The Econometrics Journal

Econometrics Journal (2022), volume **00**, pp. 1–26.
https://doi.org/10.1093/ectj/utac002

Debiased machine learning of global and local parameters using regularized Riesz representers

VICTOR CHERNOZHUKOV, WHITNEY K. NEWEY AND RAHUL SINGH

Deparment of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge MA 02142, USA.
Email: vchern@mit.edu, wnewey@mit.edu, rahul.singh@mit.edu

First version received: 23 February 2021; final version accepted: 21 October 2021.

Summary: We provide adaptive inference methods, based on ℓ_1 regularization, for regular (semiparametric) and nonregular (nonparametric) linear functionals of the conditional expectation function. Examples of regular functionals include average treatment effects, policy effects, and derivatives. Examples of nonregular functionals include average treatment effects, policy effects, and derivatives conditional on a covariate subvector fixed at a point. We construct a Neyman orthogonal equation for the target parameter that is approximately invariant to small perturbations of the nuisance parameters. To achieve this property, we include the Riesz representer for the functional as an additional nuisance parameter. Our analysis yields weak 'double sparsity robustness': either the approximation to the regression or the approximation to the representer can be 'completely dense' as long as the other is sufficiently 'sparse'. Our main results are nonasymptotic and imply asymptotic uniform validity over large classes of models, translating into honest confidence bands for both global and local parameters.

Keywords: *Neyman orthogonality*, *Gaussian approximation*, *sparsity*.

JEL codes: *C14*, *C21*, *C45*.

1. INTRODUCTION

Many statistical objects of interest can be expressed as a linear functional of a regression function (or projection, more generally). Examples include global parameters: average treatment effects, policy effects from changing the distribution of or transporting regressors, and average directional derivatives, as well as their local versions defined by taking averages over regions of shrinking volume. This variety of important examples motivates the problem of learning linear functionals of regressions. Global parameters are typically regular (estimable at $1/\sqrt{n}$ rate), and local parameters are nonregular (estimable at slower than $1/\sqrt{n}$ rates). Global parameters can also be nonregular under weak identification (for example, in average treatment effects, when propensity scores accumulate mass near zero or one, along a given sequence of models).

Often the regression is high dimensional, depending on many variables such as covariates in a treatment effect model. Plugging a machine learner into a functional of interest can give a badly biased estimator. To avoid such bias, we use debiased/'double' machine learning (DML) based on Neyman orthogonal scores that have zero derivative with respect to each first step learner (e.g., Neyman, 1959; Belloni et al., 2014, 2015; Chernozhukov et al., 2016, 2018a; Foster and Syrgkanis, 2019). Note that the word 'double' emphasizes the connection to double robustness,

a property that orthogonal scores have in this case. Such scores are constructed by adding a bias correction term: the average product of the regression residual with a learner of the functional's Riesz representer (RR). This construction builds upon and is directly inspired by Newey (1994), where such scores arise in the computation of the semiparametric efficiency bound for regular functionals. We also remove overfitting bias (high entropy bias) by using cross-fitting, an efficient form of sample splitting, where we average over data observations different from those used by the nonparametric learners. See, for example, Schick (1986) for early use and Chernozhukov et al. (2018a) for more recent use in the context of debiased machine learning.

Using closed-form solutions for RRs in several examples, Chernozhukov et al. (2016, 2018a) defined DML estimators in high-dimensional settings and established their good properties. In comparison, the new approach proposed in this paper has the following advantages and some limitations:

(1) We provide a novel algorithm based on ℓ_1 regularization to *automatically estimate* the RR from the empirical analog of equations that implicitly characterize it.
(2) Even when a closed-form solution for the RR is available, the method avoids estimating each of its components. For example, the method avoids explicit density derivative estimation for the average derivative, and it avoids inverting estimated propensity scores for average treatment effects.
(3) The adaptive inference theory covers both regular objects (estimable at the $1/\sqrt{n}$ rate) and nonregular ones (with rates $L/\sqrt{n}$, where $L \to \infty$ is the operator norm of the linear functional).
(4) As far as we know, the adaptive inference theory given here is the first nonasymptotic Gaussian approximation analysis of debiased machine learning.
(5) Our approach remains interpretable under misspecification, estimating a linear functional of the projection rather than regression. (This point is made explicit in Section 4).
(6) We provide a nonasymptotic analysis when using the ℓ_1-penalized method to learn the regression, and an asymptotic analysis when using other modern machine learning estimators to learn the regression.
(7) The current analysis focuses on linear functionals. In follow-up work, Chernozhukov et al. (2018) extend the approach to nonlinear functionals through a linearization.

This paper is a revised version of Chernozhukov et al. (2018d) that gave an algorithm based on ℓ_1 regularization for automatically estimating the RR. This version is distinguished from Chernozhukov et al. (2018d), Chernozhukov et al. (2018a), Chernozhukov et al. (2016), and Chernozhukov et al. (2018c) in covering local objects that are estimated at a rate slower than $1/\sqrt{n}$. Providing debiased machine learning for such local objects is an important contribution of this paper.

Sections 2 and 3 present the main ideas for a general audience. In Section 2, we define global, local, and perfectly localized linear functionals of the regression, and provide orthogonal representations for these functionals. In Section 3, we present two empirical examples: local and global average treatment effects, and local and global average derivatives.

Sections 4 and 5 are theoretical. In Section 4, we provide estimation theory, demonstrating concentration and approximate Gaussianity of the DML estimator with regression and RR estimated via regularized moment conditions. We provide rates of convergence for estimating the RR, giving both fast rates under approximate sparsity. In Section 5, we demonstrate asymptotic

consistency and Gaussianity of the DML estimator with regression estimated via general machine learning.

The supplement provides supporting material. In Section S1, we give a detailed account of how our work relates to previous and contemporary work. In Section S2, we review prelimaries of functional analysis. In Section S3, we analyse the structure of the leading examples, providing bounds on operator norm, variance of the score, and kurtosis. Finally, we provide proofs for each section.

2. OVERVIEW OF TARGET FUNCTIONALS, ORTHOGONAL REPRESENTATION, ESTIMATION, AND INFERENCE

2.1. Target functionals

We consider a random element W with distribution P taking values w in its support $\mathcal{W}$. Denote the $L^q(P)$ norm of a measurable function $f : \mathcal{W} \to \mathbb{R}$ and also the $L^q(P)$ norm of random variable $f(W)$ by $\|f\|_{P,q} = \|f(W)\|_{P,q}$. For a differentiable map $x \mapsto g(x)$, from $\mathbb{R}^d$ to $\mathbb{R}^k$, we use $\partial_{x'} g$ to abbreviate the partial derivatives $(\partial/\partial x')g(x)$, and we use $\partial_{x'} g(x_0)$ to mean $\partial_{x'} g(x) \mid_{x=x_0}$, etc. We use x' to denote the transpose of a column vector x.

Let (Y, X) denote a random subvector of W taking values in their support sets, $y \in \mathcal{Y} \subset \mathbb{R}$ and $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$, where $d_x = \infty$ is allowed. Let F denote the law of X. We define

$$x \mapsto \gamma_0^\star(x) := \mathrm{E}[Y \mid X = x],$$

as the unknown regression function of Y on X. We consider the convex parameter space Γ_0 for $\gamma_0^\star$ with elements γ. (Later, in the theoretical sections, we generalize and replace the regression function by a projection).

Our goal is to construct high-quality inference methods for real-valued linear functionals of $\gamma_0^\star$. To present examples below we need to endow $\gamma_0^\star$ with a causal interpretation, which requires us to assume that it is a structural function, invariant to the changes in the distribution of X under policies described below. This property is not guaranteed for an arbitrary regression problem. For the reader who is unfamiliar with these concepts, we note that a simple sufficient condition for invariance is follows: given a stochastic process $x \mapsto Y(x)$, called potential outcomes or structural function, vector X is generated to follow distribution F independently of $x \mapsto Y(x)$ and Y is generated as $Y = Y(X)$. In this case we have $\gamma_0^\star(x) = \mathrm{E}Y(x)$ for any F. This condition is conventionally called exogeneity in econometrics and random assignment in statistics. The measurability requirement here is that $(x, \omega) \mapsto Y(x, \omega)$ is a measurable map. We refer to Imbens and Rubin (2015), Hernan and Robins (2020), and Peters et al. (2017) for the relevant formalizations that enable causal interpretation.

EXAMPLE 2.1 (AVERAGE TREATMENT EFFECT). *Let $X = (D, Z)$ and $\gamma_0^\star(X) = \gamma_0^\star(D, Z)$, where $D \in \{0, 1\}$ is the indicator of the receipt of the treatment. Define*

$$\theta_0^\star = \int (\gamma_0^\star(1, z) - \gamma_0^\star(0, z))\ell(x) dF(x),$$

where $x \mapsto \ell(x)$ is a weighting function. This statistical parameter is a weighted average treatment effect under the standard conditional exogeneity assumption, which guarantees that $\gamma_0^\star$ is invariant to changes in the distributions of D conditional on Z. The assumption requires D to be independent of the potential outcome process $d \mapsto Y(d, Z)$ and outcome to be generated

as $Y = Y(D, Z)$, so that $\gamma_0^\star(d, z) = \mathrm{E}[Y(d, Z) \mid Z = z]$. Here $\gamma_0^\star$ is invariant to changes in the conditional distributions of D, but not to the changes in the distribution of Z.

Here and below, a weighting function is a measurable function $x \mapsto \ell(x)$ such that $\int \ell dF = 1$ and $\int \ell^2 dF < \infty$. In this example, setting

(*i*) $\ell(x) = 1$ gives average treatment effect in the entire population,
(*ii*) $\ell(x) = 1(d = 1)/P(D = 1)$ gives the average treatment effect for the *treated* population,
(*iii*) $\ell(x) = 1(z \in N)/P(Z \in N)$ the average treatment effect conditional on the covariates Z being in the group or neighborhood N,

and so on. We can model small neighbourhoods N as shrinking in volume with the sample size. The local weighting and kernel weighting discussed below are applicable to all key examples. Moreover, they are combinable with other weighting functions so that, for example, we can target inference on local average treatment effects for the treated.

EXAMPLE 2.2 (POLICY EFFECT FROM CHANGING DISTRIBUTION OF X). *The average causal effect of the policy that shifts the distribution of covariates from F_0 to F_1 with the support contained in $\mathcal{X}$, when $\gamma_0^\star$ is invariant over $\{F, F_0, F_1\}$, for the weighting function $x \mapsto \ell(x)$, is given by:*

$$\theta_0^\star = \int \gamma_0^\star(x)\ell(x)dG(x); \quad G(x) = F_1(x) - F_0(x).$$

Exogeneity is a sufficient condition for the stated invariance of $\gamma_0^\star$.

EXAMPLE 2.3 (POLICY EFFECT FROM TRANSPORTING X). *A weighted average effect of changing covariates X according to a transport map $X \mapsto T(X)$, where T is deterministic measurable map from $\mathcal{X}$ to $\mathcal{X}$, with the weighting function $x \mapsto \ell(x)$, is given by:*

$$\theta_0^\star = \int [\gamma_0^\star(T(x)) - \gamma_0^\star(x)]\ell(x)dF(x).$$

This has a causal interpretation if the policy induces the equivariant change in the regression function, namely the outcome $\tilde{Y}$ under the policy obeys $\mathrm{E}[\tilde{Y}|X] = \gamma_0^\star(T(X))$. Exogeneity is a sufficient condition.

EXAMPLE 2.4 (AVERAGE DIRECTIONAL DERIVATIVE). *In the same settings as the previous example, a weighted average derivative of a continuously differentiable γ_0 with respect to component vector d in the direction $d \mapsto t(x)$ and weighed by $x \mapsto \ell(x)$ is the linear functional of the form:*

$$\theta_0^\star = \int \ell(x)t(x)'\partial_d\gamma_0^\star(d, z)dF(x).$$

In causal analysis, $\theta_0^\star$ is an approximation to $1/r$ times the average causal effect of the policy that shifts the distribution of covariates via the map $X = (D, Z) \mapsto T(X) = (D + rt(X), Z)$ for small r, weighted by $\ell(X)$. Here we require that $(d, x) \mapsto \partial_d\gamma_0^\star(x)$ exists and is continuous on $\mathcal{X}$.

In this example, consider the case when $X = (D, Z)$ consists of continuous treatment variable D and covariates Z. Further suppose $\ell(x) = \ell(d)$ and $t(x) = 1$. Then the parameter of interest is $\theta_0^* = E[\ell(D)T(D)]$, where $T(d) = E[\partial_d\gamma_0^*(D, Z)|D = d]$. When $Y = Y(D)$ for a potential outcome process $Y(d)$ that is independent of treatment D conditional on the covariates Z and differentiable in d, it was shown by Altonji and Matzkin (2005) and Florens et al. (2008)

that $T(d) = E[\partial_d Y(D)|D = d]$, which is an average treatment effect on the treated. Thus θ_0^* is a weighted average of the effect of treatment on the treated and would be equal to $T(d)$ for the perfectly localized $\ell(d) = 1(D = d)/f_D(d)$, where $f_D(d)$ is the pdf of D. Also for $\ell(d) \equiv 1$, Imbens and Newey (2009) showed that $\theta_0^* = E[T(D)] = E[\partial_d Y(D)]$, which is an average treatment effect. See also Rothenhäusler and Yu (2019).

In Example 2.4, we consider the case where the variable of differentiation is also the variable of localization. As explained above, this case corresponds to effects of continuous treatments, and it turns out to require extra care in Section S3. The other possible case is where the variable of differentiation is different from the variable of localization. Such a case turns out to be simpler and is handled by similar arguments as Examples 2.1, 2.2, and 2.3 in Section S3.

All of these statistical parameters play an important role in causal inference, counterfactual decompositions, and predictive analyses. Introduction of the weighting function $\ell(X)$ allows us to study subgroup effects and local effects, and these will be covered by our nonasymptotic results and asymptotic results. All of the above examples can be viewed as real-valued linear functionals of the regression function.

DEFINITION 2.1 (TARGET PARAMETER). *Our target is the real-valued linear functional of* $\gamma_0^\star$*:*

$$\theta_0^\star = \theta(\gamma_0^\star), \text{ where } \gamma \mapsto \theta(\gamma) := \mathrm{E} m(W, \gamma), \tag{2.1}$$

$\gamma \mapsto m(w, \gamma)$ *is a linear operator for each* $w \in \mathcal{W}$*, defined on* $\Gamma = \mathrm{span}(\Gamma_0)$*, and the map* $w \mapsto m(w, \gamma)$ *is measurable with finite second moment under P for each* $\gamma \in \Gamma$*.*

The linear operator $\gamma \mapsto \theta(\gamma)$ has the following generating function m in these examples:

2.1 $m(w, \gamma) = (\gamma(1, z) - \gamma(0, z))\ell(x)$;
2.2 $m(w, \gamma) = m(\gamma) = \int \gamma(x)\ell(x)dG(x)$; $G(x) = F_1(x) - F_0(x)$;
2.3 $m(w, \gamma) = \ell(x)(\gamma(T(x)) - \gamma(x))$;
2.4 $m(w, \gamma) = \ell(x)t(x)'\partial_d\gamma(x)$.

In these examples, we can recognize the dependency on the weighting function by writing $m(w, \gamma; \ell)$. In Examples 2.1, 2.3, and 2.4 we can decompose $m(w, \gamma; \ell) = m_0(w, \gamma)\ell(x)$.

Estimation of some parameters of the form in Definition 2.1 is very straightforward, such as $\mathrm{E}[w(X)\gamma_0(X)]$ for a known function $w(x)$. These can be estimated as the sample mean of $w(X)Y$. Such simple estimation is not possible for the causal, counterfactual parameters in Examples 2.1, 2.2, 2.3, and 2.4. The approach of this paper provides estimators for these counterfactual parameters and can be used for many others.

Our local functionals are defined by using the weight function that localizes the functionals around value d_0 of a low-dimensional vector component D. Here D is a p_1-dimensional component of vector X. We consider the weighting function

$$\ell_h(D) = \frac{1}{h^{p_1}} K\left(\frac{d_0 - D}{h}\right)/\omega, \quad \omega = \mathrm{E}\left[\frac{1}{h^{p_1}} K\left(\frac{d_0 - D}{h}\right)\right], \quad h \in \mathbb{R}_+, \tag{2.2}$$

where $K : \mathbb{R}^{p_1} \to \mathbb{R}$ in (2.2) is a kernel function of order o such that $\int K = 1$ and

$$\int (\otimes^m u)K(u)du = 0, \quad \text{for } m = 1, \ldots, \mathsf{o} - 1,$$

with its support contained in the cube $[-1, 1]^{p_1}$. The simplest example is the box kernel with $K(u) = \times_{j=1}^{p_1} 1(-1 < u_j < 1)/2$, which is of order $\mathsf{o} = 2$. To present the main results in the most clear way, we assume that ℓ_h is known, i.e., ω is known. Our main results also hold for one sided kernels. We leave to future work the application of this theory to settings with one sided limits, e.g., regression discontinuity design.

DEFINITION 2.2 (LOCAL AND LOCALIZED FUNCTIONALS). *We consider the local functional*

$$\theta(\gamma_0^\star; \ell_h) := \mathrm{E} m(W, \gamma_0^\star; \ell_h),$$

as well as the (perfectly) localized functional

$$\theta(\gamma_0^\star; \ell_0) := \lim_{h \to 0} \theta(\gamma_0^\star; \ell_h).$$

The difficulty in targeting localized functionals is that they are not pathwise differentiable. A key quantity in the analysis is the operator norm (the modulus of continuity) of $\gamma \mapsto \theta(\gamma)$ on Γ, defined as

$$L := \sup_{\gamma \in \Gamma \setminus \{0\}} |\theta(\gamma)| / \|\gamma\|_{P,2}. \tag{2.3}$$

We consider $L = \infty$ in (2.3) as nonregular cases, e.g., perfectly localized functionals. We also consider cases where $L \to \infty$ as $n \to \infty$ as nonregular. Indeed, the latter case arises from approximating the functional with $L = \infty$ by functionals where $L \to \infty$, e.g., local functionals with $h \to 0$. The $L \to \infty$ case also arises in triangular array asymptotics where P changes with n. The asymptotic thought experiment, where $L \to \infty$, approximates nonasymptotic cases where L is high. We emphasize that we derive both nonasymptotic results and their asymptotic corollaries (which lead to simplified statements conveying key qualitative features of nonasymptotic results).

2.2. *Building an orthogonal representation of the target functional*

Equation (2.1) can be thought of as a direct formulation of the target parameter. Next we introduce a dual formulation and finally an orthogonal formulation. Towards this end, we define the RR α_0.

DEFINITION 2.3 (LINEAR AND MINIMAL LINEAR REPRESENTER). *A linear representer (also called a Riesz representer) for the linear functional $\gamma \mapsto \theta(\gamma)$ is $\alpha_0 \in L^2(F)$ such that*

$$\theta(\gamma) = \mathrm{E}\gamma(X)\alpha_0(X), \text{ for all } \gamma \in \Gamma. \tag{2.4}$$

If $\alpha_0 \in \bar{\Gamma} :=$ closure(Γ) in $L^2(F)$, we call it the minimal representer and denote it by $\alpha_0^\star$; if not, we call it a representer. Any representer can be reduced to the minimal representer by projecting it onto $\bar{\Gamma}$.

A minimal linear representer exists if and only if $L < \infty$, as a consequence of the Riesz–Frechet theorem; see Lemma 2.1 below. Therefore, when $L < \infty$, we define the following dual linear representation for the target parameter

$$\theta_0^\star = \theta(\alpha_0^\star); \quad \theta(\alpha) := \mathrm{E}[\alpha(X)Y]. \tag{2.5}$$

To motivate the upcoming orthogonal representation, we note that either the direct (2.1) or the dual (2.5) identification strategies can be used for direct plug-in estimation, but this does not give good estimators, as explained in the following technical remark.

REMARK 2.1 (NON-ORTHOGONALITY OF DIRECT AND DUAL FORMULATIONS). Even if we knew expectation operator E and use $\theta(\hat{\gamma})$ or $\theta(\hat{\alpha})$ as the estimator for $\theta_0^\star$, this estimator would have high biases. Indeed, neither $\gamma \mapsto \theta(\gamma)$ nor $\alpha \mapsto \theta(\alpha)$ are orthogonal to local perturbations $h \in \Gamma$ of $\gamma_0^\star$ or $\bar{h} \in \Gamma$ of $\alpha_0^\star$, namely

$$\partial_t \theta(\gamma_0^\star + th)\Big|_{t=0} = \mathrm{E} m(W, h) \neq 0, \quad \partial_t \theta(\alpha_0^\star + t\bar{h})\Big|_{t=0} = \mathrm{E}\gamma_0^\star(X)\bar{h}(X) \neq 0.$$

Consequently, the quantities $\mathrm{E}m(W, \hat{\gamma} - \gamma_0^\star)$ and $\mathrm{E}\gamma_0^\star(\hat{\alpha} - \alpha_0^\star)$ are first order biases for $\theta(\hat{\gamma})$ and $\theta(\hat{\alpha})$. The regularized estimators $\hat{\gamma}$ or $\hat{\alpha}$ exploit structure of $\gamma_0^\star$ and $\alpha_0^\star$ to estimate them well in high-dimensional problems, but they exhibit biases that vanish at rates slower than $1/\sqrt{n}$, which makes $\theta(\hat{\gamma})$ and $\theta(\hat{\alpha})$ converge at the same slow rate.

Therefore we proceed to construct another representation for $\theta_0^\star$ that has the required Neyman orthogonality structure.

DEFINITION 2.4 (ORTHOGONAL REPRESENTATION FOR THE TARGET FUNCTIONAL). *We have*

$$\theta_0^\star = \theta(\alpha_0^\star, \gamma_0^\star); \quad \theta(\alpha, \gamma) := \mathrm{E}[m(W, \gamma) + \alpha(X)(Y - \gamma(X))], \tag{2.6}$$

where (α, γ) *are the nuisance parameters with the true value* $(\alpha_0^\star, \gamma_0^\star)$.

Unlike the direct or dual representations for the functional, this representation is Neyman orthogonal to perturbations $(\bar{h}, h) \in \Gamma^2$ of $(\alpha_0^\star, \gamma_0^\star)$ such that

$$\frac{\partial}{\partial t}\theta(\alpha_0^\star + t\bar{h}, \gamma_0^\star + th)\Big|_{t=0} = \mathrm{E} m(W, h) - \mathrm{E}\alpha_0^\star(X)h(X) + \mathrm{E}[(Y - \gamma_0^\star(X))\bar{h}(X)] = 0. \tag{2.7}$$

In fact, a stronger property holds

$$\theta(\alpha, \gamma) - \theta(\alpha_0^\star, \gamma_0^\star) = -\int (\gamma - \gamma_0^\star)(\alpha - \alpha_0^\star) dF, \tag{2.8}$$

which implies (2.7) as well as double robustness. The quantity in (2.8) is also known as the remainder of the von Mises expansion of the functional.

The Neyman orthogonality property states that the representation of the target parameter θ_0 in terms of the nuisance parameters (α, γ) is invariant to the local perturbations of the values of the nuisance parameters. This property makes the orthogonal representation an excellent basis for constructing high-quality point and interval estimators of $\theta_0^\star$ in modern high-dimensional settings when we will be plugging-in biased estimators in lieu of $\gamma_0^\star$ and $\alpha_0^\star$, where the bias occurs because of the regularization (see, e.g., Chernozhukov et al., 2016, 2018a).

Both γ_0^* and α_0^* are identified, γ_0^* as $E[Y|X]$ and α_0^* by virtue of the consistent estimator we give in Section 2.5. Identification allows us to use the orthogonal representation to estimate target parameters.

2.3. The case of finite-dimensional linear regression

It is instructive to consider the case of linear finite-dimensional regression. Consider $x \mapsto b(x) = \{b_j(x)\}_{j=1}^p$ as a p-dimensional dictionary of basis functions with $b_j \in L^2(F)$ for each $j = 1, \ldots, p$. The regression function is assumed to obey the linear functional form $\gamma_0^\star = b'\beta_0$ for some β_0. Also define

$$G = \mathrm{E} b(X)b(X)', \quad M = \mathrm{E} m(W, b).$$

First, observe that for $\gamma = b'\beta$,

$$\theta(\gamma) = \mathrm{E}m(W, b'\beta) = \mathrm{E}m(W, b)'\beta = M\beta.$$

For instance, in Examples 2.1, 2.2, 2.3, and 2.4:

$$2.1\ M = \mathrm{E}(b(1, Z) - b(0, Z))\ell(X) \quad 2.2\ M = \int b\ell(dF_1 - dF_0),$$

$$2.3\ M = \mathrm{E}(b(T(X)) - b(X))\ell(X) \quad 2.4\ M = \mathrm{E}\partial_d b(D, Z)t(X)\ell(X).$$

Second, we make a guess that the linear representer $\alpha_0^\star$ to be of the form $\alpha_0^\star(x) = b(x)'\rho_0$, for ρ_0 defined below. We can define the parameters β_0 and ρ_0 as any minimal ℓ_1-norm solutions to the system of equations:

$$\min \|\beta\|_1 + \|\rho\|_1 : \quad G\beta = \mathrm{E}Yb(X), \quad G\rho = M. \tag{2.9}$$

In particular, if G in (2.9) is full rank, the solutions are $\beta_0 = G^{-1}\mathrm{E}b(X)Y$ and $\rho_0 = G^{-1}M$.

We now verify the representation property for our guess:

$$\mathrm{E}\gamma(X)\alpha_0^\star(X) = \mathrm{E}\beta' b(X)b(X)'\rho_0 = \beta' G\rho_0 = \beta' M = \theta(\gamma),$$

for all β's and hence all γ's. The operator norm of $\theta(\gamma) = M'\beta$ is given by

$$L = \sup_{\beta\in\mathbb{R}^p\setminus\{0\}} \frac{|M'\beta|}{\sqrt{\beta' G\beta}} = \sup_{\beta\in\mathbb{R}^p\setminus\{0\}} \frac{|\beta' G\rho_0|}{\sqrt{\beta' G\beta}} = \sqrt{\rho_0' G\rho_0} < \infty.$$

We conclude that direct, dual, and orthogonal representations are given by

$$\theta(\gamma) = M'\beta; \quad \theta(\alpha) = \rho'\mathrm{E}b(X)Y; \quad \theta(\gamma, \alpha) = M'\beta + \rho'\mathrm{E}b(X)Y - \rho' G\beta,$$

where β is γ's parameter and ρ is α's parameter. These representations appear to be both novel and useful.

2.4. *The case of infinite-dimensional regression*

In the infinite-dimensional case, we can employ the Riesz–Frechet representation theorem and Hahn–Banach extension theorem to establish existence of the linear RR.

LEMMA 2.1 (EXTENDED RIESZ REPRESENTATION). *(i) If $L < \infty$, there exists a unique minimal representer $\alpha_0^\star \in \bar{\Gamma}$ and $L = \|\alpha_0^\star\|_{P,2}$. (ii) If there exists a linear representer α_0 on Γ with $\|\alpha_0\|_{P,2} < \infty$, then $L = \|\alpha_0^\star\|_{P,2} \leq \|\alpha_0\|_{P,2} < \infty$, where $\alpha_0^\star$, obtained by projecting α_0 onto $\bar{\Gamma}$, is the unique minimal representer. In both cases $\gamma \mapsto \theta(\gamma)$ can be extended to $\bar{\Gamma}$ or to the entire $L^2(F)$ with the modulus of continuity L.*

The first part of the lemma shows (implicit) existence of a linear representer when $L < \infty$. Our estimation results will rely *only on the existence* of minimal representers. In some cases, however, we may utilize the closed-form solutions for linear representers (see, e.g., Section S3 for the key examples), to improve the basis functions for estimating the minimal representers. There is also an efficiency reason to work with minimal representers rather than any linear representer, as highlighted in Section 4 analysing semiparametric efficiency.

2.5. Informal preview of estimation and inference results

Our estimation and inference will exploit empirical analogs of both the orthogonal representation of the parameter (2.6) and the equation defining the RR property (2.4).

To approximate the regression function and the RR, we consider the p-vector of dictionary functions b, where the dimension p of the dictionary can be large, potentially much larger than n. We approximate $\alpha_0^\star$ by a linear form $b'\rho_0$, and we approximate $\gamma_0^\star$ by a linear form $b'\beta_0$, and estimate the parameters using the algorithms below.

(1) Let $(W_i)_{i=1}^n = (Y_i, X_i)_{i=1}^n$ denote i.i.d. copies of data vector W. We use cross-fitting to avoid biases from overfitting that can arise in high-dimensional settings. To this end, let $(I_1, \ldots, I_K)$ be a partition of the observation index set $\{1, \ldots, n\}$ into K distinct subsets of about equal size. Let $\mathbb{E}_A f = \mathbb{E}_A f(W)$ denote the empirical average of $f(W)$ over $i \in A \subset \{1, \ldots, n\}$: $\mathbb{E}_A f := \mathbb{E}_A f(W) = |A|^{-1} \sum_{i \in A} f(W_i)$.

(2) For each block $k = 1, \ldots, K$, we obtain generalized Dantzig selector (GDS) estimates $\hat{\alpha}_k = b'\hat{\rho}_k$ and $\hat{\gamma}_k = b'\hat{\beta}_k$, where

$$\begin{aligned} \hat{\rho}_k &= \arg\min_{\rho \in \mathbb{R}^p} \|\rho\|_1 : \|\hat{D}^{-1} \left\{ \mathbb{E}_{I_k^c} m(W, b) - \mathbb{E}_{I_k^c} b(X) b(X)' \rho \right\} \|_\infty \leq \lambda_\rho, \\ \hat{\beta}_k &= \arg\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 : \|\hat{D}^{-1} \left\{ \mathbb{E}_{I_k^c} (Y - b(X)'\beta) b(X) \right\} \|_\infty \leq \lambda_\beta, \end{aligned} \tag{2.10}$$

where $I_k^c = \{1, \ldots, n\} \setminus I_k$ is the set of observation indices leaving I_k out, λ's are tuning parameters, and $\hat{D}$ is a scaling detailed in Section S5. Typically λ's scale like $\sqrt{\log(p \vee n)/n}$; Section 3 provides concrete choices.

(3) The DML estimator is an average of estimated orthogonal representations over k:

$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \{ m(W_i, \hat{\gamma}_k) + \hat{\alpha}_k(X_i)[Y_i - \hat{\gamma}_k(X_i)] \}. \tag{2.11}$$

The estimator of its asymptotic variance is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \{ m(W_i, \hat{\gamma}_k) + \hat{\alpha}_k(X_i)[Y_i - \hat{\gamma}_k(X_i)] - \hat{\theta} \}^2. \tag{2.12}$$

We remark that the RR estimator in step 2 (2.10) is of Dantzig selector type, but is not exactly the Dantzig selector, requiring some new analysis. We use the GDS rather than series or spline estimation to accommodate high dimensional specifications for the regression and RR.

The dictionary $b(x)$ is very important for the GDS estimator. This dictionary should be chosen so that linear combinations of $b(x)$ can approximate in mean square any element of Γ. For example if Γ is the set of linear combinations of an infinite sequence of regressors, as for a high dimensional regression, then $b(x)$ could be chosen as the first p elements of that sequence. Also p can be chosen flexibly, because p will be allowed to grow faster than the sample size, as specified in the asymptotic theory to follow. In practice multiple choices of p could be tried.

Next, we state the key concentration and approximate Gaussianity results informally. Key quantities in the analysis are the 'true' score and its moments:

$$\psi_0^\star(W) := \theta_0^\star - m(W, \gamma_0^\star) - \alpha_0^\star(X)(Y - \gamma_0^\star(X)), \quad \sigma^2 := \mathrm{E}\psi_0^2(W), \quad \kappa^3 := \mathrm{E}|\psi_0^3(W)|.$$

We establish conditions under which

$$\|\hat{\gamma}_k - \gamma_0^\star\|_{P,2} + \|\hat{\alpha}_k - \alpha_0^\star\|_{P,2}/\sigma \to 0, \quad \sqrt{n}\int (\hat{\gamma}_k - \gamma_0^\star)(\hat{\alpha}_k - \alpha_0^\star) dF/\sigma \to 0. \quad (2.13)$$

These include a bound on the ℓ_1 norm of coefficients and that either the regression function or the RR is approximately sparse with the effective dimension s less than $\sqrt{n}$.

Given that (2.13) holds, we establish that the resulting debiased (or 'double') machine learning (DML) estimator $\hat{\theta}$ in (2.11) and (2.12) is *adaptive*, namely it is approximated up to the error $o(\sigma/\sqrt{n})$ by the oracle estimator

$$\bar{\theta} := \theta_0^\star - n^{-1}\sum_{i=1}^n \psi_0(W_i),$$

where the *oracle* knows the scores ψ_0. Hence the approximate deviation of $\hat{\theta}$ from $\theta_0^\star$ is determined by $\|\psi_0\|_{P,2}/\sqrt{n}$, which is the standard deviation of the oracle estimator.

Consequently, $\hat{\theta}$ concentrates in a $\sigma/\sqrt{n}$ neighbourhood of the target with deviations controlled by the normal laws,

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}(\sqrt{n}(\hat{\theta} - \theta_0^\star)/\sigma \le t) - \mathrm{P}(N(0,1) \le t) \right| \le A(\kappa/\sigma)^3/\sqrt{n} + \mathrm{error}_n \to 0,$$

where the error_n bound is nonasymptotic and tends to zero as $n \to \infty$. Of course, $\sigma/\sqrt{n} \to 0$ is required for concentration. The nonasymptotic bound automatically implies the uniform validity of results over large classes of probability laws P for W.

There are two cases to consider:

(1) REGULAR CASE: *the parameters σ, κ/σ, and L are bounded, leading to $1/\sqrt{n}$ concentration, adaptation, and Gaussian approximation.*
(2) NONREGULAR CASE, *the parameters σ, κ/σ, and L diverge, so that we need*

$$\sigma/\sqrt{n} \to 0, L/\sqrt{n} \to 0, (\kappa/\sigma)/\sqrt{n} \to 0,$$

for $\sigma/\sqrt{n}$ concentration, adaptation, and Gaussian approximation.

As we show in Section S3, in the case of local functionals, the latter condition can be more succinctly stated as

$$(\kappa/\sigma) \lesssim \sigma \asymp L, \quad L/\sqrt{n} \to 0.$$

Finally, we establish that we can *transfer* learning and inference guarantees for local functionals to those for the *(perfectly) localized* functionals if the localization bias is sufficiently small, namely

$$\sqrt{n}(\theta(\gamma_0^\star; \ell_h) - \theta(\gamma_0^\star; \ell_0)/\sigma \to 0.$$

We think it is remarkable that a single inference theory covers both regular and nonregular cases, and provides uniform validity over large classes of P.

Table 1. Average treatment effect of 401(k) eligibility on net financial assets in US dollars. Localized average treatment effects are reported by income quintile groups. The regression is estimated by GDS or Lasso. Standard errors are reported in parentheses.

Income quintile	N treated	N untreated	GDS		Lasso	
All	3682	6187	7607.95	(1394.92)	7733.31	(1416.46)
1	272	1702	4500.33	(924.12)	4477.43	(920.31)
2	527	1447	1051.60	(1501.03)	1119.06	(1500.78)
3	755	1219	5204.93	(1199.87)	4919.65	(1200.10)
4	962	1012	9515.58	(2141.92)	8837.39	(2150.58)
5	1166	807	19354.00	(7934.70)	14138.37	(8310.59)

3. APPLICATIONS

3.1. Global and local effects of 401(k) eligibility on net financial assets

First, we use our method to answer a question in household finance: what is the average treatment effect of 401(k) eligibility on net financial assets (over a horizon of about two years)? 401(k) is a retirement savings and investing plan that gives employees a tax break on money they contribute. We follow the identification strategy of Poterba et al. (1995), who assume selection on observables. The authors assume that when 401(k) was introduced, workers ignored whether a given job offered 401(k) and instead made employment decisions based on income and other observable job characteristics; after conditioning on income and job characteristics, 401(k) eligibility was exogenous at the time. This empirical question corresponds to Example 2.1.

We use data from the 1991 US Survey of Income and Program Participation (Chernozhukov et al., 2018b), using sample selection and variable construction as in Abadie (2003) and Chernozhukov and Hansen (2004). The outcome Y is net financial assets defined as the sum of individual retirement account (IRA) balances, 401(k) balances, checking accounts, US saving bonds, other interest-earning accounts, stocks, mutual funds, and other interest-earning assets minus nonmortgage debt. The treatment D is an indicator of eligibility to enroll in a 401(k) plan. The raw covariates X are age, income, years of education, family size, marital status, two-earner status, benefit pension status, IRA participation, and home-ownership. We impose common support of the propensity score for the treated and untreated groups based on these covariates, yielding $n = 9869$ observations. We consider the fully-interacted specification $b(D, X)$ of Chernozhukov et al. (2018a) with $p = 277$ including polynomials of continuous covariates, interactions among all covariates, and interactions between covariates and treatment status.

Tables 1 and 2 summarize results for the entire population and for each quintile of the income distribution. We use $K = 5$ folds in cross-fitting. To estimate the RR, we use the GDS procedure introduced in the present work. To estimate the regression, we use GDS, Lasso, random forest, or neural network. GDS is implemented using the tuning procedure described in Section S5. Lasso is implemented using the tuning procedure described in Chernozhukov et al. (2018). Random forest and neural network are implemented with the same settings as Chernozhukov et al. (2018a), i.e., with 1000 trees or a single hidden layer of eight neurons, respectively. We find average treatment effect (ATE) of 7608 (1395) using GDS for both the RR and the regression. This ATE estimate is stable across different choices of regression estimator. We find that localized ATE is not statistically significant for the second quintile, and it is statistically significant, positive, and

Table 2. Average treatment effect of 401(k) eligibility on net financial assets in US dollars. Localized average treatment effects are reported by income quintile groups. The regression is estimated by random forest or neural network. Standard errors are reported in parentheses.

Income quintile	N treated	N untreated	Random forest		Neural network	
All	3682	6187	8638.15	(1621.78)	7364.66	(1844.39)
1	272	1702	4874.49	(937.86)	4664.61	(1309.59)
2	527	1447	1957.72	(1738.61)	1635.69	(1603.19)
3	755	1219	3973.11	(1474.72)	5106.03	(1287.70)
4	962	1012	10056.79	(2375.44)	9529.03	(2205.61)
5	1166	807	21168.13	(8015.79)	20138.57	(7506.92)

strongly heterogeneous for the other quintiles. Interpreting the relatively high effect of 401(k) eligibility for the first quintile is a question for future research.

For comparison, Chernozhukov et al. (2018a) report ATE of 7170 (1398) by DML, which estimates the RR by estimating the propensity score and plugging it into the RR functional form. Though these two estimators are asymptotically equivalent under correct specification, our estimator avoids the estimated propensity score in the denominator which could cause numerical instability. The ATE results are broadly consistent with Poterba et al. (1995), who use a simpler specification motivated by economic reasoning. The localized ATE estimates by income quintile group appear to be new empirical results and are of interest in their own right. In Section S5 we report analogous estimates without debiasing. Without debiasing, the GDS and Lasso estimates of ATE are attenuated due to regularization. The bias is smaller for the random forest and neural network estimates of ATE.

3.2. Global and local price elasticity of petrol demand

Second, we use our method to estimate the average price elasticity of household petrol demand: the percentage change in demand due to a unit percentage change in price. This parameter is critical for assessing the welfare consequences of tax changes, and it has been studied in Hausman and Newey (1995), Schmalensee and Stoker (1999), Yatchew and No (2001), and Blundell et al. (2012). Formally, the parameter of interest is the average derivative of log demand with respect to log price holding income and demographic characteristics fixed. The exact version of this empirical question corresponds to Example 2.4. The approximate version of this empirical question corresponds to Example 2.3.

We use data from the 1994–1996 Canadian National Private Vehicle Use Survey (Semenova and Chernozhukov, 2021b), using sample selection and variable construction as in Yatchew and No (2001) and Belloni et al. (2019). The outcome Y is log petrol consumption. The variable D with respect to which we differentiate is log price per litre. The raw covariates X are log age, log income, and log distance as well as geographical, time, and household composition dummies. In total we have $n = 5001$ observations. We consider the specification $b(D, X)$ previously considered by Semenova and Chernozhukov (2021a) augmented with additional interactions. The Semenova and Chernozhukov (2021a) specification includes polynomials of continuous covariates, and interactions of log price (and its square) with time and household composition dummies. We further include interactions of log price (and its square) with log age, log age squared, log income, and log income squared to allow for heterogeneity. Altogether, $p = 99$.

Table 3. Estimated average derivative (price elasticity) of petrol demand. Localized average derivatives are reported by income quintile groups. The regression is estimated by GDS, Lasso, random forest, or neural network. Standard errors are reported in parentheses.

Income quintile	N	GDS		Lasso		Random forest		Neural network	
All	5001	−0.28	(0.06)	−0.16	(0.05)	−0.01	(0.06)	0.15	(0.05)
1	1001	−0.84	(0.13)	−0.44	(0.12)	−0.37	(0.14)	0.06	(0.12)
2	1000	−0.36	(0.12)	−0.27	(0.11)	−0.13	(0.13)	0.42	(0.12)
3	1000	−1.40	(0.15)	−0.91	(0.13)	−0.60	(0.13)	−0.28	(0.13)
4	1000	−1.06	(0.14)	−0.79	(0.14)	−0.32	(0.15)	0.13	(0.14)
5	1000	−0.11	(0.14)	−0.03	(0.11)	0.16	(0.12)	0.58	(0.10)

Table 3 summarizes results for the entire population and for each quintile of the income distribution. We use $K = 5$ folds in cross-fitting. To estimate the RR, we use the GDS procedure introduced in the present work. To estimate the regression, we use GDS, Lasso, random forest, or neural network. Again, GDS is implemented using the tuning procedure described in Section S5, Lasso is implemented using the tuning procedure described in Chernozhukov et al. (2018), and random forest and neural network are implemented with the same settings as Chernozhukov et al. (2018a). We find average price elasticity of −0.28 (0.06) using GDS for both the RR and the regression. Lasso gives similar results. Note that random forest is not differentiable, and the derivative of a neural network may be difficult to extract from a black-box implementation. When using these estimators, we implement a partial difference approximation of the derivative, detailed in Section S5. We conjecture that this approximation explains why the results using random forest appear attenuated and why the results using neural network appear positive or statistically insignificant. Using GDS, we find that localized average price elasticity is statistically significant and negative in each income quintile, with substantial heterogeneity.

For comparison, OLS regression of log consumption on log price, log age, log income, and log distance as well as geographical, time, and household composition dummies yields an estimate of 0.14 (0.06). The linear specification leads to a positive elasticity estimate, contradicting economic intuition (since it says there would be more petrol consumption when prices are higher). Our localized average price elasticity results using GDS are broadly consistent with Semenova and Chernozhukov (2021a), who more explicitly consider the relationship between average price elasticity and income. In Section S5 we report analogous estimates without debiasing. Without debiasing, the GDS and Lasso estimates of quintile elasticities are attenuated due to regularization. The bias is smaller for the random forest and neural network estimates of quintile elasticities.

4. ESTIMATION AND INFERENCE FOR HIGH DIMENSIONAL APPROXIMATELY LINEAR MODELS

4.1. Best linear approximations for the regression function and the Riesz representer

To approximate the regression function, we consider the p-vector of dictionary functions

$$x \mapsto b(x) = (b_j(x))_{j=1}^p, \quad b_j \in L^2(F).$$

The dimension p of the dictionary can be large, potentially much larger than n. Let Γ_b be the linear subspace of $L^2(F)$ generated by b. We assume that as $n \to \infty$ we have that $p \to \infty$ and $\Gamma_b \to \bar{\Gamma} := \text{closure}(\Gamma)$, where $\bar{\Gamma}$ is a linear subspace of $L^2(F)$ with the basis functions $\{\tilde{b}_j\}_{j=1}^{\infty}$. Here convergence means that any convergent sequence in Γ_b has its limit in $\bar{\Gamma}$ and for each $\gamma \in \bar{\Gamma}$ we have a sequence in Γ_b converging to it, with respect to the $L^2(F)$ norm. Note that this setup allows the dictionary $b = b_n$ to change with n, as for example with b-splines.

Here we define $\gamma_0^\star$ as a projection of Y onto $\bar{\Gamma}$, i.e., $\gamma_0^\star$ is the projection of Y on the infinite set of variables $\{\tilde{b}_j(X)\}_{j=1}^{\infty}$. This setup is slightly more general than in the introduction, where $\gamma_0^\star$ was the conditional expectation function. Of course, if the latter is an element of $\bar{\Gamma}$, it automatically coincides with $\gamma_0^\star$.

We approximate $\gamma_0^\star$ by the finite-dimensional best linear predictor (BLP) γ_0 via

$$\gamma_0^\star = \gamma_0 + r_\gamma := b'\beta_0 + r_\gamma : \mathrm{E}[b(X)r_\gamma(X)] = 0,$$

where r_γ is the approximation error, and $\gamma_0 := b'\beta_0$ is the BLP of Y and best linear approximation to $\gamma_0^\star$. We define β_0 as a minimal ℓ_1-norm solution to the system of equations

$$\min \|\beta\|_1 : \mathrm{E}[b(X)(\gamma_0^\star(X) - b(X)'\beta)] = 0,$$

when $G = \mathrm{E}b(X)b(X)'$ is not full rank.

Similarly, we approximate the RR $\alpha_0^\star$, which exists by Lemma 2.1 whenever $L < \infty$, via the best linear approximation α_0:

$$\alpha_0^\star = \alpha_0 + r_\alpha = b'\rho_0 + r_\alpha : \mathrm{E}[r_\alpha(X)b(X)] = 0.$$

We define ρ_0 as a minimal ℓ_1-norm solution to the system of equations

$$\min \|\rho\|_1 : \mathrm{E}[(\alpha_0^\star(X) - b(X)'\rho)b(X)] = 0.$$

Using that $\mathrm{E}\alpha_0^\star(X)b(X) = \mathrm{E}m(W, b)$, we note that

$$0 = \mathrm{E}[r_\alpha(X)b(X)] = \mathrm{E}((\alpha_0^\star(X) - b(X)'\rho_0)b(X)) = \mathrm{E}m(W, b) - \mathrm{E}\alpha_0(X)b(X). \tag{4.1}$$

Hence α_0 is the RR for $\mathrm{E}m(W, \gamma)$ for each $\gamma \in \Gamma_b$. Here we can interpret Γ_b as the collection of test functions on which the representation property (4.1) holds.

DEFINITION 4.1 (PENULTIMATE AND ULTIMATE TARGET PARAMETERS). *Our penultimate target is the linear functional applied to the BLP γ_0:*

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

Our ultimate target is the linear functional applied to $\gamma_0^\star$

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

If the approximation errors are such that

$$(\sqrt{n}/\sigma) \int r_\alpha r_\gamma dF \to 0, \tag{4.2}$$

our inference will target the ultimate parameter. In the nonregular setup, the second order error condition $\int r_\alpha r_\gamma dF \leq \sigma/\sqrt{n}$ in (4.2) is weaker than what is usually required for pathwise differentiable functionals (since $\sigma \to \infty$ is the nonregular case); there is a lower bar for oracle rates in nonregular problems. This phenomenon was also noted by Foster and Syrgkanis (2019)

and Kennedy (2020). Otherwise our inference will target an interpretable penultimate parameter. We shall formally refer to the latter case as the *misspecified case*.

LEMMA 4.1 (BASIC PROPERTIES OF THE SCORE). *Our DML estimator of θ_0 will be based on the following score function:*

$$\psi(W, \theta; \beta, \rho) = \theta - m(W, b)'\beta - \rho' b(X)(Y - b(X)'\beta),$$

which has the following properties:

$$\partial_\beta \psi(W, \theta; \beta, \rho) = -m(W, b) + \rho' b(X) b(X)', \quad \partial_\rho \psi(W, \theta; \beta, \rho) = -b(X)(Y - b(X)'\beta),$$

$$\partial^2_{\beta\beta'} \psi(W, \theta; \beta, \rho) = \partial^2_{\rho\rho'} \psi(W, \theta; \beta, \rho) = 0, \quad \partial^2_{\beta\rho'} \psi(W, \theta; \beta, \rho) = b(X) b(X)'.$$

This score function is Neyman orthogonal at (β_0, ρ_0):

$$\mathrm{E}[\partial_\beta \psi(W, \theta; \beta, \rho_0)] = -\mathrm{E}[m(W, b)] + G\rho_0 = 0,$$

$$\mathrm{E}[\partial_\rho \psi(W, \theta; \beta_0, \rho)] = \mathrm{E}[-b(X)(Y - b(X)'\beta_0)] = -\mathrm{E}[b(X)\gamma_0(X)] + G\beta_0 = 0.$$

The second claim of the lemma is immediate from the definition of (β_0, ρ_0) and the first follows from elementary calculations. The orthogonality property above says that the score function is invariant to small perturbations of the nuisance parameters ρ and β around their 'true values' ρ_0 and β_0. This invariance property plays a crucial role in removing the impact of biased estimation of nuisance parameters ρ_0 and β_0 on the estimation of the main parameters θ_0.

4.2. Estimators

Estimation will be carried out using the following Dantzig selector-type estimators (Candes and Tao, 2007). In a follow-up work, Chernozhukov et al. (2018) consider Lasso-type estimators.

DEFINITION 4.2 (GENERALIZED DANTZIG SELECTOR ESTIMATOR). *Consider a parameter $t \in T \subset \mathbb{R}^p$, where T is a convex set. Consider the moment functions $t \mapsto g(t)$ and the estimated moment functions $t \mapsto \hat{g}(t)$, mapping $\mathbb{R}^p$ to $\mathbb{R}^p$:*

$$g(t) = Gt - M; \quad \hat{g}(t) = \hat{G}t - \hat{M},$$

where G and $\hat{G}$ are p by p nonnegative-definite matrices and M and $\hat{M}$ are p-vectors. Define t_0 as a minimal ℓ_1-norm solution to $g(t) = 0$ and assume $t_0 \in T$. Define the GDS estimator $\hat{t}$ by solving

$$\hat{t} \in \arg\min \|t\|_1 : \|\hat{g}(t)\|_\infty \leq \lambda, \quad t \in T,$$

where λ is chosen such that $\|\hat{g}(t_0) - g(t_0)\|_\infty \leq \lambda$, with probability at least $1 - \epsilon$.

Here we record the possibility of convex restrictions on the parameter space by placing t in a convex parameter space T. If parameter restrictions are correct, then this can potentially improve theoretical guarantees by weakening the requirements on G and other primitives.

DEFINITION 4.3 (GDS FOR BLP: DANTZIG SELECTOR). *Given a diagonal positive-definite normalization matrix D_β, define $\hat{\beta}_A = D_\beta \hat{t}$, where $\hat{t}$ is the GDS estimator for $t_0 = D_\beta^{-1}\beta_0$ with*

$$G = \mathrm{E} b(X) b(X)', \hat{G} = \mathbb{E}_A b(X) b(X)', M = D_\beta^{-1} \mathrm{E} Y b(X), \hat{M} = D_\beta^{-1} \mathbb{E}_A Y b(X); T_\beta \subset \mathbb{R}^p.$$

In this setting, our estimator specializes to the original Dantzig selector. In practice, we use $T_\beta = \mathbb{R}^p$, although when we are interested in average derivative functionals, it is theoretically helpful to impose the convex restrictions of the sort $T = \{t \in \mathbb{R}^p : \sup_{x\in\mathcal{X}} |\partial_d b(x)'t| \leq B\}$, where B is some a priori known upper bound on the derivative. Ideally, D_β is chosen such that $\text{diag}(Var(D_\beta^{-1}(\hat{G}\beta_0 - \hat{M})) = I$. Our practical algorithm given in Section S5 estimates D_β from the data.

DEFINITION 4.4 (GDS FOR RIESZ REPRESENTER). *Given a diagonal positive-definite normalization matrix D_ρ, define $\hat{\rho}_A = D_\rho\hat{t}$, where $\hat{t}$ is the GDS estimator of the parameter $t_0 = D_\rho^{-1}\rho_0$ with*

$$G = \text{E}b(X)b(X)', \hat{G} = \mathbb{E}_A b(X)b(X)', M = D_\rho^{-1}\text{E}m(W, b), \hat{M} = D_\rho^{-1}\mathbb{E}_A m(W, b); T_\rho \subset \mathbb{R}^p.$$

In this setting, our estimator is a generalization of the original Dantzig selector. In practice, we are using $T_\rho = \mathbb{R}^p$, even though it is possible to exploit some structured restrictions on the problem motivated by the nature of the universal RRs. Ideally, D_ρ is chosen such that $\text{diag}(Var(D_\rho^{-1}(\hat{G}\rho_0 - \hat{M})) = I$. Our practical algorithm given in Section S5 estimates D_ρ from the data.

We now define the DML estimator with RRs, which makes use of cross-fitting.

DEFINITION 4.5 (DML WITH RR). *Consider the partition of $\{1, \ldots, n\}$ into $K \geq 2$ blocks $(I_k)_{k=1}^K$, with $m = \lfloor n/K \rfloor$ observations in I_k, for $k < K$ and $\lceil n/K \rceil$ remaining in I_K. For each $k = 1, \ldots, K$, let $\hat{\beta}_k$ and $\hat{\rho}_k$ denote GDS estimators obtained using data $(W_i)_{i\in I_k^c}$, where $I_k^c = \{1, \ldots, n\} \setminus I_k$, and let estimator $\hat{\theta}_k$ be defined as*

$$\hat{\theta}_k = \mathbb{E}_{I_k}[m(W, b)'\hat{\beta}_k + \hat{\rho}_k' b(X)(Y - b(X)'\hat{\beta}_k)].$$

Define the DML estimator $\hat{\theta}$ as the average:

$$\hat{\theta} = \sum_{k=1}^K \hat{\theta}_k w_k; \quad w_k = \frac{\lfloor n/K \rfloor}{n} \text{ if } k < K, \quad w_K = \frac{\lceil n/K \rceil}{n}.$$

4.3. Properties of DML: Main result

We provide a single nonasymptotic result that allows us to cover both global and local functionals, implying uniformly valid rates of concentration and normal approximations over large sets of P.

Consider the oracle estimator based upon the true score functions:

$$\bar{\theta} := \theta_0 - n^{-1}\sum_{i=1}^n \psi_0(W_i), \quad \psi_0(W) := \psi(W, \theta_0; \beta_0, \rho_0).$$

We seek to establish minimal conditions under which the DML estimator approximates the oracle estimator, and is approximately normal with distribution

$$N(0, \sigma^2/n), \quad \sigma := \|\psi_0\|_{P,2}.$$

For regular functionals σ is bounded, giving $1/\sqrt{n}$ concentration around θ_0, and for nonregular functionals $\sigma \propto L \to \infty$ requring $L/\sqrt{n} \to 0$ to get concentration. Our normal approximation is accurate if kurtosis of ψ_0 does not grow too fast:

$$(\kappa/\sigma)^3/\sqrt{n} \text{ is small}, \quad \kappa := \|\psi_0\|_{P,3}.$$

In the regular case $(\kappa/\sigma)^3$ is bounded, but for the nonregular cases it can scale as fast as L, again requiring $L/\sqrt{n} \to 0$.

Fix all of these sequences and the constants. Define the guarantee set:

$$\mathsf{S} = \left\{ (u, v) \in \mathbb{R}^{2p} : \sqrt{u'Gu} \leq r_1, \sqrt{v'Gv} \leq \sigma r_2, \quad |u'Gv| \leq \sigma r_3, \beta_0 + u \in T_\beta, \rho_0 + v \in T_\rho \right\},$$

We will take $u = \hat{\beta}_k - \beta_0$ and $v = \hat{\rho}_k - \rho_0$. As such, r_1 measures the nonasymptotic mean square rate for the BLP; r_2 measures the nonasymptotic mean square rate for the RR; and r_3 measures how the estimation errors interact. Note the presence of σ acting on r_2 and r_3, which accommodates nonregular functionals. We will instantiate (r_1, r_2, r_3) as fast and slow rates by analysing the GDS estimator, in Theorem 4.3 below.

Next, define μ to be the smallest modulus of continuity such that on $(u, v) \in \mathsf{S}$

$$\sqrt{Var}((-m(W, b) + \rho_0{}'b(X)b(X))'u) \leq \mu\sigma\|b'u\|_{P,2}, \sqrt{Var}((Y - b(X)'\beta_0)b(X)'v) \leq \mu\|b'v\|_{P,2},$$

$$\sqrt{Var}(u'b(X)b(X)'v) \leq \mu(\|b'u\|_{P,2} + \|b'v\|_{P,2}).$$

In typical applications, the modulus of continuity μ is bounded. Indeed, if elements of the dictionary are bounded with probability one, $\|b(X)\|_\infty \leq C$, then we can select $\mu = CB$ for many functionals of interest, so the assumption is plausible. If $b(X) = X$ are sub-Gaussian, then this assumption is also easily satisfied; however, this case is not of central interest to us. See Chernozhukov et al. (2021) for a more general discussion.

Consider P that satisfies the following conditions.

R(δ) *With probability* $1 - \varepsilon$, *the estimation errors* $\{(\hat{\beta}_k - \beta_0, \hat{\rho}_k - \rho_0)\}_{k=1}^K$ *take values in* S^K, *with quality of the guarantee obeying*

$$\sigma^{-1}(\sqrt{m}\sigma r_3 + \mu r_1(1 + \sigma) + \mu\sigma r_2) \leq \delta.$$

$R(\delta)$ is a requirement on how the sequences (r_1, r_2, r_3) evolve relative to (σ, μ, m). We will formally verify $R(\delta)$ for the approximately sparse setting, in Corollary 4.4 below. $R(\delta)$ is the key condition for our main result, Theorem 4.1.

THEOREM 4.1 (ADAPTIVE ESTIMATION AND APPROXIMATE GAUSSIAN INFERENCE). *Suppose K divides n for simplicity. Under condition $R(\delta)$, we have the adaptivity property, namely the difference between the DML and the oracle estimator is small: for any $\Delta \in (0, 1)$,*

$$|\sqrt{n}(\hat{\theta} - \bar{\theta})/\sigma| \leq \sqrt{K}4\delta/\Delta,$$

with probability at least $1 - \varepsilon - \Delta^2$.

As a consequence, $\hat{\theta}$ concentrates in a $\sigma/\sqrt{n}$ neighborhood of θ_0, with deviations approximately distributed according to the Gaussian law $\Phi(z) = \mathrm{P}(N(0, 1) \leq z)$:

$$\sup_{z\in\mathbb{R}} \left|\mathrm{P}(\sigma^{-1}\sqrt{n}(\hat{\theta}_0 - \theta_0) \leq z) - \Phi(z)\right| \leq A(\kappa/\sigma)^3 n^{-1/2} + \sqrt{K}2\delta/\Delta + \varepsilon + \Delta^2,$$

where $A < 1/2$ is the sharpest absolute constant in the Berry–Esseen bound.

The conclusions of this result are distinguished from those of Chernozhukov et al. (2018a) and Chernozhukov et al. (2018) in applying to local, nonparametric objects, in providing finite sample bounds, and in being uniform over the parameter space. The conclusions are similar to

this previous work in relying on a rate condition that is the product of rates of estimation for two distinct functions, here the regression and the RR.

The constants can be chosen to yield an asymptotic result.

COROLLARY 4.1 (UNIFORM ASYMPTOTIC ADAPTIVITY AND GAUSSIANITY). *Let $\mathcal{P}_n$ be any nondecreasing set of probability laws P that obey condition $R(\delta_n)$ where $\delta_n \to 0$ is a given sequence. Then the DML estimator $\hat{\theta}$ is uniformly asymptotically equivalent to the oracle estimator $\bar{\theta}$, that is*

$$|\sqrt{n}(\hat{\theta} - \bar{\theta})/\sigma| = O_P(\delta_n),$$

uniformly in $P \in \mathcal{P}_n$ as $n \to \infty$. In addition, if for each $P \in \mathcal{P}_n$ the kurtosis of ψ_0 does not grow too fast, namely:

$$(\kappa/\sigma)^3/\sqrt{n} \leq \delta_n,$$

we have that $\sqrt{n}(\hat{\theta} - \theta_0)/\sigma$ is asymptotically Gaussian uniformly in $P \in \mathcal{P}_n$:

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_n} \sup_{z\in\mathbb{R}} \left| \mathrm{P}_P(\sqrt{n}(\hat{\theta}_0 - \theta_0)/\sigma \leq z) - \Phi(z) \right| = 0.$$

Hence the DML estimator of the linear functionals of the BLP function γ_0 enjoys good properties under the stated regularity conditions. This result does not distinguish between inference on global functionals from inference on local functionals, as long as the latter are not perfectly localized. We state a separate result for perfectly localized functionals below.

COROLLARY 4.2 (INFERENCE ON THE ULTIMATE PARAMETER $\theta_0^\star$). *Suppose that, in addition to conditions of Corollary 4.1, P satisfies the small approximation error condition:*

$$(\sqrt{n}/\sigma)|\theta_0 - \theta_0^\star| = (\sqrt{n}/\sigma) \left| \int r_\alpha r_\gamma dF \right| \leq \delta. \tag{4.3}$$

Then conclusions of Theorem 4.1 hold with $\theta_0^\star$ replacing θ_0, with $\sqrt{K}4\delta/\Delta$ increased by δ, and the same probability. Conclusions of Corollary 4.1 continue to hold with $\theta_0^\star$ replacing θ_0 for a class of probability laws $\mathcal{P}_n$, provided each $P \in \mathcal{P}_n$ satisfies the conditions of Corollary 4.1 and (4.3) for the given $\delta = \delta_n \to 0$.

The approximation bias for the ultimate target can be plausibly small due to the fact that many rich function classes admit regularized linear approximations with respect to conventional dictionaries b. For instance, Tsybakov (2012) and Belloni et al. (2014) show small approximation bias using Fourier bases as dictionaries, and using Sobolev and rearranged Sobolev balls, respectively, as the function classes.

COROLLARY 4.3 (INFERENCE ON THE PERFECTLY LOCALIZED PARAMETER). *Suppose that, in addition to conditions of Corollary 4.1, P satisfies the small approximation error condition:*

$$\sqrt{n}|\theta_0(\gamma_0; \ell_h) - \theta_0(\gamma_0^\star; \ell_h)|/\sigma = \sqrt{n} \left| \int r_\alpha r_\gamma dF \right| /\sigma \leq \delta, \tag{4.4}$$

and the localization bias is small:

$$\sqrt{n}|\theta_0(\gamma_0^\star; \ell_h) - \theta_0(\gamma_0^\star; \ell_0)|/\sigma \leq \delta. \tag{4.5}$$

Then conclusions of Theorem 4.1 hold with $\theta_0(\gamma_0^\star; \ell_0)$ replacing θ_0, with $\sqrt{K}4\delta/\Delta$ increased by 2δ, and the same probability. Conclusions of Corollary 4.1 continue to hold with $\theta_0^\star(\gamma^\star; \ell_0)$

replacing $\theta_0 = \theta_0(\gamma_0; \ell_h)$ for a class of probability laws $\mathcal{P}_n$, provided each $P \in \mathcal{P}_n$ satisfies the conditions of Corollary 4.1 and (4.4)–(4.5) for the given $\delta = \delta_n \to 0$.

4.4. Semiparametric efficiency

Below we use concepts from semiparametric efficiency, as presented in Bickel et al. (1993) and Van der Vaart (2000); we do not recall them here for brevity.

The DML estimator $\hat{\theta}$ will be asymptotically efficient for estimating $\theta_0^\star$, defined as a functional of $\gamma_0^\star$, the projection of Y on $\bar{\Gamma}$. The distribution of a data observation is unrestricted in this case, so that there will only be one influence function for each functional of interest, and the estimator is asymptotically linear with that influence function. The standard semiparametric efficiency results then imply that our estimator will have the smallest asymptotic concentration among estimators that are locally regular; see Bickel et al. (1993) and Van der Vaart (2000).

Our formal result stated below only implies efficiency for the regular case, where the operator norm of the function L is bounded, holding P fixed. We expect that a similar result continues to hold with $L \to \infty$, by developing an appropriate formalization that handles P changing with n and rules out super-efficiency phenomena. However, this formalization requires a separate major development, which we leave to future research. In what follows, the notation $\gamma_{0,P}^\star$ emphasizes the dependence of the projection $\gamma_0^\star$ on P.

THEOREM 4.2 (EFFICIENCY). *Let $\psi_0^\star(W) := \theta_0^\star - m(W, \gamma_0^\star) - \alpha_0^\star(X)(Y - \gamma_0^\star(X))$. Suppose that $\mathrm{E}[Y^2] < \infty$, $\mathrm{E}[\psi_0^\star(W)^2] < \infty$, and $m(W, \gamma)$ is mean square continuous in γ under P. Then $\theta_{0,P} := \int m(w, \gamma_{0,P}^\star) dP(w)$ is differentiable at P, in the sense that*

$$\lim_{\tau \searrow 0} \frac{\theta_{0,P_\tau} - \theta_{0,P}}{\tau} = \mathrm{E}_P \delta(W) \psi_0^\star(W),$$

where $\psi_0^\star$ is called the influence function and is unique, and the directional perturbation P_τ is defined as $dP_\tau = dP[1 + \tau\delta]$, where the direction δ is any element of the tangent set $\{\delta$ measurable $: \mathcal{W} \to \mathbb{R} : \int \delta dP = 0, \|\delta\|_\infty < M\}$ for each $0 < M < \infty$. Consequently, the asymptotic variance of every regular sequence of estimators is bounded below by $\|\psi_0^\star\|_{P,2}$. Further, since the tangent set is a convex cone, other conclusions of theorems 25.20 and 25.21 of Van der Vaart (2000) also hold, namely the convolution and the minimax characterization of the efficiency.

4.5. Properties of GDS estimators

Our goal is to verify that the guarantee $R(\delta)$ holds. In particular we have to analyse (r_1, r_2, r_3) by bounding the population prediction norm $v \mapsto \sqrt{v'Gv}$. This is a more nuanced problem than bounding the empirical prediction norm $v \mapsto \sqrt{v'\hat{G}v}$, which has been accomplished in a variety of prior analyses done on Dantzig-type and Lasso-type estimators.

We begin with the following condition, which only controls the max of error rates and controls the ℓ_1 norm of true parameters:

MD *We have that $t_0 \in T$ and $\|t_0\|_1 \leq B$, where $B \geq 1$, and the empirical moments obey the following bounds with probability at least $1 - \varepsilon$, for $\bar{\lambda} \geq \lambda$*

$$\|\hat{G} - G\|_\infty \leq \bar{\lambda}, \ \|\hat{G}t_0 - \hat{M}\|_\infty \leq \lambda.$$

The bounds on ℓ_1 norm of coefficients are naturally motivated, for example, by working in Sobolev or rearranged Sobolev spaces (see, Tsybakov, 2012 and Belloni et al., 2014, respectively). Rearranged Sobolev spaces allow the first p regression coefficients in the series expansion to be arbitrarily rearranged, allowing a much greater degree of oscillatory behaviours than in the original Sobolev spaces. The complexity of these function classes are also different. Sobolev spaces are Donsker sets under sufficient smoothness, whereas rearranged Sobolev spaces have the covering entropy bounded below by log p and are not Donsker if $p \to \infty$.

At the core of this approach is the restricted set

$$S(t_0, \nu) := \{\delta : \|G\delta\|_\infty \leq \nu, \|t_0 + \delta\|_1 \leq \|t_0\|_1, t + \delta \in T\},$$

where ν is the noise level. As demonstrated in the proof of Lemma 4.3, the GDS estimator belongs to this set with high probability $1 - \epsilon$ for the noise level $\nu = 4B\bar{\lambda}$, where λ is the penalty level of GDS (ν scales like $\sqrt{\log(p \vee n)}/\sqrt{n}$ in our problems).

DEFINITION 4.6 (EFFECTIVE DIMENSION). *Define the effective dimension of t_0 at the noise level $\nu > 0$ as:*

$$s(t_0) := s(t_0; \nu) := \sup_{\delta \in S(t_0, \nu)} |\delta' G \delta| / \nu^2.$$

The effective dimension is defined in terms of the population (rather than sample) covariance matrix G, which makes it easy to verify regularity conditions. Note that if $G = I$ and $\|t_0\|_0 = s$, then $s(t_0) \leq s$. More generally, $s(t_0)$ measures the effective difficulty of estimating t_0 in the prediction norm, created by design G and the structure of t_0. The condition imposes no conditions on the restricted or sparse eigenvalues of G. For example, take $G = 11'$, a rank 1 matrix, and suppose $\|t_0\|_0 = 1$. Then $s(t_0) \leq 1$ holds in this case, giving useful and intuitive performance bounds, while the standard restricted eigenvalues and cone invertibility factors are all zero in this case, yielding no bounds on the performance in the population prediction norm. This type of example illustrates the possibility of accommodation of overcomplete (multiple or amalgamated) dictionaries in b, whose use in conjunction with ℓ_1- penalization has been advocated by Donoho et al. (2005). Of course, the bounds on effective dimension follow from the bounds on cone-invertibility factors and restricted eigenvalues.

Given a vector $\delta \in \mathbb{R}^p$, let δ_A denote a vector with the j-th component set to δ_j if $j \in A$ and 0 if $j \notin A$.

LEMMA 4.2 (BOUND ON EFFECTIVE DIMENSION IN APPROXIMATELY SPARSE MODEL). *Suppose that t_0 is approximately sparse, namely*

$$|t_0|_j^* \leq Aj^{-a} \quad j = 1, \ldots, p,$$

for some finite positive constants A and $a > 1$, where $(|t_0|_j^*)_{j=1}^p$ is the nonincreasing rearrangement of $(|t_{0j}|)_{j=1}^p$. Let $t_0^{\mathcal{M}} := t_0(1(|t_0| > \nu) := (t_{0j} 1(|t_{0j}| > \nu))_{j=1}^p$ denote the vector with components smaller than ν trimmed to 0. Then

$$s(t_0; \nu) \leq s \times \left(k^{-1} \vee \frac{6a}{a-1}\right), \quad \|t_0^{\mathcal{M}}\|_0 \leq s := (A/\nu)^{1/a},$$

k is the cone invertibility factor:

$$k := \inf \frac{|\mathcal{M}| \|G\delta\|_\infty}{\|\delta\|_1} : \delta \neq 0, \quad \|\delta_{\mathcal{M}^c}\|_1 \leq 2\|\delta_{\mathcal{M}}\|_1,$$

$\mathcal{M} = support(t_0^{\mathcal{M}})$, $\mathcal{M}^c = \{1, \ldots, p\} \setminus \mathcal{M}$, *and* $|\mathcal{M}| \leq s$.

The cone invertibility factor is a generalization of the restricted eigenvalue condition of Bickel et al. (2009), proposed by Ye and Zhang (2010).

Since approximate sparsity is a simple condition that implies a bound on effective dimension, we pause and interpret approximate sparsity in the context of a motivating example from causal inference. In particular, we revisit ATE (Example 2.1). For simplicity, consider the global parameter and assume that the function $\mathrm{E}[Y|D, Z]$ is an element of $\bar{\Gamma}$, so that $\gamma_0^\star(D, Z) = \mathrm{E}[Y|D, Z]$ and $\alpha_0^\star(D, Z) = D/\pi_0^\star(Z) - (1 - D)/(1 - \pi_0^\star(Z))$ where $\pi_0^\star(Z) = \mathrm{E}[D|Z]$ is the propensity score. Consider the dictionary $b(d, z) = (dq(z)', (1 - d)q(z)')'$ where $\{q_j(z)\}_{j=1}^{p/2}$ are the initial $p/2$ elements of a sequence of basis functions that approximates the functions $\mathrm{E}[Y|1, Z]$, $\mathrm{E}[Y|0, Z]$, $1/\pi_0^\star(Z)$, and $1/(1 - \pi_0^\star(Z))$.

Suppose the minimal ℓ_1-norm mean square projections of $\mathrm{E}[Y|1, Z]$ and $\mathrm{E}[Y|0, Z]$ onto $\{q_j(z)\}_{j=1}^{p/2}$ are approximately sparse after rescaling appropriately by D_β^{-1}. (Note that if $\mathrm{E}[Y|1, Z]$ and $\mathrm{E}[Y|0, Z]$ are already approximately sparse then so are their projections.) It follows that the minimal ℓ_1-norm mean square projection of $\gamma_0^\star$ is approximately sparse and $s_\beta := s(D_\beta^{-1}\beta_0; \nu)$ is small.

Suppose instead that the minimal ℓ_1-norm mean square projections of $1/\pi_0^\star(Z)$ and $1/(1 - \pi_0^\star(Z))$ onto $\{q_j(z)\}_{j=1}^{p/2}$ are approximately sparse after rescaling appropriately by D_ρ^{-1}. (Note that if $1/\pi_0^\star(Z)$ and $1/(1 - \pi_0^\star(Z))$ are already approximately sparse then so are their projections.) It follows that the minimal ℓ_1-norm mean square projection of $\alpha_0^\star$ is approximately sparse and $s_\rho := s(D_\rho^{-1}\rho_0; \nu)$ is small.

The concept of the effective dimension does not split t_0 into a sparse component and a small dense component, as is done in the now standard analysis of ℓ_1-regularized estimators of approximately sparse t_0. The effective dimension is simply stated in terms of t_0 alone.

LEMMA 4.3 (NON-ASYMPTOTIC BOUND FOR GDS IN POPULATION PREDICTION NORM). *Suppose that MD holds. Then with probability* $1 - 2\varepsilon$ *the estimator* $\hat{t}$ *exists and obeys:*

$$(\hat{t} - t_0)'G(\hat{t} - t_0) \leq (s(t_0; \nu)\nu^2) \wedge (2B\nu).$$

The bound is a minimum of what is called the 'fast rate bound' and the 'slow rate' bound. This result tightens the result in Chatterjee (2013), who established a 'slow rate' bound (in the context of Lasso) that applies under no assumptions on G. If the effective dimension is not too big, as in the examples above, the 'fast rate' $s(t_0; \nu)\nu^2$ provides a tighter bound under weak assumptions on G. It is important to emphasize that the result is stated in terms of the population prediction norm rather than the empirical norm.

We now apply this result to GDS estimators of the RR and the BLP. We impose the following conditions. Let $\mathbb{G}_A$ denote the empirical process over $f \in \mathcal{F} : \mathcal{W} \to \mathbb{R}^p$ and $i \in A$, namely

$$\mathbb{G}_A f := \mathbb{G}_A f(W) := |I|^{-1/2} \sum_{i \in A} (f(W_i) - Pf), \quad Pf := Pf(W) := \int f(w) dP(w).$$

The following is a sufficient condition that will deliver the guarantee $R(\delta)$ for $\delta \to 0$. Let $\tilde{\ell}$ denote a positive constant (that increases to ∞ as $n \to \infty$ in the asymptotic results).

SC *(a) The* ℓ_1 *norms of coefficients are bounded as* $\|D_\rho^{-1}\rho_0\|_1 \leq B$ *and* $\|D_\beta^{-1}\beta_0\|_1 \leq B$, *for* $B \geq 1$, *and the scaling matrices obey* $\|D_\rho v\| \leq \mu_D \sigma \|v\|$ *for* $D_\rho^{-1} v \in S(D_\rho^{-1}\rho_0, \nu)$ *and* $\|D_\beta u\| \leq \mu_D \|u\|$ *for* $D_\beta^{-1} u \in S(D_\beta^{-1}\beta_0, \nu)$ *for* $\nu = 4B\tilde{\ell}/\sqrt{n}$. *(b) Given a random*

subset A of $\{1,\dots,n\}$ of size $m \geq n - \lfloor n/K \rfloor$, dictionary b obeys with probability at least $1-\epsilon$, $\|\mathbb{G}_A bb'\|_\infty \leq \tilde{\ell}$. (c) The penalty levels λ_ρ and λ_β are chosen such that with probability at least $1-\epsilon$, $\|D_\beta^{-1}(\mathbb{G}_A bb'\beta_0 - \mathbb{G}_A Yb(X))\|_\infty/\sqrt{m} \leq \lambda_\rho, \|D_\rho^{-1}(\mathbb{G}_A bb'\beta_0 - \mathbb{G}_A m(W,b))\|_\infty/\sqrt{m} \leq \lambda_\beta$, and are not overly large, $\lambda_\beta \vee \lambda_\rho \leq \tilde{\ell}/\sqrt{m}$.

SC(a) records a restriction on the ℓ_1 norm of β_0 and ρ_0. For instance, in Examples 2.1, 2.2, and 2.3, $D_\rho \asymp \sigma I \asymp LI$, which requires the ℓ_1-norm of ρ_0 to increase at most at the speed $L \asymp \sigma$.

SC(b) is a weak assumption: the bound $\bar{\lambda}$ and the penalty level λ can be chosen proportionally to $\sqrt{\log(p \vee n)}/\sqrt{n}$, that is

$$\tilde{\ell} \asymp \sqrt{\log(p \vee n)},$$

using self-normalized moderate deviation bounds (Jing et al., 2003; Belloni et al., 2014) or high-dimensional central limit theorems (Chernozhukov et al., 2017), under mild moment conditions, without requiring sub-Gaussianity. For instance, Belloni et al. (2014) employ these tools to show that, for the bounded design case $\|b\|_\infty \leq C$, λ can be chosen as in the Gaussian error case, provided that errors follow $t(2+\delta)$ distribution (having above 2 bounded moments), and get the error bounds similar to the Gaussian case. Here we state a general condition as our working assumption, instead of focusing on more specific condition that get us Gaussian-type conclusions.

THEOREM 4.3 (GDS FOR BLP AND RR). *Suppose SC holds. Then with probability at least $1 - K4\epsilon$, we have that $u = \hat{\beta}_A - \beta_0$ and $v = \hat{\rho}_A - \rho_0$ obey, for some absolute constant C,*

$$u'Gu \leq r_1^2, \quad v'Gv \leq \sigma^2 r_2^2, \quad |u'Gv| \leq \sigma r_3,$$

$$r_1^2 = C\mu_D^2(B^2\tilde{\ell}^2 s_\beta/n) \wedge (B^2\tilde{\ell}/\sqrt{n}), \quad r_2^2 = C\mu_D^2(B^2\tilde{\ell}^2 s_\rho/n) \wedge (B^2\tilde{\ell}/\sqrt{n}), \quad r_3 = r_1 r_2,$$

where s_β and s_ρ are the effective dimensions for parameters $D_\beta^{-1}\beta_0$ and $D_\rho^{-1}\rho_0$ for the noise level $\nu = 4B\tilde{\ell}/\sqrt{n}$.

In other words, we have instantiated (r_1, r_2, r_3) for approximately sparse models in the guarantee set S. We have the following corollary, which verifies $R(\delta)$ for approximately sparse models and hence provides sufficient conditions for Theorem 4.1.

COROLLARY 4.4 (SUFFICIENT CONDITION FOR $R(\delta)$). *Suppose SC holds. The guarantee $R(\delta)$ holds with $\varepsilon = 1 - K4\epsilon$, provided*

$$\text{either } Cs_\beta \leq \sqrt{n}\delta^2/(\tilde{\ell}^3\mu^2\mu_D^2) \text{ or } Cs_\rho \leq \sqrt{n}\delta^2/(\tilde{\ell}^3\mu^2\mu_D^2),$$

for some large enough constant C that only depends on B and K.

REMARK 4.1 (SHARPNESS OF CONDITIONS: DOUBLE SPARSITY ROBUSTNESS). This gives sufficient conditions such that (ignoring slowly growing term $\tilde{\ell}$) the condition $R(o(1))$ holds if

$$\text{either } s_\beta \ll \sqrt{n} \text{ or } s_\rho \ll \sqrt{n},$$

where s_β and s_ρ are measures of the effective dimensions of parameters $D_\beta^{-1}\beta_0$ and $D_\rho^{-1}\rho_0$. In well-behaved exactly sparse models, these effective dimensions are proportional to the sparsity indices divided by restricted eigenvalues. The latter possibility allows one of the parameter values to have unbounded effective dimension, in which case this parameter can be estimated at some 'slow' rate $n^{-1/4}$. These types of conditions appear to be rather sharp, matching similar conditions used in Javanmard and Montanari (2018) in the case of inference on a single coefficient in Gaussian exactly sparse linear regression models.

5. ESTIMATION AND INFERENCE USING GENERAL REGRESSION LEARNERS

In this section we generalize the previous analysis to allow for any regression learner $\hat{\gamma}$ of $E[Y|X]$ to be used in the construction of the estimator. As we have done in preceding sections we continue to include local functionals in our analysis, so that the results apply to nonregular objects as well as regular ones that can be estimated $\sqrt{n}$-consistently.

Compared to the global case, the local case may have smaller regularization and model selection biases relative to the variance. Nonetheless, bias correction is important for inference in theory and in practice. Theoretically, the local case begins to resemble the global case as the number of dimensions being integrated increases. Empirically, we provide local estimates without bias correction in Section S5. The differences can be substantial.

The only conditions we will impose on the regression learner are certain L^2 convergence properties that we will specify in this section. These properties will allow for a wide variety of learners, including GDS, Lasso, neural nets, boosting, and others. Thus we provide estimators of local functions that can be constructed using many regression learners.

We continue to consider estimators that use cross-fitting and have the form

$$\hat{\theta} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\{m(W_i, \hat{\gamma}_k) + \hat{\alpha}_k(X_i)[Y_i - \hat{\gamma}_k(X_i)]\},$$

where $\hat{\gamma}_k$ denotes the regression learner computed from observations not in I_k and $\hat{\alpha}_k(x) = b(x)'\hat{\rho}_k$ is the GDS learner of the RR described in previous sections.

To allow for as many regression learners as possible under as weak conditions as possible we focus on asymptotic analysis in this section. The fundamental property we will require of $\hat{\gamma}$ is that it have some mean square convergence rate as an estimator of the true conditional mean $\gamma_0^\star$. Specifically we require that there is $r_1^\star$ converging to zero such that for each k,

$$\left\|\hat{\gamma}_k - \gamma_0^\star\right\|_{P,2} = O_p(r_1^\star). \tag{5.1}$$

For purposes of formulating regularity conditions it is also useful to work with $\alpha_0^\star$ rather than α_0. We will also require that

$$\left\|\alpha_0^\star(\hat{\gamma}_k - \gamma_0^\star)\right\|_{P,2} = o_p(\sigma), \quad \left\|m(\cdot, \hat{\gamma}_k - \gamma_0^\star)\right\|_{P,2} = o_p(\sigma). \tag{5.2}$$

In the regular case these conditions generally follow from the mean square consistency of $\hat{\gamma}_k$ under boundedness of $\alpha_0^\star$. In nonregular cases they may impose additional conditions. For example, under the conditions of Lemma 4.3 it will be sufficient for these conditions to hold that

$$h^{-p_1/2}\left\|\hat{\gamma}_k - \gamma_0^\star\right\|_{P,2} \to_{\mathrm{P}} 0.$$

This condition will hold as long as h grows slowly enough relative to the mean square convergence rate of each $\hat{\gamma}_k$.

Recall from Theorem 4.3 that r_2 is the convergence rate of $\sigma^{-1}\left\|\hat{\alpha}_k - \alpha_0\right\|_{P,2}$. Let $r_2^\star = r_2 + \sigma^{-1}\left\|\alpha_0 - \alpha_0^\star\right\|_{P,2}$ and

$$\psi_0^\star(W) = \theta_0 - m(W, \gamma_0^\star) - \alpha_0^\star(X)[Y - \gamma_0^\star(X)].$$

THEOREM 5.1 (ASYMPTOTIC GAUSSIAN INFERENCE WITH GENERAL REGRESSION LEARNER). *Suppose $Var(Y|X)$ is bounded; $r_1^\star \to 0$ and $r_2^\star \to 0$; equations (5.1) and (5.2)*

are satisfied; and $\sqrt{n}r_1^\star r_2^\star \to 0$. *Then*

$$\hat{\theta} = \theta_0 - \frac{1}{n}\sum_{i=1}^{n} \psi_0^\star(W_i) + o_p\left(\frac{\sigma}{\sqrt{n}}\right), \quad \textit{hence} \quad \frac{\sqrt{n}}{\sigma}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1).$$

This result shows that asymptotic linearity of the estimator $\hat{\theta}$ will result if $r_2^\star \to 0$ fast enough relative to $r_1^\star$. Asymptotic linearity implies asymptotic Gaussian inference by standard central limit theorem arguments. As in regular doubly robust estimation problems it allows for a tradeoff between the speed of convergence $r_2^\star$ of the RR and $r_1^\star$ of the regression. It only requires a mean square convergence rate for the regression learner $\hat{\gamma}_k$ and so allows for a wide variety of first step machine learning estimators.

We could also formulate a nonasymptotic analogue to this asymptotic result. This would depend on the availability of nonasymptotic results for the learner $\hat{\gamma}_k$. To the best of our knowledge such results are not available for many learners, such as neural nets and random forests. To allow the results of this section to include as many first steps as possible we focus here on the asymptotic result and reserve the nonasymptotic result to future work.

ACKNOWLEDGEMENTS

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

REFERENCES

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231–63.

Altonji, J. G. and R. L. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–102.

Belloni, A., V. Chernozhukov, D. Chetverikov and I. Fernández-Val (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics 213*(1), 4–29.

Belloni, A., V. Chernozhukov and K. Kato (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika 102*(1), 77–94.

Belloni, A., V. Chernozhukov and L. Wang (2014). Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics 42*(2), 757–88.

Bickel, P. J., C. A. Klaassen, Y. Ritov and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

Bickel, P. J., Y. Ritov and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705–32.

Blundell, R., J. L. Horowitz and M. Parey (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics 3*(1), 29–51.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–51.

Chatterjee, S. (2013). Assumptionless consistency of the lasso. arXiv:1303.5817.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–68.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018b). Double/debiased machine learning for treatment and structural parameters: Replication package. *The Econometrics Journal 21*(1), Data deposited at https://doi.org/10.1111/ectj.12097.

Chernozhukov, V., D. Chetverikov and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability 45*(4), 2309–52.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey and J. M. Robins (2016). Locally robust semiparametric estimation. arXiv:1608.00033.

Chernozhukov, V. and C. Hansen (2004). The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *Review of Economics and Statistics 86*(3), 735–51.

Chernozhukov, V., W. K. Newey and J. Robins (2018d). Double/de-biased machine learning using regularized Riesz representers. Technical report, cemmap Working Paper, No. CWP15/18.

Chernozhukov, V., W. K. Newey and R. Singh (2018c). Automatic debiased machine learning of causal and structural effects. arXiv:1809.05224.

Chernozhukov, V., W. K. Newey and R. Singh (2021). A simple and general debiased machine learning theorem with finite sample guarantees. arXiv:2105.15197.

Donoho, D. L., M. Elad and V. N. Temlyakov (2005). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory 52*(1), 6–18.

Florens, J.-P., J. J. Heckman, C. Meghir and E. Vytlacil (2008). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica 76*(5), 1191–206.

Foster, D. J. and V. Syrgkanis (2019). Orthogonal statistical learning. arXiv:1901.09036.

Hausman, J. A. and W. K. Newey (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica 63*, 1445–76.

Hernan, M. A. and J. M. Robins (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–512.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

Javanmard, A. and A. Montanari (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics 46*(6A), 2593–622.

Jing, B.-Y., Q.-M. Shao and Q. Wang (2003). Self-normalized Cramér-type large deviations for independent random variables. *The Annals of Probability 31*(4), 2167–215.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. arXiv:2004.14497.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–82.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander (Ed.), *Probability and Statistics*, pp. 416–44. New York, NY: Wiley.

Peters, J., D. Janzing and B. Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge: MIT press.

Poterba, J. M., S. F. Venti and D. A. Wise (1995). Do 401(k) contributions crowd out other personal saving? *Journal of Public Economics 58*(1), 1–32.

Rothenhäusler, D. and B. Yu (2019). Incremental causal effects. arXiv:1907.13258.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics 14*(3), 1139–51.

Schmalensee, R. and T. M. Stoker (1999). Household gasoline demand in the United States. *Econometrica 67*(3), 645–62.

Semenova, V. and V. Chernozhukov (2021a). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal 24*(2), 264–89.

Semenova, V. and V. Chernozhukov (2021b). Debiased machine learning of conditional average treatment effects and other causal functions: Replication package. *The Econometrics Journal 24*(2), Data deposited at https://doi.org/10.1093/ectj/utaa027.

Tsybakov, A. B. (2012). *Introduction to Nonparametric Estimation*. New York: Springer.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Yatchew, A. and J. A. No (2001). Household gasoline demand in Canada. *Econometrica 69*(6), 1697–709.

Ye, F. and C.-H. Zhang (2010). Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research 11*(Dec), 3519–40.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

Co-editor Jaap Abbring handled this manuscript.