gamma$ this derivative just evaluates $m(W_i,\gamma)$ at $\gamma=\alpha$ giving the $m(W,\alpha)$ of subsection 2.2.

To obtain $\hat{\alpha}_{j\ell}(X_j)$ we also make use of an estimated derivative $\hat{v}_{\rho j}(W_i)$ of $\rho_j(W,\gamma_j)$ with respect to $\gamma_j$ at $\hat{\gamma}_{j\ell}$. Then $\hat{\alpha}_{j\ell}$ is given by

$$\hat{\alpha}_{j\ell}=\arg\min_{\alpha_j\in\mathcal{A}_n^j}\left\{\sum_{i\notin I_\ell}[-2\hat{D}_j(W_i,\alpha_j)-\hat{v}_{\rho j}(W_i)\alpha_j(X_{ji})^2]\right\}, \tag{3.12}$$

where $\mathcal{A}_n^j$ is the set of approximating functions for $\alpha_j$. As with linear $m(w,\gamma)$ this $\hat{\alpha}_{j\ell}$ depends just on $m(w,\gamma)$ and the first step. Thus $\hat{\alpha}_{j\ell}$ is automatic, in the same way as in section 2, in only requiring $m(w,\gamma)$ and the regression residual $\rho_j(W_i,\gamma_j)$ for its construction.

Below we give two examples of this setting:

EXAMPLE 5 (Marginal Effect in a Generalized Regression Model). Suppose that $X = (D, Z)$, where $D$ is a continuous treatment or policy variable and $Z$ are covariates. We are interested in $\theta_0 = \mathrm{E}[m(W, \gamma_0)]$ with

$$m(W, \gamma) = \partial_a \mu(\gamma(X)) \partial_d \gamma(X).$$

The function $\gamma_0 \in \Gamma$ is assumed to satisfy the orthogonality condition (3.2) for $\rho(W, \gamma) = \lambda(\gamma(X))[Y - \mu(\gamma(X))]$. This is the first order condition of a Generalized Regression Model with link function $\mu(a)$ (Nelder and Wedderburn (1972)). For example, when $Y$ is binary, $\mu(a)$ is a CDF and $\lambda(a) = \partial_a \mu(a)/[\mu(a)(1 - \mu(a))]$, and $\Gamma$ is the mean square closure of finite linear combinations of $X$, this corresponds to a high dimensional, approximately sparse binary response model.

In this example, $J = 1$, $\hat{D}(W, \alpha) = \partial_a^2(\hat{\gamma}(X))\partial_d \hat{\gamma}(X)\alpha(X) + \partial_a \mu(\hat{\gamma}(X))\partial_d \alpha(X)$ and $\hat{v}_\rho(X) = \partial_a \lambda(\hat{\gamma}(X))[Y - \mu(\hat{\gamma}(X))] - \lambda(\hat{\gamma}(X))\partial_a \mu(\hat{\gamma}(X))$. In the Logit case, $\lambda(a) \equiv 1$, and so this simplifies to $\hat{v}_\rho(X) = -\partial_a \mu(\hat{\gamma}(X))$. The weighted Riesz regression estimator $\hat{\alpha}_\ell$ can be found by evaluating (3.12) at these quantities, and then used to build the Neyman orthogonal score (3.11).

EXAMPLE 6 (Inverse Logit Propensity Score Weighting). Suppose now that $Y$ is a continuous or discrete outcome, $X = (D, Z)$ where $D$ is a binary treatment and $Z$ are covariates, and the parameter of interest is $\theta_0 = \mathrm{E}[m(W, \gamma_0)]$ with

$$m(W, \gamma) = \frac{DY}{\Lambda(\gamma(Z))},$$

where $\Lambda(a)$ is the standard logistic CDF. The parameter $\gamma_0 \in \Gamma$ satisfies the orthogonality condition (3.2) for $\rho(W, \gamma) = D - \Lambda(\gamma(Z))$. This corresponds to inverse propensity score weighting of the outcome, where the propensity score is modelled by a flexible Logit specification. For example, if $\Gamma$ is the mean square closure of finite linear combinations of $X$, this corresponds to a high dimensional, approximately sparse logit model; if $\Gamma$ is the space of all square-integrable functions, the model is essentially unrestricted.

In this example, $J = 1$, $\hat{D}(W, \alpha) = -DY\partial_a \Lambda(\hat{\gamma}(X))\alpha(X)/\Lambda(\hat{\gamma}(X))^2$ and $\hat{v}_\rho(X) = -\partial_a \Lambda(\hat{\gamma}(X))$. The debiasing function $\hat{\alpha}_\ell$ can be found by evaluating (3.12) at these quantities, and then used to build the Neyman orthogonal score (3.11).

It is straightforward to obtain a convergence rate for $\hat{\alpha}_{j\ell}$ analogous to Theorem 3.2. The following result does so while accounting for the presence of $\hat{\gamma}$ in $\hat{D}_j(W_i, \alpha_j)$. For notational convenience we suppress the $j$ subscripts.

ASSUMPTION 9. The estimate $\hat{D}$ satisfies that:

$$|\,\mathrm{E}[\hat{D}(W, \alpha) - D(W, \alpha)]| \leq \epsilon_{mn} \|\alpha\| \tag{3.13}$$

THEOREM 3.4. *If the conditions of Theorem 3.2 and Assumption 9 is satisfied then it follows that with probability* $1 - \zeta$, *for some universal constant* $C$,

$$\|\hat{\alpha} - \alpha_0\|^2 \leq C\left(\frac{M}{\lambda^2}\delta_n^2 + \frac{1}{\lambda}\|\alpha_* - \alpha_0\|^2 + \frac{1}{\lambda^2}\left(\|\hat{v}_\rho - v_\rho\|_X^2 + \epsilon_{mn}^2\right) + \frac{M\ln(1/\zeta)}{\lambda n}\right).$$

The construction of $\hat{\theta}$ is analogous to that in subsection 2.2 with the bias correction term being the sum of terms for each $\gamma_j$ in $\gamma$. That is,

$$\hat{\theta} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i \in I_\ell}\{m(W_i, \hat{\gamma}_\ell) + \sum_{j=1}^{J}\hat{\alpha}_{j\ell}(X_{ji})\rho_j(W_i, \hat{\gamma}_{j\ell})\}, \tag{3.14}$$

$$\hat{V} = \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\left\{m(W_i,\hat{\gamma}_\ell) - \hat{\theta} + \sum_{j=1}^{J}\hat{\alpha}_{j\ell}(X_{ji})\rho_j(W_i,\hat{\gamma}_{j\ell})\right\}^2.$$

It is straightforward to specify conditions for asymptotic normality of $\hat{\theta}$ and consistency of $\hat{V}$ by combining the conditions of section 3.1 with the convergence rate result of Corollary 2.2. For relative simplicity we give a result only for neural net learners. We also assume for simplicity that each $X_j$ has the same dimension $d$.

ASSUMPTION 10. $\mathrm{E}[m(W,\gamma_0)^2] < \infty$ and for each $j$, (i) $\mathrm{E}[\rho_j(W,\gamma_{j0})^2 \mid X]$ is bounded (ii), $\mathrm{E}[\{\rho_j(W,\gamma_j) - \rho_j(W,\gamma_0)\}^2] \to 0$ if $\|\gamma_j - \gamma_{j0}\| \to 0$; (iii) there is $C > 0$ such that for all $\|\gamma_j - \gamma_{j0}\|$ small enough $\mathrm{E}[\{\bar{\rho}_j(X_j,\gamma_j) - \bar{\rho}_j(X,\gamma_{j0})\}^2] \le C\|\gamma_j - \gamma_{j0}\|^2$, where $\bar{\rho}_j(X_j,\gamma_j) = \mathrm{E}[\rho_j(W,\gamma_j) \mid X_j]$.

This condition is analogous to Assumption 7.

ASSUMPTION 11. $n^{1/4}\|\hat{\gamma}_{j\ell} - \gamma_{j0}\| \xrightarrow{p} 0$ for each $j$ and there are $C, \varepsilon > 0$ such that for

$$\left|\mathrm{E}[m(W,\gamma) - \theta_0 + \sum_{j=1}^{J}\alpha_{j0}(X_j)\rho(W,\gamma_j)]\right| \le C\sum_{j=1}^{J}\|\gamma_j - \gamma_{j0}\|^2.$$

whenever $\|\gamma_j - \gamma_{j0}\|^2 \le \varepsilon$ for all $j = 1,\ldots,J$.

This condition is analogous to Assumption 8.

THEOREM 3.5. *If Assumptions 1, 4, 10 and 11 are satisfied for each $j = 1,\ldots,J$, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{p} N(0,V),\ \hat{V} \xrightarrow{p} V.$$

*where* $V = \mathrm{E}[\{m(W,\gamma_0) - \theta_0 + \sum_{j=1}^{J}\alpha_{j0}(X_j)\rho_j(W,\gamma_{j0})\}^2]$.

## 4. Empirical Application.

To illustrate our methods, we study whether applicant race is a significant predictor of banks' mortgage denial decisions. Following Munnell et al. (1996), we use the publicly available Boston Home Mortgage Disclosure Act (HDMA) dataset. The dataset contains information on 2,925 mortgage applications made in 1990 in the Greater Boston metropolitan area. We restrict attention to black and white applicants in single-family households (excluding other racial minorities and multi-family residences), which reduces our sample size to 2,380 observations.

Our outcome of interest is an indicator $Y = 1$ if the mortgage application was denied. Our regressor of interest is an indicator $D = 1$ if the applicant is black. We also have access to a vector of covariates, which we denote by $Z$, containing financial and other characteristics of the applicant that banks may factor into their mortgage denial decisions. These include monthly debt to income (DTI) ratio; monthly housing expenses to income (HTI) ratio; loan to assessed property value (LTV) ratio; a categorical variable for "bad" consumer credit score with 6 categories (1 if no slow payments or delinquencies, 2 if one or two slow payments or delinquencies, 3 if more than two slow payments or delinquencies, 4 if insufficient credit history for determination, 5 if delinquent credit history with payments 60 days overdue, and 6 if delinquent credit history with payments 90 days overdue); a categorical variable for "bad" mortgage credit score with 4 categories (1 if no late mortgage payments, 2 if no mortgage payment history, 3 if one or two late mortgage payments, and 4 if more than two late mortgage

TABLE 1
*Summary Statistics*

| | Full Sample | | Black | | White | |
|---|---|---|---|---|---|---|
| | mean | sd | mean | sd | mean | sd |
| Deny | 0.12 | 0.32 | 0.28 | 0.45 | 0.09 | 0.29 |
| Monthly DTI Ratio | 0.19 | 0.01 | 0.19 | 0.01 | 0.19 | 0.01 |
| Monthly HTI Ratio | 0.12 | 0.01 | 0.12 | 0.01 | 0.12 | 0.01 |
| LTV Ratio | 0.37 | 0.03 | 0.38 | 0.01 | 0.37 | 0.03 |
| Consumer Credit Ind. | 2.12 | 1.67 | 3.02 | 2.01 | 1.97 | 1.55 |
| Mortgage Credit Ind. | 1.72 | 0.54 | 1.88 | 0.42 | 1.69 | 0.55 |
| Public Record | 0.07 | 0.26 | 0.18 | 0.38 | 0.06 | 0.23 |
| Denied Insurance | 0.02 | 0.14 | 0.05 | 0.22 | 0.02 | 0.12 |
| Self-Employed | 0.12 | 0.32 | 0.07 | 0.26 | 0.12 | 0.33 |
| Single | 0.39 | 0.49 | 0.52 | 0.50 | 0.37 | 0.48 |
| High School | 0.98 | 0.13 | 0.97 | 0.18 | 0.99 | 0.12 |
| Industry Unemp. | 3.77 | 2.03 | 3.45 | 1.50 | 3.83 | 2.10 |
| Condominium | 0.29 | 0.45 | 0.49 | 0.50 | 0.25 | 0.44 |
| *N* | 2,380 | | 339 | | 2,041 | |

payments); an indicator for public record of credit problems including bankruptcy, charge-offs, and collective actions; an indicator for denial of application for mortgage insurance; three indicators for self-employed, single, and high school graduate, the 1989 Massachusetts unemployment rate in the applicant's industry, and an indicator for whether the unit is a condominium.

Table 1 reports the sample means and standard deviations of the variables used in the analysis. The probability of being denied a mortgage is 19 percentage points higher for black applicants than for white applicants. However, black applicants are also more likely to have financial and socio-economic characteristics linked to mortgage denial, as Table 1 shows. For example, black applicants have higher (worse) consumer and mortgage credit indices on average, and are more likely to have a public record of credit problems and to be single. We would like to test whether the racial differences in probability of mortgage denial persist once we control for these covariates.

To showcase the versatility of our method, we present results for three estimands:

1. *Difference in Probability of Mortgage Denial:*

$$\theta_0 = \mathrm{E}[\gamma_0(1,Z) - \gamma_0(0,Z)], \quad \text{where} \quad \gamma_0(D,Z) = \Pr(Y=1 \mid D,Z) = \mathrm{E}[Y \mid D,Z].$$

   This is an average linear effect for a conditional mean (Section 2). This parameter can be interpreted as an average difference in probability of mortgage denial between a black and a white applicant with the same value of covariates $Z$.

2. *Average Difference in Log-Odds of Mortgage Denial:*

$$\theta_0 = \mathrm{E}[\gamma_0(1,Z) - \gamma_0(0,Z)], \quad \text{where} \quad \gamma_0(D,Z) = \ln \frac{\Pr(Y=1 \mid D,Z)}{\Pr(Y=0 \mid D,Z)}.$$

   This is an average non-linear effect for a generalized regression (Section 3.2). Because $\ln(a) - \ln(b) \approx (a-b)/b$ when $(a-b)/b$ is small, this parameter can be interpreted as an approximate average percentage difference in odds of mortgage denial between a black and a white applicant with the same value of covariates $Z$. As discussed in Section 3.2, this $\gamma_0$ minimizes the logistic regression loss function,

$$\ell(W,\gamma) = Y \ln \Lambda(\gamma(X)) + (1-Y) \ln[1 - \Lambda(\gamma(X))],$$

with corresponding generalized residual

$$\rho(W,\gamma) = Y - \Lambda(\gamma(X)),$$

for $\Lambda(t) := 1/(1+e^{-t})$, the standard logistic CDF.

3. *Average Difference in Odds of Mortgage Denial:*

$$\theta_0 = \mathrm{E}[e^{\gamma_0(1,Z)} - e^{\gamma_0(0,Z)}], \quad \text{where} \quad \gamma_0(D,Z) = \ln\frac{\Pr(Y=1 \mid D,Z)}{\Pr(Y=0 \mid D,Z)}.$$

This is an average non-linear effect for a generalized regression (Section 3.2). It can be interpreted as an average difference in odds of mortgage denial between a black and a white applicant with the same value of covariates $Z$.

We estimate these parameters using AutoDML, where both $\hat{\gamma}$ and $\hat{\alpha}$ are neural net learners. For the difference in probability, we have $v_\rho(X) = -1$, since $\gamma_0$ is a conditional mean. For the average difference in log-odds and the average difference in odds, $v_\rho(X) = -\lambda(\gamma_0(X))$, where $\lambda(t) := e^{-t}/(1-e^{-t})^2$ is the standard logistic PDF, which we estimate by replacing $\gamma_0$ with a preliminary estimate $\hat{\gamma}_{\text{prel}}$, also based on a neural net learner. We describe the architecture and training hyperparameter choice in detail in Appendix B.

TABLE 2
*Empirical Application Results: Racial Differences in Probability, Average Log-Odds and Average Odds of Mortgage Denial*

| | Probability | | Log-Odds | | Odds | |
|---|---|---|---|---|---|---|
| | est | se | est | se | est | se |
| Main Spec. | 0.080 | (0.021) | 0.829 | (0.152) | 0.157 | (0.044) |

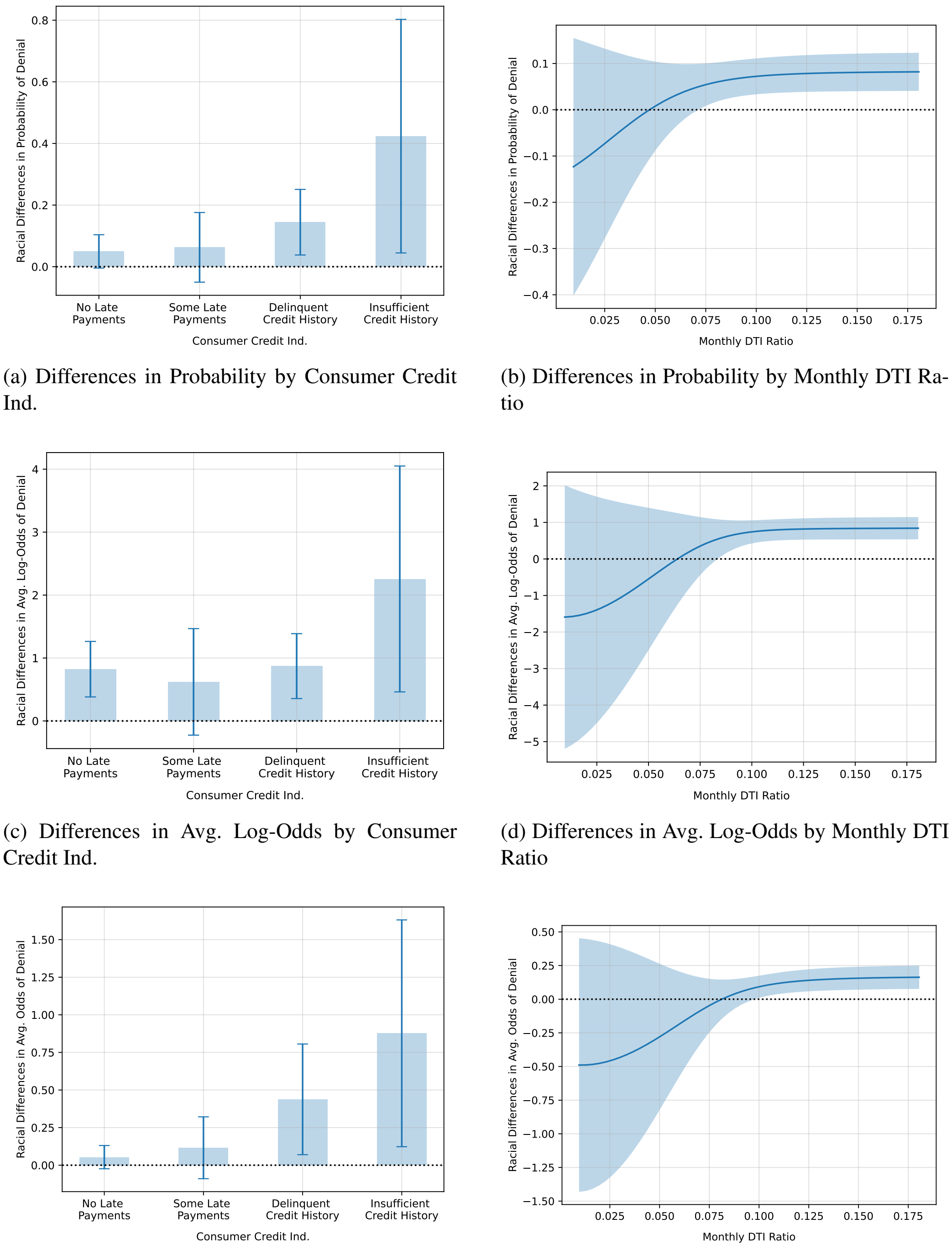
Table 2 presents the results of our main analysis. Once we control for covariates, the difference in probability of mortgage denial decreases from 19 to 8 percentage points. If we look at the average log-odds or odds instead, we observe differences of 0.829 or 0.157, respectively. These differences are all estimated to be statistically different from 0 at the 1% significance level.

A slight modification of our method allows us to estimate average differences for subgroups of applicants with certain characteristics (analogous to conditional average treatment effects or CATEs). Suppose we want to estimate an average effect for applicants with $Z_j = z$ for a particular covariate $Z_j$. To obtain these, we weight the Neyman orthogonal estimating equation (3.14) as follows:

$$\hat{\theta}(z) = \frac{1}{\sum_i^n \omega_i(z)} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \omega_i(z)\{m(W_i,\hat{\gamma}_\ell) + \hat{\alpha}_\ell(X_i)\rho(W_i,\hat{\gamma}_\ell)\}.$$

When $Z_j$ is a categorical variable, we take $\omega_i(z) = 1\{Z_j = z\}$. When $Z_j$ is continuous, we take $\omega_i(z) = K((Z_i - z)/h)$ for a kernel function $K$ and a small but fixed bandwith $h$.[3] Figure 1 presents the racial differences in probability, average log-odds and average odds by values of the consumer credit index and the monthly DTI ratio. Remarkably, we estimate the racial differences in all three estimands to be higher for applicants with a delinquent credit history or with insufficient credit history (although the latter is quite imprecisely estimated). The racial differences appear to be constant for most of the range of the monthly DTI ratio variable, except values below 0.075 for which it is also imprecisely estimated.

[3] Chernozhukov, Newey and Singh (2019) analyze a localized version of this parameter, that is, the limit as $h \to 0$, which is beyond the scope of this paper.



(a) Differences in Probability by Consumer Credit Ind.

(b) Differences in Probability by Monthly DTI Ratio

(c) Differences in Avg. Log-Odds by Consumer Credit Ind.

(d) Differences in Avg. Log-Odds by Monthly DTI Ratio

(e) Differences in Avg. Odds by Consumer Credit Ind.

(f) Differences in Avg. Odds by Monthly DTI Ratio

Fig 1: Heterogeneous Effects

## 5. Monte Carlo Simulations.

5.1. *Based on the Empirical Application.* First, we analyze the performance of our method in the setting of our empirical application. We redraw the covariates $Z$ based on a generative adversarial network (GAN) trained on the real mortgage data. We use an elastic-net Logit, with penalties chosen by cross-validation, to estimate the outcome regression $\Pr(Y=1 \mid D, Z)$ and the propensity score $\Pr(D=1 \mid Z)$ in the real mortgage data, which we take as ground truth in our simulations. We present results for the three estimands on interest in Section 4: the difference in probability, the average difference in log-odds and the average difference in odds.

Table 3 presents simulation results over 1,000 draws for $n = 2,000$ and $n = 10,000$. The first column shows the non-parametric $R^2$ for $\gamma$, defined as $R^2(\gamma) = 1 - \mathrm{E}[(\hat{\gamma}(X) - \gamma_0(X))^2]/\mathrm{Var}(\gamma_0(X))$, where the expectation is evaluated over a test set not used to estimate $\hat{\gamma}$. The second column shows the same non-parametric $R^2$ metric for $\alpha$. We also give the mean absolute error (MAE), bias, standard deviation (sd), the average standard error to standard deviation ratio (se/sd) and the coverage of a 95$\%$ confidence interval (covg.).

For the estimands we study, the Riesz regression $\alpha_0$ can be characterized explicitly based on the Riesz representer for the Average Treatment Effect (ATE),

$$v_m(X) = \frac{D}{\Pr(D=1 \mid Z)} - \frac{1-D}{1-\Pr(D=1 \mid Z)}.$$

For the difference in probability estimand, which is an average linear effect for a conditional mean (Section 2), $\alpha_0(X) = v_m(X)$. For the average difference in log-odds, which is an average linear effect for a generalized regression (Section 3.1), the weighted Riesz regression is $\alpha_0(X) = v_m(X)/(-v_\rho(X))$, where $v_\rho(X) = -\lambda(\gamma_0(X))$. Finally, for the average difference in odds, which is an average non-linear effect for a generalized regression (Section 3.2), we have $\alpha_0(X) = e^{\gamma_0(X)} v_m(X)/(-v_\rho(X))$.

Our automatic debasing method does not make use of this explicit characterization of $\alpha_0$. To benchmark our results, we compare its performance to an estimator that uses the explicit characterization of $\alpha_0$. In the ATE setting, this is known as an Augmented Inverse Propensity Weighting (AIPW) estimator, so we will refer to these as AIPW-like. To build the AIPW-like estimator of $\theta_0$ we plug learners of the outcome and treatment propensities $\Pr(Y=1 \mid D, Z)$ and $\Pr(D=1 \mid Z)$ into the formula for $\alpha_0$; we try both a non-parametric version based on neural nets (the same architecture and hyperparameters as our main specification) and a "well-specified" version, where we use the same elastic-net Logit that we used to build the ground truth.

A comparison between auto-DML and the AIPW-like benchmark sheds light on the advantages of our automatic approach. For large sample sizes, $n = 10,000$, both methods perform comparably well. There is no loss in efficiency between our main specification (which uses a non-parametric, NN-based method) and the correctly-specified AIPW estimator (which uses correctly specified parametric learners for $\gamma$ and $\alpha$). Our automatic debiasing method achieves close to nominal coverage, whereas the AIPW-like method gets worse coverage when we use the non-parametric, neural net specification. A reason for that could be that, in the explicit characterization of $\alpha_0$, we are plugging numbers that are close to zero into a denominator (such as the propensity score $\Pr(D=1 \mid Z)$ or the logit pdf $v_\rho(X)$), so that estimation error amplifies. This is reflected into the large negative non-parametric $R^2$ for $\alpha$ when $n = 2,000$.

Consistent with these results, in work that followed up on the first version of this paper (Chernozhukov et al. (2022a)), we found that our automatic debiasing method using neural net and random forest Riesz regressions performed much better than state of the art methods based on inverse propensity score weighting in Monte Carlo experiments. Singh and Sun (2023) also found that automatic debiased estimators performed better than plugin-based estimators in the setting of local average treatment effects.

TABLE 3
*Simulation Results: Based on the Empirical Application*

| | Probability | | | | | | |
|---|---|---|---|---|---|---|---|
| | $R^2(\gamma)$ | $R^2(\alpha)$ | MAE | bias | sd | se/sd | covg. |
| $n = 2,000$ | | | | | | | |
| Main Spec. | 0.639 | 0.820 | 0.021 | 0.005 | 0.027 | 0.938 | 0.931 |
| AIPW, NN | 0.622 | -1e7 | 0.022 | 0.005 | 0.028 | 0.890 | 0.923 |
| AIPW, well spec. | 0.807 | 0.912 | 0.021 | 0.003 | 0.026 | 0.956 | 0.932 |
| $n = 10,000$ | | | | | | | |
| Main Spec. | 0.898 | 0.969 | 0.009 | 0.002 | 0.012 | 0.973 | 0.946 |
| AIPW, NN | 0.892 | 0.969 | 0.009 | 0.002 | 0.012 | 0.970 | 0.938 |
| AIPW, well spec. | 0.963 | 0.988 | 0.009 | 0.001 | 0.012 | 0.984 | 0.949 |

| | Log-Odds | | | | | | |
|---|---|---|---|---|---|---|---|
| | $R^2(\gamma)$ | $R^2(\alpha)$ | MAE | bias | sd | se/sd | covg. |
| $n = 2,000$ | | | | | | | |
| Main Spec. | 0.583 | 0.368 | 0.203 | -0.010 | 0.265 | 0.871 | 0.899 |
| AIPW, NN | 0.585 | -2e7 | 0.197 | 0.019 | 0.260 | 0.840 | 0.893 |
| AIPW, well spec. | 0.767 | -3e3 | 0.191 | -0.019 | 0.243 | 0.909 | 0.915 |
| $n = 10,000$ | | | | | | | |
| Main Spec. | 0.889 | 0.914 | 0.086 | 0.000 | 0.107 | 0.910 | 0.925 |
| AIPW, NN | 0.888 | 0.925 | 0.084 | 0.007 | 0.104 | 0.927 | 0.928 |
| AIPW, well spec. | 0.959 | 0.959 | 0.083 | -0.007 | 0.104 | 0.977 | 0.938 |

| | Odds | | | | | | |
|---|---|---|---|---|---|---|---|
| | $R^2(\gamma)$ | $R^2(\alpha)$ | MAE | bias | sd | se/sd | covg. |
| $n = 2,000$ | | | | | | | |
| Main Spec. | 0.589 | 0.653 | 0.039 | 0.015 | 0.055 | 0.786 | 0.932 |
| AIPW, NN | 0.585 | -9e6 | 0.044 | 0.018 | 0.061 | 0.820 | 0.931 |
| AIPW, well spec. | 0.767 | -2e10 | 0.199 | 0.174 | 4.567 | 0.022 | 0.899 |
| $n = 10,000$ | | | | | | | |
| Main Spec. | 0.889 | 0.830 | 0.015 | 0.003 | 0.019 | 0.948 | 0.947 |
| AIPW, NN | 0.888 | 0.867 | 0.016 | 0.006 | 0.020 | 0.960 | 0.936 |
| AIPW, well spec. | 0.959 | 0.866 | 0.017 | 0.005 | 0.021 | 0.994 | 0.95 |

5.2. *Additional Simulations.* We present an additional set of simulations when $\gamma_0$ is a quantile of the conditional distribution of $Y \mid X$ for a continuous outcome $Y$. For simplicity, we will focus on the conditional median. As discussed in Section 3.1, this corresponds to the generalized residual $\rho(W, \gamma) = 0.5 - 1(Y < \gamma(X))$, which is a sub-derivative of the "check" loss function $\ell(W, \gamma) = 0.5|Y - \gamma(X)|$.

We will consider four simulation settings. In the first setting, the object of interest will be the average difference in conditional median for a binary treatment $D$, that is, $\theta_0 = \mathrm{E}[\gamma_0(1, Z) - \gamma_0(0, Z)]$. We assume that $\Pr(D = 1 \mid Z) = \text{logit}(-0.1 + 0.5Z_1 - 0.2Z_2)$, and that $Y = \mu_Y(X) + \epsilon$ with $\mu_Y(X) = 0.5D - 0.2D \times Z_3 + 0.3Z_2$ and $\epsilon \sim N(0, 1)$. The second setting considers the same data generating process, but we use a twice-differentiable quantile loss based on Epanechnikov kernel smoothing, due to He et al. (2023), rather than the non-differentiable check loss function. In the third setting, our object of interest is the average derivative of the conditional median with respect to a continuous treatment $D$, that is, $\theta_0 = \mathrm{E}[\partial_d \gamma_0(D, Z)]$, where we draw $D = \mu_D(Z) + \eta$ for $\mu_D(Z) = -0.1 + 0.5Z_1 - 0.2Z_2$ and $\eta \sim N(0, 1)$; the distribution of $Y \mid X$ is as before. The last setting uses the same

TABLE 4
*Simulation Results: Additional Designs*

| | Quantiles | | | | | | |
|---|---|---|---|---|---|---|---|
| | $R^2(\gamma)$ | $R^2(\alpha)$ | MAE | bias | sd | se/sd | covg. |
| $n = 2,000$ | | | | | | | |
| Main Spec. | 0.737 | 0.981 | 0.045 | -0.003 | 0.055 | 1.035 | 0.96 |
| $n = 10,000$ | | | | | | | |
| Main Spec. | 0.890 | 0.993 | 0.020 | 0.001 | 0.025 | 1.022 | 0.957 |

| | Quantiles, Smooth Loss | | | | | | |
|---|---|---|---|---|---|---|---|
| | $R^2(\gamma)$ | $R^2(\alpha)$ | MAE | bias | sd | se/sd | covg. |
| $n = 2,000$ | | | | | | | |
| Main Spec. | 0.757 | -8.562 | 0.044 | -0.006 | 0.116 | 0.469 | 0.957 |
| $n = 10,000$ | | | | | | | |
| Main Spec. | 0.898 | 0.993 | 0.019 | 0.000 | 0.024 | 0.984 | 0.948 |

| | Quantiles, Smooth Loss, Continuous Treatment | | | | | | |
|---|---|---|---|---|---|---|---|
| | $R^2(\gamma)$ | $R^2(\alpha)$ | MAE | bias | sd | se/sd | covg. |
| $n = 2,000$ | | | | | | | |
| Main Spec. | 0.869 | 0.944 | 0.021 | -0.002 | 0.026 | 0.968 | 0.928 |
| $n = 10,000$ | | | | | | | |
| Main Spec. | 0.952 | 0.979 | 0.010 | -0.002 | 0.012 | 0.941 | 0.922 |

| | Quantiles, Smooth Loss, Continuous Treatment, Non-Linear Effect | | | | | | |
|---|---|---|---|---|---|---|---|
| | $R^2(\gamma)$ | $R^2(\alpha)$ | MAE | bias | sd | se/sd | covg. |
| $n = 2,000$ | | | | | | | |
| Main Spec. | 0.976 | 0.926 | 0.036 | 0.013 | 0.043 | 0.933 | 0.926 |
| $n = 10,000$ | | | | | | | |
| Main Spec. | 0.990 | 0.964 | 0.016 | 0.002 | 0.019 | 0.947 | 0.938 |

data generating process, but it focuses on a non-linear parameter, the average derivative squared $\theta_0 = \mathrm{E}[(\partial_d \gamma_0(D,Z))^2]$. This parameter is useful, for example, in testing whether $\partial_d \gamma_0(D,Z) = 0$ with probability 1. In all settings, to work with a realistic data generating process for the covariates, we will draw $Z$ from the same GAN trained on the real mortgage data that we used in the previous subsection.

The results over 1,000 simulation draws are presented in Table 4. In all the settings we consider, our estimator performs well, with low bias and coverage confidence intervals close to the nominal 95% level.

# APPENDIX A: PROOFS

PROOF OF THEOREM 2.1. This is a special case of Theorem 3.2 with $v_\rho(X) = \hat{v}_\rho(X) = -1$ and $\epsilon_{\rho n} = 0$. We prove the more general version below. □

PROOF OF COROLLARY 2.2. An upper bound for the critical radius of a MLP neural net is given in equation (A.10) of Farrell, Liang and Misra (2021a). Using the fact that the number of parameters given there is bounded by $CK^2m$ it follows that

$$\delta_n \leq C\sqrt{\frac{K^2m^2\ln(K^2m)\ln(n)}{n}}, \tag{A.1}$$

where $C$ denotes a generic positive constant. Let $\epsilon_n = \inf_{\alpha\in\mathcal{A},x\in\mathcal{X}} |\alpha(x) - \alpha_0(x)|$ . It follows by the uniform approximating bounds given in FLM, in particular in the first inequality on the top of p. 206, that

$$K^2m^2\ln(K^2m) \leq C\epsilon_n^{-2d/\beta}(\ln(1/\epsilon_n)+1)^7.$$

It follows that for any $\varepsilon > 0$ and $n$ large enough,

$$\epsilon_n \leq C\{Km\sqrt{\ln(K^2m)}\}^{-\beta/d+\varepsilon},$$

where the presence of $\varepsilon$ allows us to ignore the $(\ln(1/\epsilon_n)+1)^7$ term. It follows that

$$\|\alpha^* - \alpha_0\| \leq C\epsilon_n \leq C\{Km\sqrt{\ln(K^2m)}\}^{-\beta/d+\varepsilon}. \tag{A.2}$$

The conclusion then follows from Theorem 1 and squaring and plugging in the inequalities from equations (A.1) and (A.2). □

PROOF OF THEOREM 2.3. This is a special case of Theorem 3.3 with $\rho(W,\gamma) = Y - \gamma(X)$. Note that Assumption 3 implies Assumption 7 (i). Assumptions 7 (ii), (iii) and 8 are obviously satisfied for this choice of $\rho(W,\gamma) = Y - \gamma(X)$. □

PROOF OF COROLLARY 2.4. By Corollary 2.2, $\|\hat{\alpha}_\ell - \alpha_0\| = O_p(\epsilon_{\alpha n})$, which satisfies the rate conditions of Assumption 4. The conclusion follows by Theorem 2.3. □

PROOF OF THEOREM 3.2. Throughout this proof, let $C > 0$ denote a generic constant (possibly different each time it appears), and let $\mathbb{E}_n[\cdot]$ denote the empirical expectation over a sample of size $n$, i.e. $\mathbb{E}_n[Z] = \frac{1}{n}\sum_{i=1}^n Z_i$,

$$\begin{aligned} L_n(\alpha, v) =&\ \mathbb{E}_n[-2m(W,\alpha) - v(W)\alpha(X)^2], \\ L(\alpha, v) =&\ \mathrm{E}[-2m(W,\alpha) - v(W)\alpha(X)^2] \end{aligned}$$

Note that

$$\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}_n} L_n(\alpha, \hat{v}_\rho).$$

Since $\Gamma$ is a closed linear space and $\alpha_0$ is defined as the minimizer of $L(a)$ over $\Gamma$, then we have the first-order condition that for all $\nu \in \Gamma$:

$$\left.\frac{\partial}{\partial\tau}L(\alpha_0 + \tau\nu, v_\rho)\right|_{\tau=0} = 0 \tag{A.3}$$

Moreover, note that by the linearity of the moment $m$ and linearity of expectation:

$$\begin{aligned} \left.\frac{\partial}{\partial\tau}L(\alpha_0 + \tau\nu, v)\right|_{\tau=0} =&\ \mathrm{E}\left[-2m(W;\nu) - 2v(W)\alpha_0(X)\nu(X)\right] \\ =&\ \mathrm{E}\left[-2m(W;\nu) - 2\,\mathrm{E}[v(W) \mid X]\alpha_0(X)\nu(X)\right] \qquad \text{(A.4)} \end{aligned}$$

Thus we have:

$$\begin{aligned}\frac{\partial}{\partial\tau}L(\alpha_0+\tau\nu,v)\Big|_{\tau=0} =&\ \frac{\partial}{\partial\tau}L(\alpha_0+\tau\nu,v)\Big|_{\tau=0} - \frac{\partial}{\partial\tau}L(\alpha_0+\tau\nu,v_\rho)\Big|_{\tau=0}\\ =&\ -2\,\mathrm{E}[\alpha_0(X)\,(\mathrm{E}[v(W)\mid X]-v_\rho(X))\,\nu(X)]\end{aligned}$$

Define:

$$\hat{\hat{v}}_\rho(X)=\mathrm{E}[\hat{v}_\rho(W)\mid X]$$

By a Taylor expansion, with $\nu=\alpha-\alpha_0$ and some $\bar{\tau}\in[0,1]$:

$$\begin{aligned}L(\alpha,\hat{v}_\rho)-L(\alpha_0,\hat{v}_\rho) =&\ \frac{\partial}{\partial\tau}L(\alpha_0+\tau\nu,\hat{v}_\rho)\Big|_{\tau=0}+\frac{\partial^2}{\partial\tau^2}L(\alpha_0+\tau\nu,\hat{v}_\rho)\Big|_{\tau=\bar{\tau}}\\ =&\ -2\,\mathrm{E}[\alpha_0(X)(\hat{\hat{v}}_\rho(X)-v_\rho(X))\,\nu(X)]-2\,\mathrm{E}[\hat{\hat{v}}_\rho(X)\nu(X)^2]\end{aligned}$$

By a Cauchy-Schwarz inequality and an AM-GM inequality and since $\alpha_0(X)$ is bounded:

$$|2\,\mathrm{E}[\alpha_0(X)(\hat{\hat{v}}_\rho(X)-v_\rho(X))\,\nu(X)]|\le\ C\,\|\hat{\hat{v}}_\rho-v_\rho\|\,\|\nu\|\le\frac{C^2}{2\lambda}\|\hat{\hat{v}}_\rho-v_\rho\|^2+\frac{\lambda}{2}\|\nu\|^2$$

Moreover, by our assumption $|\hat{v}_\rho(W)|\le C\implies|\hat{\hat{v}}_\rho(X)|\le C$. Thus:

$$\begin{aligned}2C\|\nu\|^2\ge-2\,\mathrm{E}[\hat{\hat{v}}_\rho(X)\nu(X)^2]\ge&\ -2\,\mathrm{E}[v_\rho(X)\nu(X)^2]-|\,\mathrm{E}[(\hat{\hat{v}}_\rho(X)-v_\rho(X))\nu(X)^2]|\\ \ge&\ -2\,\mathrm{E}[v_\rho(X)\nu(X)^2]-\|\hat{\hat{v}}_\rho-v_\rho\|\sqrt{\mathrm{E}[\nu(X)^4]}\\ \ge&\ -2\,\mathrm{E}[v_\rho(X)\nu(X)^2]-C\|\hat{\hat{v}}_\rho-v_\rho\|\sqrt{\mathrm{E}[\nu(X)^2]}\\ =&\ -2\,\mathrm{E}[v_\rho(X)\nu(X)^2]-C\|\hat{\hat{v}}_\rho-v_\rho\|\,\|\nu\|\end{aligned}$$

Let $\nu_*=\alpha-\alpha_*$ and $\nu_0=\alpha_*-\alpha_0$, such that $\nu=\nu_*+\nu_0$. Since $-v_\rho(X)\ge0$, and $-\,\mathrm{E}[v_\rho(X)\nu_*(X)^2]\ge\lambda\,\mathrm{E}[\nu_*(X)^2]$, we have:

$$\begin{aligned}-\,\mathrm{E}[v_\rho(X)\nu(X)^2]=&\ \mathrm{E}[|v_\rho(X)|\,\nu_*(X)^2]+2\,\mathrm{E}[|v_\rho(X)|\,\nu_*(X)\nu_0(X)]+\mathrm{E}[|v_\rho(X)|\,\nu_0(X)^2]\\ \ge&\ \mathrm{E}[|v_\rho(X)|\,\nu_*(X)^2]-2\,\mathrm{E}[|v_\rho(X)|\,|\nu_*(X)\nu_0(X)|]\\ \ge&\ \mathrm{E}[|v_\rho(X)|\,\nu_*(X)^2]-\frac{1}{2}\,\mathrm{E}[|v_\rho(X)|\nu_*(X)^2]-2\,\mathrm{E}[|v_\rho(X)|\nu_0(X)^2]\\ \ge&\ \mathrm{E}[|v_\rho(X)|\,\nu_*(X)^2]-\frac{1}{2}\,\mathrm{E}[|v_\rho(X)|\nu_*(X)^2]-2C\,\mathrm{E}[\nu_0(X)^2]\\ \ge&\ \frac{1}{2}\,\mathrm{E}[|v_\rho(X)|\,\nu_*(X)^2]-2C\,\mathrm{E}[\nu_0(X)^2]\\ \ge&\ \frac{\lambda}{2}\,\mathrm{E}[\nu_*(X)^2]-2C\,\mathrm{E}[\nu_0(X)^2]\end{aligned}$$

Combining the last two inequalities:

$$\begin{aligned}2C\|\nu\|^2\ge-2\,\mathrm{E}[\hat{\hat{v}}_\rho(X)\nu(X)^2]\ge&\ \lambda\|\nu_*\|^2-4C\|\nu_0\|^2-C\,\|\hat{v}_\rho-v_\rho\|\,\|\nu\|\\ \ge&\ \lambda\|\nu_*\|^2-4C\|\nu_0\|^2-C\,\|\hat{\hat{v}}_\rho-v_\rho\|\,(\|\nu_*\|+\|\nu_0\|)\\ \ge&\ \frac{\lambda}{2}\|\nu_*\|^2-5C\|\nu_0\|^2-\left(C+\frac{C^2}{2\lambda}\right)\|\hat{\hat{v}}_\rho-v_\rho\|^2\end{aligned}$$

We conclude that for some constant $C$, for any $\alpha \in \Gamma$:

(A.5)
$$\frac{C}{\lambda}\|\hat{\hat{v}}_\rho - v_\rho\|^2 + C\|\nu\|^2 \geq L(\alpha, \hat{v}_\rho) - L(\alpha_0, \hat{v}_\rho) \geq \frac{\lambda}{2}\|\nu_*\|^2 - C\|\nu_0\|^2 - \frac{C}{\lambda}\|\hat{\hat{v}}_\rho - v_\rho\|^2$$

Next, by Lemma 11 of Foster and Syrgkanis (2019), the fact that $-2m(W,\alpha) - v(W)\alpha(X)^2$ is Lipschitz with respect to the vector $(m(W,\alpha), \sqrt{|v(W)|}\alpha(X))$ and by choosing $\delta := \delta_n + c_0\sqrt{\ln(c_1/\zeta)/n}$, where $\delta_n$ is an upper bound on the critical radius of star$(\sqrt{\mathcal{V}_n} \cdot (\mathcal{A} - \alpha_*))$ and star$(m \circ \mathcal{A} - m \circ \alpha_*)$, then with probability $1 - \zeta$, for all $\alpha \in \mathcal{A}_n$ and $v \in \mathcal{V}_n$:

$$\begin{aligned}
&|L_n(\alpha, v) - L_n(\alpha_*, v) - (L(\alpha, v) - L(\alpha_*, v))| \\
&\quad \leq O\left(\delta\left(\sqrt{\mathrm{E}[|v(W)|(\alpha(X) - \alpha_*(X))^2]} + \sqrt{\mathrm{E}[(m(W,\alpha) - m(W,\alpha_*))^2]}\right) + \delta^2\right)
\end{aligned}$$

By MSE-continuity of the moment and the fact that $|\hat{v}_\rho(W)|$ is upper bounded by a constant:

$$|L_n(\alpha, \hat{v}_\rho) - L_n(\alpha_*, \hat{v}_\rho) - (L(\alpha, \hat{v}_\rho) - L(\alpha_*, \hat{v}_\rho))| = O\left(\delta\sqrt{M}\|\alpha - \alpha_*\| + \delta^2\right) =: \epsilon_1(\alpha)$$

Finally, since $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}_n} \hat{L}_n(\alpha)$, we have that:

$$L_n(\hat{\alpha}, \hat{v}_\rho) - L_n(\alpha_*, \hat{v}_\rho) \leq 0$$

Combined with the concentration inequality, yields:

$$L(\hat{\alpha}, \hat{v}_\rho) - L(\alpha_*, \hat{v}_\rho) \leq L(\hat{\alpha}, \hat{v}_\rho) - L(\alpha_*, \hat{v}_\rho) - (L_n(\hat{\alpha}, \hat{v}_\rho) - L_n(\alpha_*, \hat{v}_\rho)) \leq \epsilon_1(\hat{\alpha})$$

Invoking Equation (A.5) at $\alpha = \hat{\alpha}$:

$$\begin{aligned}
\frac{\lambda}{2}\|\hat{\alpha} - \alpha_*\|^2 &\leq L(\hat{\alpha}, \hat{v}_\rho) - L(\alpha_0, \hat{v}_\rho) + C\|\nu_0\|^2 + \frac{C}{\lambda}\|\hat{\hat{v}}_\rho - v_\rho\|^2 \\
&\leq L(\hat{\alpha}, \hat{v}_\rho) - L(\alpha_*, \hat{v}_\rho) + L(\alpha_*, \hat{v}_\rho) - L(\alpha_0, \hat{v}_\rho) + C\|\nu_0\|^2 + \frac{C}{\lambda}\|\hat{\hat{v}}_\rho - v_\rho\|^2
\end{aligned}$$

(by Equation (A.5) at $\alpha = \alpha_*$)

$$\begin{aligned}
&\leq L(\hat{\alpha}, \hat{v}_\rho) - L(\alpha_*, \hat{v}_\rho) + 2C\|\nu_0\|^2 + \frac{2C}{\lambda}\|\hat{\hat{v}}_\rho - v_\rho\|^2 \\
&\leq \epsilon_1(\hat{\alpha}) + 2C\|\alpha_* - \alpha_0\|^2 + \frac{2C}{\lambda}\|\hat{\hat{v}}_\rho - v_\rho\|^2
\end{aligned}$$

By the AM-GM inequality:

$$\frac{\lambda}{2}\|\hat{\alpha} - \alpha_*\|^2 \leq \frac{\lambda}{4}\|\hat{\alpha} - \alpha_*\|^2 + O\left(\frac{M}{\lambda}\delta^2 + \|\alpha_* - \alpha_0\|^2 + \frac{1}{\lambda}\|\hat{\hat{v}}_\rho - v_\rho\|^2\right)$$

Re-arranging yields:

$$\|\hat{\alpha} - \alpha_*\|^2 \leq O\left(\frac{M}{\lambda^2}\delta^2 + \frac{1}{\lambda}\|\alpha_* - \alpha_0\|^2 + \frac{1}{\lambda^2}\|\hat{\hat{v}}_\rho - v_\rho\|^2\right)$$

Finally, note that:

$$\begin{aligned}
\|\hat{\alpha} - \alpha_0\|^2 &\leq 2\|\hat{\alpha} - \alpha_*\|^2 + 2\|\alpha_* - \alpha_0\|^2 \\
&\leq O\left(\frac{M}{\lambda^2}\delta^2 + \left(1 + \frac{1}{\lambda}\right)\|\alpha_* - \alpha_0\|^2 + \frac{1}{\lambda^2}\|\hat{\hat{v}}_\rho - v_\rho\|^2\right). \qquad \square
\end{aligned}$$

Finally, note that by definition $\|\hat{\hat{v}}_\rho - v_\rho\| = \|\hat{v}_\rho - v_\rho\|_X$.

PROOF OF THEOREM 3.3. Throughout this proof, let $C > 0$ denote a generic constant (possibly different each time it appears). To show the first conclusion we verify Assumptions 1–3 of (Chernozhukov et al., 2022b, CEINR), with $g(w, \gamma, \theta)$ and $\phi(w, \gamma, \alpha, \theta)$ there given by $m(w, \gamma) - \theta$ and $\alpha(x)\rho(w, \gamma)$ respectively. By Assumption 1 and $\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0$,

$$\int \|g(w, \hat{\gamma}_\ell, \theta_0) - g(w, \gamma_0, \theta_0)\|^2 F_0(dw) = \int \{m(w, \hat{\gamma}_\ell) - m(w, \gamma_0)\}^2 F_0(dw)$$

$$\leq M \|\hat{\gamma}_\ell - \gamma_0\|^2 \xrightarrow{p} 0. \tag{A.6}$$

By Assumption 7 (i), (iii) and $\|\hat{\gamma}_\ell - \gamma_0\| \xrightarrow{p} 0$,

$$\int \|\phi(w, \hat{\gamma}_\ell, \alpha_0, \theta_0) - \phi(w, \gamma_0, \alpha_0, \theta_0)\|^2 F_0(dw) = \int \alpha_0(x)^2 \{\rho(w, \hat{\gamma}_\ell) - \rho(w, \gamma_0)\}^2 F_0(dw)$$

$$\leq C \int \{\rho(w, \hat{\gamma}_\ell) - \rho(w, \gamma_0)\}^2 F_0(dw) \xrightarrow{p} 0. \tag{A.7}$$

By Assumption 7 (i), since $\|\hat{\alpha}_\ell - \alpha_0\| \xrightarrow{p} 0$, iterated expectations gives

$$\int \left\|\phi(w, \gamma_0, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \phi(w, \gamma_0, \alpha_0, \theta_0)\right\|^2 F_0(dw) = \int \{\hat{\alpha}_\ell(x) - \alpha_0(x)\}^2 \rho(W, \gamma_0)^2 F_0(dw)$$

$$\leq C \|\hat{\alpha}_\ell - \alpha_0\|^2 \xrightarrow{p} 0. \tag{A.8}$$

Therefore, Assumption 1 (i), (ii), and (iii) of CEINR is satisfied.

Next note that:

$$\hat{\Delta}_\ell(w) := \phi(w, \hat{\gamma}_\ell, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \phi(w, \gamma_0, \hat{\alpha}_\ell, \tilde{\theta}_\ell) - \phi(w, \hat{\gamma}_\ell, \alpha_0, \theta_0) + \phi(w, \gamma_0, \alpha_0, \theta_0)$$
$$= \{\hat{\alpha}_\ell(x) - \alpha_0(x)\}\{\rho(w, \hat{\gamma}_\ell) - \rho(w, \gamma_0)\}.$$

Let $\bar{\rho}(X, \gamma) = \mathrm{E}[\rho(W, \gamma) \mid X]$. Then by iterated expectations, the Cauchy-Schwartz inequality, and Assumptions 7 and 4,

$$\int \hat{\Delta}_\ell(w) F_0(dw) = \int \{\hat{\alpha}_\ell(x) - \alpha_0(x)\}\{\rho(w, \hat{\gamma}_\ell) - \rho(w, \gamma_0)\} F_0(dx)$$
$$\leq \|\hat{\alpha}_\ell - \alpha_0\| \, \|\bar{\rho}(\cdot, \hat{\gamma}_\ell) - \bar{\rho}(\cdot, \gamma_0)\|$$

$$\leq C \|\hat{\alpha}_\ell - \alpha_0\| \, \|\hat{\gamma}_\ell - \gamma_0\| = o_p(n^{-1/2}). \tag{A.9}$$

Since $\hat{\alpha}_\ell(x)$ and $\alpha_0(x)$ are bounded,

$$\int \left\|\hat{\Delta}_\ell(w)\right\|^2 F_0(dw) = \int \{\hat{\alpha}_\ell(x) - \alpha_0(x)\}^2 \{\rho(w, \hat{\gamma}_\ell) - \rho(w, \gamma_0)\}^2 F_0(dw)$$

$$\leq C \|\hat{\gamma}_\ell - \gamma_0\|^2 \xrightarrow{p} 0, \tag{A.10}$$

as in equation (A.7). By equations (A.9) and (A.10) it follows that Assumption 2 (i) of CEINR is satisfied.

Assumption 3 of CEINR follows by Assumption 8. Therefore each of Assumptions 1–3 of CEINR are satisfied, so the first conclusion follows by Lemma 15 of CEINR and the Lindeberg-Lévy central limit theorem.

Finally, by the first conclusion $\hat{\theta} \xrightarrow{p} \theta_0$ and thus

$$\int \{m(w, \hat{\gamma}_\ell) - \hat{\theta} - m(w, \gamma_0) + \theta_0\}^2 F_0(dw) \xrightarrow{p} 0,$$

so that the hypotheses of Lemma 16 of CEINR are satisfied, giving the second conclusion. □

PROOF OF 3.4. he proof would be identical to Theorem 3.2, with the only difference being that $v$ now contains two nuisance parameters $(D, v_\rho)$ and:

$$\begin{aligned}\frac{\partial}{\partial\tau}L(\alpha_0+\tau\nu,\hat{v})\bigg|_{\tau=0} &= \frac{\partial}{\partial\tau}L(\alpha_0+\tau\nu,\hat{v})\bigg|_{\tau=0} - \frac{\partial}{\partial\tau}L(\alpha_0+\tau\nu,v_0)\bigg|_{\tau=0}\\ &= \mathrm{E}[\hat{D}(W;\nu)-D(W;\nu)] - 2\,\mathrm{E}[\alpha_0(X)\,(\mathrm{E}[v(W)\mid X]-v_\rho(X))\,\nu(X)]\end{aligned}$$

The first part can then be bounded as:

$$|\,\mathrm{E}[\hat{D}(W;\nu)-D(W;\nu)]| \le \epsilon_{mn}\|\nu\|$$

The proof then follows identically to the proof of Theorem 3.2. □

PROOF OF 3.5. It follows exactly as in the proof of Lemma 15 of CEINR that for each $j$

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}[\hat{\alpha}_{j\ell}(X_{ji})\rho_j(W_i,\hat{\gamma}_{j\ell})-\alpha_{j0}(X_{ji})\rho_j(W_i,\gamma_{j0})]\\ &=\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}[\hat{\alpha}_{j\ell}(X_{ji})-\alpha_{j0}(X_{ji})]\rho(W_i,\gamma_{j0})+\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}\alpha_{j0}(X_{ji})[\rho(W_i,\hat{\gamma}_{j\ell})-\rho(W_i,\gamma_{j0})]+o_p(1)\\ &=\frac{n_\ell}{\sqrt{n}}\int\alpha_{j0}(x_j)[\rho(w,\hat{\gamma}_{j\ell})-\rho(w,\gamma_{j0})]F_0(dw)+o_p(1),\end{aligned}$$

$$\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}[m(W_i,\hat{\gamma}_\ell)-m(W_i,\gamma_0)]=\frac{n_\ell}{\sqrt{n}}\int[m(w,\hat{\gamma}_\ell)-\theta_0]F_0(dw)+o_p(1).$$

Also by Assumption 11 it is the case that $\|\hat{\gamma}_{j\ell}-\gamma_{j0}\|<\varepsilon$ for all $j$ with probability approaching one, so that by the triangle inequality and Assumption 9 iii) we have

$$\begin{aligned}&\left|\frac{1}{\sqrt{n}}\sum_{i\in I_\ell}[m(W_i,\hat{\gamma}_\ell)-\theta_0+\sum_{j=1}^{J}\hat{\alpha}_{j\ell}(X_{ji})\rho_j(W_i,\hat{\gamma}_{j\ell})-\psi(W_i,\gamma_0,\alpha_0,\theta_0)]\right|\\ &\le\frac{n_\ell}{\sqrt{n}}\left|\int[m(w,\hat{\gamma}_\ell)-\theta_0+\sum_{j=1}^{J}\alpha_{j0}(x_j)\rho_j(w,\hat{\gamma}_{j\ell})]F_0(dw)\right|+o_p(1)\\ &\le\sqrt{n}C\sum_{j=1}^{J}\|\hat{\gamma}_j-\gamma_{j0}\|^2=\sqrt{n}o_p((n^{-1/4})^2)=o_p(1),\end{aligned}$$

where

$$\psi(w,\gamma_0,\alpha_0,\theta_0):=m(w,\gamma_0)-\theta_0+\sum_{j=1}^{J}\alpha_{j0}(x_j)\rho_j(w,\gamma_{j0}).$$

The first conclusion then follows by the triangle inequality and the central limit theorem. The second conclusion follows in analogous way, treating each $j$ separately, using the arguments in Lemma 16 of CEINR. □

## APPENDIX B: HYPERPARAMETERS

Here we give details on the hyperparameters in the architecture and training of neural nets used in our main specification in Sections 4 and 5.

The regression learner $\hat{\gamma}$ and the debiasing function learner $\hat{\alpha}$ are both parametrized as neural nets with two hidden layers and ReLU activation function. The width of the hidden layers, the learning rate and the training L2 penalty are tuned on a grid based on the out-of-sample loss on a test set (30% of the data). We train the parameters of the neural net using the Adam optimizer of PyTorch, with a batch size of 128. During training, we randomly drop some layers out with a dropout probability of 0.05. We also do early stopping to avoid overfitting, where we end the training process if the loss on a separate validation set (also 30% of the data) decreases by less than $10^{-5}$ in 5 consecutive rounds.

## REFERENCES

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A* 381–393.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705 – 1732. https://doi.org/10.1214/08-AOS620

BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and adaptive estimation for semiparametric models* **4**. Springer.

CHEN, X. and WHITE, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory* **45** 682–691.

CHERNOZHUKOV, V., NEWEY, W. and SINGH, R. (2019). De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers. https://doi.org/10.48550/ARXIV.1802.08667

CHERNOZHUKOV, V., NEWEY, W. K. and SINGH, R. (2022). Automatic Debiased Machine Learning of Causal and Structural Effects. *Econometrica* **90** 967-1027. https://doi.org/10.3982/ECTA18515

CHERNOZHUKOV, V., NEWEY, W., SINGH, R. and SYRGKANIS, V. (2020). Adversarial Estimation of Riesz Representers. *arXiv preprint arXiv:2101.00009*.

CHERNOZHUKOV, V., NEWEY, W. K., QUINTAS-MARTINEZ, V. and SYRGKANIS, V. (2022a). RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests. In *ICML 2022*.

CHERNOZHUKOV, V., ESCANCIANO, J. C., ICHIMURA, H., NEWEY, W. K. and ROBINS, J. M. (2022b). Locally robust semiparametric estimation. *Econometrica* **90** 1501–1535.

FARRELL, M. H., LIANG, T. and MISRA, S. (2021a). Deep Neural Networks for Estimation and Inference. *Econometrica* **89** 181-213.

FARRELL, M. H., LIANG, T. and MISRA, S. (2021b). Deep Learning for Individual Heterogeneity: An Automatic Inference Framework. *arXiv preprint arXiv:2010.14694*.

FOSTER, D. J. and SYRGKANIS, V. (2019). Orthogonal Statistical Learning. *arXiv preprint, arXiv:1901.09036*.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1978). On the nonparametric estimation of functionals. In *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics* 41–51.

HE, X., PAN, X., TAN, K. M. and ZHOU, W.-X. (2023). Smoothed quantile regression with large-scale inference. *Journal of Econometrics* **232** 367–388.

HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics* **49** 3206 – 3227. https://doi.org/10.1214/21-AOS2080

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* **47** 663–685.

ICHIMURA, H. and NEWEY, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics* **13** 29–61.

IMBENS, G. W. and NEWEY, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* **77** 1481–1512.

LEVIT, B. Y. (1975). On efficiency of a class of non-parametric estimates. *Teoriya Veroyatnostei i ee Primeneniya* **20** 738–754.

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Optimal individualized treatments in resource-limited settings. *The International Journal of Biostatistics* **12** 283–303.

LUO, Y., SPINDLER, M. and KÜCK, J. (2022). High-Dimensional $L_2$Boosting: Rate of Convergence. *arXiv:1602.08927*.

MUNNELL, A. H., TOOTELL, G. M., BROWNE, L. E. and MCENEANEY, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review* 25–53.

NELDER, J. A. and WEDDERBURN, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135** 370–384.

NEWEY, W. K. (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica* **62** 1349–1382.

NEWEY, W. K., HSIEH, F. and ROBINS, J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* **72** 947–962.

NEWEY, W. K. and ROBINS, J. R. (2018). Cross-fitting and Fast Remainder Rates for Semiparametric Estimation. *arXiv preprint arXiv:1801.09138*.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90** 122–129.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70** 41–55.

SCHMIDT-HIEBER, J. (2020). Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function. *The Annals of Statistics* **48** 1875–1897.

SINGH, R. (2021). Debiased Kernel Methods. *arXiv preprint arXiv:2102.11076*.

SINGH, R. and SUN, L. (2023). Double robustness for complier parameters and a semi-parametric test for complier characteristics. *The Econometrics Journal* utad019. https://doi.org/10.1093/ectj/utad019

SINGH, R., XU, L. and GRETTON, A. (2022). Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves. *arXiv preprint arXiv:2010.04855*.

SYRGKANIS, V. and ZAMPETAKIS, M. (2020). Estimation and inference with trees and forests in high dimensions. In *Conference on learning theory* 3453–3454. PMLR.

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data* **4**. Springer.

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**.

VAN DER VAART, A. (1991). On differentiable functionals. *The Annals of Statistics* 178–204.

VAN DER VAART, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge university press.

YAROTSKY, D. (2018). Optimal Approximation of Continuous Functions by Very Deep ReLU Networks. In *Conference on Learning Theory* 639–649. PMLR.