# DOUBLE/DE-BIASED MACHINE LEARNING OF GLOBAL AND LOCAL PARAMETERS USING REGULARIZED RIESZ REPRESENTERS

V. CHERNOZHUKOV, W. NEWEY, J. ROBINS, R. SINGH

Abstract. We provide adaptive inference methods, based on $\ell_1$ regularization methods, for regular (semi-parametric) and non-regular (nonparametric) linear functionals of the conditional expectation function. Examples of regular functionals include average treatment effects, policy effects from covariate distribution shifts and stochastic transformations, and average derivatives. Examples of non-regular functionals include the local linear functionals defined as local averages that approximate perfectly localized quantities: average treatment, average policy effects, and average derivatives, conditional on a covariate subvector fixed at a point. Our construction relies on building Neyman orthogonal equations for the target parameter that are approximately invariant to small perturbations of the nuisance parameters. To achieve this property we include the linear Riesz representer for the functionals in the equations as the additional nuisance parameter.

We use $\ell_1$-regularized methods to learn approximations to the regression function and the linear representer, in settings where dimension of (possibly overcomplete) dictionary of basis functions $p$ is much larger than $n$. We then estimate the linear functional by the solution to the empirical analog of the orthogonal equations. Our key result is that under weak assumptions the estimator of the functional concentrates in a $L/\sqrt{n}$ neighborhood of the target with deviations controlled by the Gaussian law, provided $L/\sqrt{n} \to 0$; $L$ is the operator norm of the functional, measuring the degree of its non-regularity, with $L$ diverging for local functionals (or under weak identification of the global functionals). Further conditions are needed to control bias if perfectly localized quantities are the target. For $\ell_1$ regularization methods, our construction and analysis yield weak requirements: either the approximation to the regression function or the approximation to the representer can be "completely dense" or "mildly sparse" as long as the other one is sufficiently "sparse". Our main results are non-asymptotic and imply asymptotic uniform validity over large classes of models, translating into honest confidence bands for both global and local parameters.

## 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 of 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 non-regular (estimable at slower than $1/\sqrt{n}$ rates). Global parameters can also be non-regular under weak identification (for example, in average treatment effects, when propensity scores can approach zero or one).

Often the regression is high dimensional, depending on many variables such as covariates that need to be controlled for. 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., Belloni et al. [2014c,d], Chernozhukov et al. [2016, 2018a]). 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). 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 than those used by the nonparametric learners.[1]

Using closed-form solutions for Riesz representers in several examples, Chernozhukov et al. [2016, 2018a] defined DML estimators in high dimensional settings and established their good properties. Compared to this approach, the new approach proposed in this paper has the following advantages and some limitations:

(1) The theory covers both regular (estimable at the $1/\sqrt{n}$ rate ) objects and nonregular ones (with rates $L/\sqrt{n}$, where $L \to \infty$ is the operator norm of the linear functional).

(2) The method automatically estimates the Riesz representer from the empirical analog of equations that implicitly characterize it.

(3) When a closed-form solution for the Riesz representer 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.

(4) Our approach remains interpretable under misspecification, estimating a linear functional of the projection rather than regression.

(5) While the current paper focuses only on sparse regression methods, the approach readily extends to cover other modern machine learning estimators of regression, just like Chernozhukov et al. [2018a]; we demonstrate this formally in Chernozhukov et al. [2018b].

(6) The current analysis focuses on linear functionals. In follow-up work we extend the approach to nonlinear functionals through a linearization; see Chernozhukov et al. [2018b].

The paper also builds upon ideas in classical semi- and nonparametric learning theory with low-dimensional $X$, using traditional smoothing methods [Van Der Vaart et al. [1991]; Newey [1994a]; Bickel et al. [1993]; Robins and Rotnitzky [1995]; Van der Vaart [2000]], that do not apply to the current high-dimensional setting. Our paper also builds upon and contributes to the literature on modern orthogonal/debiased estimation and inference [Zhang and Zhang [2014]; Belloni et al. [2011, 2014a,b,c,d]; Javanmard and Montanari [2014a,b,

[1]see, e.g, Schick [1986] for early use and Chernozhukov et al. [2018a] for more recent, in the context of debiased machine learning.

2018]; Van de Geer et al. [2014]; Ning and Liu [2017]; Chernozhukov et al. [2015]; Neykov et al. [2018]; Ren et al. [2015]; Jankova and Van De Geer [2015, 2016, 2018]; Bradic and Kolar [2017]; Zhu and Bradic [2017, 2018]], which focuses on coefficients in high-dimensional linear and generalized linear regression models, without considering the general linear functionals analyzed here.

The functionals we consider are different than those analyzed in Cai and Guo [2017]. The continuity properties of functionals we consider provide additional structure that we exploit, namely the Riesz representer, an object that is not considered in Cai and Guo [2017]. 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], Toth and van der Laan [2016], and Zheng et al. (2016). Here we provide DML learners via regularized RR, which are relatively simple to implement and analyze, and which directly target functionals of interest.

We build on previous work on debiased estimating equations constructed by adding an influence function. Hasminskii and Ibragimov [1979] and Bickel and Ritov [1988] suggest such estimators for functionals of a density. Doubly robust estimating equations as in Robins et al. [1995] and Robins and Rotnitzky [1995] have this structure. Newey et al. [1998, 2004] further develop theory in this vein, in low-dimensional nonparametric setting. In the regular case, Chernozhukov et al. [2016, 2018a] analyze the doubly robust learners in several high-dimensional settings. However, analysis requires an explicit formula for the Riesz representer, used in its estimation, which is often unavailable in closed form (or may be inefficient when restrictions such as additivity are used). In contrast, here we estimate the Riesz representer *automatically* from the moment conditions that characterize it, and extend the analysis to cover non-regular functionals.

The Athey et al. [2018] estimator of the (global) average treatment effect (ATE) is based on sparse linear regresssion and on approximate balancing weights when the regression is linear and strongly sparse. Our results apply to a much broader class of linear functionals and allow the regression learner to converge at relatively slow rates, including the dense case or approximately sparse case. Zhu and Bradic [2017] showed that it is possible to attain root-n consistency for the coefficients of a partially linear model when the regression function is dense. Our results apply to a much broader class of functionals, and allow for tradeoffs in accuracy of estimating the regression function and the Riesz representer. Hirshberg and Wager [2019] build upon the present work by considering the problem of learning regular functionals when the regression function belongs to a Donsker class. They utilize the orthogonal representations proposed in this paper and Chernozhukov et al. [2016], and extend the initial version of the paper Hirshberg and Wager [2017] that had only considered the ATE example. Our approach does not require a Donsker class assumption, which is too restrictive in our setting. The recent paper by Rothenhäusler and Yu [2019] also builds upon our work, analyzing global average derivative functionals, and proposing practical Lasso-type solvers for estimating the RR. Our approach is also practical; the RR estimation is based on a Dantzig-selector type estimator, which is easy to compute by linear programming methods. In the companion work Chernozhukov et al. [2018b], we also considered different Lasso-type solvers for estimating RR. Compared to Rothenhäusler and Yu [2019], our analysis covers a much broader collection of functionals, and deals with both local and global versions. A recent paper by Hirshberg and Wager [2018] also considers average derivative functional in a single index model, analyzing a variant of the estimator proposed here and in Chernozhukov et al. [2018b], adapted to the single-index regression structure.

A new development incorporated in this version of the present paper is the inclusion of local and localized functionals, such as average treatment/policy effects and derivatives localized to certain neighborhoods of a value of a low-dimensional covariate subvector. In low-dimensional nonparametrics, the study of such functionals, called "partial means" goes back, e.g., to Newey [1994b]. In contrast, here we treat the case where the ambient covariate space is very high-dimensional, but we localize with respect to a value of a low-dimensional subvector. Moreover, we must rely on orthogonalized estimating equations to eliminate the regularization biases arising due to the high-dimensional ambient space. This line of inquiry appears to be promising, with only few studies beginning to look at this. Independently and contemporaneously to the present version of the paper, Fan et al. [2019] and Zimmert and Lechner [2019] define and study perfectly localized average treatment effects with high-dimensional confounders; our development is complementary as it covers a much broader collection of functionals. Guo and Zhang [2019] study inference on the regression derivative $\partial\gamma_1(d)$ at a point $d$ in a high-dimensional regression model, $\gamma(D,Z) = \gamma_1(D)+\gamma_2(Z)$, where $D$ is univariate covariate of interest and $Z$ is a high-dimensional vector of control covariates. Our analysis is again complementary: it covers objects like this, but also covers more general functionals like $\mathrm{E}[\partial_d\gamma(D,Z) \mid D = d]$, either without additivity structure or without requiring $D$ to be one-dimensional. Moreover, we cover local effects, that are not perfectly localized, which may be more robust objects from an inferential point of view, as argued in Genovese and Wasserman [2008]. In a somewhat related development, Chernozhukov and Semenova [2017], apply low-dimensional series regression estimators on top of the pre-estimated unbiased orthogonal signal of treatment and partial derivative effects, where pre-estimation of the orthogonal signal is done in the high-dimensional setting. Our analysis has a rather different structure, and kernels are used for localization instead of series. Finally, Lee [2019] develops inference on perfectly localized average potential outcomes with continuous treatment effects, using a different approach than what we develop here.

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 provide estimation theory, demonstrating concentration and approximate Gaussianity of the DML estimator with Riesz representer estimated via regularized moment conditions. We provide rates of convergence for estimating the Riesz representer, giving both fast rates under approximate sparsity and slow rates under the dense model. In Section 4, we analyze the structure of the leading examples, providing bounds on operator norm, variance of the score, and kurtosis.

## 2. Target Functionals and Orthogonal Representations

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 sub-vector of $W$ taking values in their support sets, $y \in \mathcal{Y} \subset \mathbb{R}$ and $x \in \mathcal{X} \subset \mathbb{R}^d$. The law of $X$ is denoted by $F$. 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 $\Gamma_0$ for $\gamma_0^\star$ with with elements $\gamma$. (Later, in the estimation section, we can 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.[2] We refer to Imbens and Rubin [2015], Hernan and Robins [2019], and Peters et al. [2017] for the relevant formalizations that enable causal interpretation.

**Example 1** (**Average Treatment Effects**). *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 $Z$.*[3]

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

- $\ell(x) = 1$ gives average treatment effect in the entire population,
- $\ell(x) = 1(d = 1)/P(D = 1)$ gives the average treatment effect for the *treated* population,
- $\ell(x) = 1(x \in N)/P(X \in N)$ gives a localized average treatment effect for neighborhood or group $N$,

and so on. We can model small neighborhoods $N$ as shrinking in volume with the sample size. The localization and kernel weighting discussed below are applicable to all key examples.

[2] 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.

[3] 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 $Z$.

Moreover, they are combinable with other weighting functions so that, for example, we can target inference on localized average treatment effects for the treated.

**Example 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 weighing function $x \mapsto \ell(x)$, is given by:*

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

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

**Example 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 $\mathrm{E}[\tilde{Y}|X] = \gamma_0^\star(X + T(X))$. Exogeneity is a sufficient condition.*

**Example 4** (**Average Directional Derivative**)**.** *In the same settings as the previous example, a weighted average derivative of a continuously differentiable $\gamma_0$ 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 $\mathcal{X}$.*

All of these statistical parameters play an important role in causal, 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 finite-sample results and asymptotic results.

All of the above examples can be viewed as real-valued linear functionals of the regression function.

**Definition 1** (**Target Parameter**)**.** *Our target is the real-valued linear functional of $\gamma_0^\star$:*

$$\theta_0^\star = \theta(\gamma_0^\star), \text{ where } \gamma \mapsto \theta(\gamma) := \mathrm{E}m(W,\gamma), \tag{1}$$

*$\gamma \mapsto m(w,\gamma)$ 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:

(1) $m(w,\gamma) = (\gamma(1,z) - \gamma(0,z))\ell(x)$;

(2) $m(w,\gamma) = m(\gamma) = \int \gamma(x)\ell(x)d\mu(x); \mu(x) = F_1(x) - F_0(x)$;
(3) $m(w,\gamma) = \ell(x)(\gamma(T(x)) - \gamma(x))$;
(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

$$m(w,\gamma;\ell).$$

and, in examples 1,3, and 4, we can decompose

$$m(w,\gamma;\ell) = m_0(w,\gamma)\ell(x).$$

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

$$\ell_h(D) = \frac{1}{h^{p_1}} K\left(\frac{d_0 - D}{h}\right)/w, \quad w = \mathrm{E}\left[\frac{1}{h^{p_1}}K\left(\frac{d_0-D}{h}\right)\right], \quad h \in \mathbb{R}_+, \tag{2}$$

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

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

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 $\ell_h$ is known, i.e. $w$ is known.

**Definition 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*

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

2.2. **Building an Orthogonal Representation of the Target Functional.** A key quantity in the analysis is the operator norm (the modulus of continuity) of $\gamma \mapsto \theta(\gamma)$ on $\Gamma$, defined as

$$L = \sup_{\gamma\in\Gamma\setminus\{0\}} |\theta(\gamma)|/\|\gamma\|_{P,2}. \tag{3}$$

**Definition 3** (**Linear and Minimal Linear Representer**). *A linear representer for the linear functional $\gamma$ is $\alpha_0 \in L^2(F)$ such that*

$$\theta(\gamma) = \mathrm{E}\gamma(X)\alpha_0(X), \ \textit{for all } \gamma \in \Gamma. \tag{4}$$

*If $\alpha_0 \in \bar{\Gamma} :=$ closure$(\Gamma)$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 always exists when $L < \infty$ as a consequence of the Riesz-Frechet theorem, see Lemma 1 below. We can also generate linear representers in each example by various tools described below. When a linear representer exists, we define the following dual linear representation for the target parameter

$$\theta_0^\star = \theta(\alpha_0^\star); \quad \theta(\alpha) := \mathrm{E}[\alpha(X)Y]. \tag{5}$$

**Remark 1** (**On Direct and Dual Formulations**). *To motivate the upcoming orthogonal representation, we note that the direct and dual identification strategies can be used for direct plug-in estimation, but this does not give good estimators. 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 oder 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.*

**Definition 4** (**Orthogonal Representation for the Target Functional**). *We proceed to construct another representation for* $\theta_0^\star$ *that has the required Neyman orthogonality structure:*

$$\theta_0^\star = \theta(\alpha_0^\star, \gamma_0^\star); \quad \theta(\alpha, \gamma) := \mathrm{E}[m(W, \gamma) + \alpha(X)(Y - \gamma(X))]. \tag{6}$$

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

$$\frac{\partial}{\partial t}\theta(\alpha_0^\star + t\bar{h}, \gamma_0^\star + th)\Big|_{t=0} = \mathrm{E}m(W, h) - \mathrm{E}\alpha_0^\star(X)h(X) + \mathrm{E}[(Y - \gamma_0^\star(X))\bar{h}(X)] = 0. \tag{7}$$

In fact, a stronger property holds

$$\theta(\alpha, \gamma) - \theta(\alpha_0^\star, \gamma_0^\star) = -\int (\gamma - \gamma_0^\star)(\alpha - \alpha_0^\star) dF, \tag{8}$$

which implies double robustness [Chernozhukov et al., 2016, Proposition 5].

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] and Chernozhukov et al. [2018a]).

2.3. **The Case of Linear Regression.** We first consider the case of linear regression, where linearity holds with respect to a collection of basis functions. Suppose

$$\Gamma_0 = \Gamma_b = \{x \mapsto \gamma(x) = b(x)'\beta, \beta \in \mathbb{R}^p\},$$

where $x \mapsto b(x) = \{b_j(x)\}_{j=1}^p$ is a $p$-dimensional dictionary of basis functions with $b_j \in L^2(F)$ for each $j = 1, ..., p$. Define

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

so that $\theta(\gamma) = M'\beta$. For instance, in Examples 1-4:

(1) $M = \mathrm{E}(b(1, Z) - b(0, Z))\ell(X)$,

(2) $M = \int b\ell(dF_1 - dF_0)$,
(3) $M = \mathrm{E}(b(T(X)) - b(X))\ell(X)$,
(4) $M = \mathrm{E}\partial_d b(D, Z)t(X)\ell(X)$.

For this space we can take the linear representer $\alpha_0^\star$ in the same space $\Gamma$ in the form of

$$\alpha_0^\star(x) = b(x)'\rho_0.$$

We can define the parameters $\beta_0$ and $\rho_0$ as minimal $\ell_1$-norm solutions to the system of equations:

$$\min \|\beta\|_1 + \|\rho\|_1 : \quad G\beta = \mathrm{E}Yb(X), \quad G\rho = M. \tag{9}$$

Of course, if $G$ is full rank, then $\gamma_0^\star(x) = b(x)'\beta_0$ with $\beta_0 = G^{-1}\mathrm{E}b(X)Y$, and $\alpha_0^\star = b'\rho_0$ with $\rho_0 = G^{-1}M$.

We see the representation property from

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

The operator norm of $\theta(\gamma) = M'\beta$ is given by

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

The 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 $\beta$ is $\gamma$'s parameter and $\rho$ is $\alpha$'s parameter.

2.4. **Existence of Linear Representers.** We employ the Riesz-Frechet representation theorem and Hahn-Banach extension theorem to generate the linear Riesz representer.

**Lemma 1** (**Extended Riesz Representation**). *(i) If $L < \infty$, there exists a unique minimal representer $\alpha_0^\star \in \bar{\Gamma}$ and $L = \|\alpha_0^{\star 2}\|_{P,2}$. (ii) If there exists a linear representer $\alpha_0$ on $\Gamma$ with $\|\alpha_0^2\|_{P,2} < \infty$, then $L = \|\alpha_0^{\star 2}\|_{P,2} \leq \|\alpha_0^2\|_{P,2} < \infty$, and $\alpha_0^\star$ 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 allows us to generate implicitly linear representers for continuous linear functionals. The second part of the lemma allows us to use any representer obtained through, for example, change of measure and integration by parts technique, to claim continuity of the functionals, and generate minimal representers. Our estimation results will rely only on existence of minimal representers, although we can utilize the knowledge of linear representers to improve the basis functions for estimating the minimal representers.

The following linear representers are derived by change of measure and integration by parts. We refer to them as universal since, when they exist, they can represent the linear functionals even when span of $\Gamma_0$ is dense in $L^2(F)$, i.e. when $\bar{\Gamma} = L^2(F)$. There is an efficiency reason to work with minimal representers rather than universal representers: Using universal representers we attain full semi-parametric efficiency only when they are minimal, i.e. when $\Gamma$ is dense in $L^2(F)$.

Consider the following candidates for universal linear representers in Examples 1-4:

$$\begin{aligned}
\alpha_0(x;\ell) &= [(1(d=1)-1(d=0))/\mathrm{P}(D=d\mid Z=z)]\ell(x); && (10)\\
\alpha_0(x;\ell) &= [d(F_1(x)-F_0(x))/dF(x)]\ell(x); && (11)\\
\alpha_0(x;\ell) &= [d(F_1(x)-F(x))/dF(x)]\ell(x),\ F_1=\mathrm{Law}(T(X)); && (12)\\
\alpha_0(x;\ell) &= -(\mathrm{div}_d(\ell(x)t(x)f(d|z))/f(d|z),\ f(d|z)= \text{ pdf of } D \text{ given } Z=z; && (13)
\end{aligned}$$

treated as formal maps $\alpha_0 : \mathcal{X} \to \mathbb{R} \cup \{\mathrm{na}\}$, where $dF_k/dF$ denotes the Radon-Nykodym derivative of measure $F_k$ with respect to $F$ on support$(\ell)$, $\mathrm{div}_d$ denotes the divergence of scalar function:

$$\mathrm{div}_d\ g(d,z) = \sum_{j=1}^{p_1} \partial_{d_j} g(d,z),$$

and na is "not available". The Radon-Nykodym derivatives exist if $F_k$ is absolutely continuous with respect to $F$ on support$(\ell)$.

**Lemma 2** (**Universal Representers for Key Examples**). *In Example 1-4, (i) If $\alpha_0(X,\ell)$ is real-valued a.s. and $\alpha_0 \in L^2(F)$, then it is the universal representer for the corresponding linear functional $\gamma \mapsto \theta(\gamma)$, and the latter is continuous. In Example 4, we require that $d \mapsto \gamma(x)\ell(x)t(x)f(d|z)$ is continuously differentiable on the support set $\mathcal{D}_z = \mathrm{support}(D|Z=z)$, and vanishes on its boundary $\partial\mathcal{D}_z$, which is assumed to be piecewise-smooth, for each $z \in \mathcal{Z}$. Further, if $\bar{\Gamma} = L^2(F)$, the representer is minimal; otherwise, minimal representer $\alpha_0^\star$ is obtained by projecting $\alpha_0$ onto $\bar{\Gamma}$. (ii) There are substantive examples of $P$, exhibited in the proof of this lemma, such that linear functionals in Examples 1-4 can be continuous on $\Gamma$, but $\alpha_0(X) = \mathrm{na}$ with positive probability.*

The first part of the lemma provides a simple sufficient condition to guarantee continuity of the target functionals. It recovers well-known sufficient conditions for nonparametric identification of various functionals. The second part of the lemma states that this condition is not necessary, and that target functionals can be continuous without these conditions.

The following is a useful result in view of the wide practical use of additive models, which model the regression function as additive in the two sets of vector components $x_1$ and $x_2$ of $x$. (There is not much loss in generality in considering two sets, rather than multiple sets). It is an important setting where $\Gamma$ is not dense in $L^2(F)$.

AM *Suppose that the regression function is additive in components $x_1$ and $x_2$ of $x$:*

$$\gamma(x) = \gamma_1(x_1) + \gamma(x_2), \quad x = (x_1', x_2')' \in \mathcal{X}$$

*where $\gamma_1 \in \Gamma_{01}$, a dense subset of $L^2(F_1)$, where $F_1$ denotes the probability law of $X_1$. The linear functional $m_0$ and the weighing function $\ell$ depends only on the first component, namely $m_0(w,\gamma) = m_0(w,\gamma_1)$ and $\ell(x) = \ell(x_1)$.*

The following lemma shows that we can construct representers for additive models by taking conditional expectation of a universal representer.

**Lemma 3** (**Order-Preserving, Contractive Representers for Additive Models**). *Work with AM and assume $L$ is finite. Then on $\gamma \in \Gamma$,*

$$\theta(\gamma) = \theta(\gamma_1) = \int \alpha_0^\star(x_1)\gamma_1(x_1)dF(x), \quad \alpha_0^\star(x_1) = \mathrm{E}[\alpha_0(X) \mid X_1 = x_1],$$

*where $\alpha_0$ is any linear representer for $\gamma \mapsto \theta(\gamma)$ on $\Gamma$. In particular, the conditional expectation operator is order-preserving, and it induces the contraction for all $L^q(P)$ norms for all $q \in [1, \infty]$:*

$$\|\alpha_0^\star\|_{P,q} \leq \|\alpha_0\|_{P,q}.$$

The latter properties are useful in characterizing the structure of the global and local functionals.

2.5. **Preview of Estimation and Inference.** Our estimation and inference will exploit an empirical analog of (6), given a random sample $(W_i)_{i=1}^n$ generated as i.i.d. copies of $W$. In place of the unknown regression $\gamma_0^\star$ and the Riesz representer $\alpha_0^\star$, we will plug-in estimators obtained using $\ell_1-$ regularization. Estimation of the regression function will be standard but estimation of the Riesz representer will be based on a novel approach: we use the minimal $\ell_1$-norm solutions to a relaxed version of the empirical analog of the moment condition (4) that defines the representation property, based on a collection of basis/test functions in $\Gamma$; see the definitions on the next section.

We shall use the empirical analog of the orthogonal representation of the target functional to estimate the parameter. We shall use sample-splitting in the form of cross-fitting to avoid biases from overfitting that can arise in high-dimensional settings.

**Definition 5** (**DML with Linear Representers**). *Let $(I_k)_{k=1}^K$ be a partition of the observation index set $\{1, ..., n\}$ into $K$ distinct subsets of about equal size. Let $\hat{\gamma}_k$ and $\hat{\alpha}_k$ be estimators constructed from data, leaving out block $I_k$, i.e. from $\{W_i\}_{i \notin I_k}$. Then*

$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in I_k} \{m(W_i, \hat{\gamma}_k) + \hat{\alpha}_k(X_i)[Y_i - \hat{\gamma}_k(X_i)]\}. \tag{14}$$

A key variable in the analysis is the "true" score

$$\psi_0(W) := m(W, \gamma_0^\star) + \alpha^\star(X)(Y - \gamma_0^\star(X)) - \theta_0^\star$$

and its moments:

$$\sigma^2 := \mathrm{E}\psi_0^2(W), \quad \kappa^3 := \mathrm{E}|\psi_0^3(W)|.$$

**Concentration and Approximate Gaussianity Result.** We establish that the resulting de-biased (or “double") machine learning (DML) estimator $\hat{\theta}$ is approximated by the oracle estimator

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

and, consequently, concentrates in a $\sigma/\sqrt{n}$ neighborhood of the target with deviations controlled by the normal laws,

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

where the $\mathrm{error}_n$ bound is non-asymptotic and tends to zero as $n \to \infty$, where $\sigma/\sqrt{n} \to 0$ is required for concentration. This implies the uniform validity of results over large classes of probability laws $P$ for $W$. The result also shows that the estimator is adaptive, since it has the approximate deviation determined by $\|\psi_0\|_{P,2}$, which is the *variance* of the oracle estimator that knows the true score.

There are two cases to consider:

Regular Case: *the parameters $\sigma$, $\kappa/\sigma$, and $L$ are bounded, leading to $1/\sqrt{n}$ concentration, adaptation, and Gaussian approximation.*

Non-Regular Case, *the parameters $\sigma, \kappa/\sigma$, 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.*

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

As we show in Section 4, in the leading non-regular cases, the latter condition can be more succinctly stated as

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

This is the case for localized functionals, discussed formally below where the source of non-regularity is the localization.

## 3. Estimation and Inference Results for High Dimensional Approximately Linear Models

### 3.1. **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 $n$.

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

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

where $r_\gamma$ is the approximation error, and $\gamma_0 := b'\beta_0$ is the best linear predictor of $Y$ and best linear approximation to $\gamma_0^\star$. We define $\beta_0$ as a minimal $\ell_1$-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. Let $\Gamma_b$ be the linear subspace of $L^3(F)$ generated by $b$.

Similarly, we approximate the Riesz representer via the best linear approximation $\alpha_0$:

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

We define $\rho_0$ as a minimal $\ell_1$-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

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

Hence $\alpha_0$ is the Riesz representer for $\mathrm{E}m(W, \gamma)$ for each $\gamma \in \Gamma_b$.

In some of the asymptotic results that follow, we can have $\Gamma_b \to \bar{\Gamma}_0$ as $p \to \infty$, in which case $\|r_\alpha\|_{P,2} \to 0$, and $\|\alpha_0 - \alpha_0^\star\|_{P,2} \to 0$.

**Definition 6** (**Penultimate and Ultimate Target Parameters**). *Our penultimate target is the linear functional applied to the BLP $\gamma_0$:*

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

*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 $(\sqrt{n}/\sigma) \int r_\alpha r_\gamma dF \to 0$ our inference will target the ultimate parameter. Otherwise, under misspecification, our inference will target an interpretable penultimate parameter.

Our DML estimator of $\theta_0$ will be based on the following score function:

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

**Lemma 4** (**Basic Properties of the Score**). *The score function 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 $(\beta_0, \rho_0)$:*

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

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

The second claim of the lemma is immediate from the definition of $(\beta_0, \rho_0)$ 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 $\rho$ and $\beta$ around their "true values" $\rho_0$ and $\beta_0$. This invariance property plays a crucial role in removing the impact of biased estimation of nuisance parameters $\rho_0$ and $\beta_0$ on the estimation of the main parameters $\theta_0$.

3.2. **Estimators.** Estimation will be carried out using the following Dantzig Selector-type estimators (Candes and Tao [2007]).[4]

**Definition 7** (**Regularized Minimum Distance 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 $p$ non-negative-definite matrices and $M$ and $\hat{M}$ are $p$-vectors. Define $t_0$ as a minimal $\ell_1$-norm solution to $g(t) = 0$ and assume $t_0 \in T$. Define the RMD estimator $\hat{t}$ by solving*

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

*where $\lambda$ 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.

Let $(W)_{i=1}^n = (Y_i, X_i)_{i=1}^n$ denote i.i.d. copies of $W$. We define the estimators of $\beta_0$ and $\rho_0$ over subset $A$ of data. Let $\mathbb{E}_A f$ denote the empirical average of $f(W)$ over $i \in A \subset \{1, ..., n\}$:

$$\mathbb{E}_A f := \mathbb{E}_A f(W) = |A|^{-1} \sum_{i \in A} f(W_i).$$

**Definition 8** (**RMD for BLP: Dantzig Selector**). *Given a diagonal positive normalization matrix $D_\beta$, define $\hat{\beta}_A = D_\beta \hat{t}$, where $\hat{t}$ is the RMD estimator for $t_0 = D_\beta^{-1}\beta_0$ with*

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

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, $D_\beta$ is chosen such that $\mathrm{diag}(Var(D_\beta^{-1}(\hat{G}\beta_0 - \hat{M})) = I$. Our practical algorithm estimates $D_\beta$ from the data.

**Definition 9** (**RMD for Riesz Representer**). *Given a diagonal positive normalization matrix $D_\rho$, define $\hat{\rho}_A = D_\rho \hat{t}$, where $\hat{t}$ is the RMD estimator of the parameter $t_0 = D_\rho^{-1}\rho_0$ with*

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

In practice, we are using $T_\rho = \mathbb{R}^p$, even though it is possible to exploit some structured restrictions on the problem motivated the nature of the universal Riesz representers. Ideally, $D_\rho$ is chosen such that $\mathrm{diag}(Var(D_\rho^{-1}(\hat{G}\rho_0 - \hat{M})) = I$. Our practical algorithm estimates $D_\rho$ from the data.

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

[4]In a follow-up work Chernozhukov et al. [2018b] we also consider Lasso-type estimators.

**Definition 10 (DML with RR).** *Consider the partition of* $\{1, ..., n\}$ *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* $k = 1, ..., K$, *let* $\hat{\beta}_k$ *and* $\hat{\rho}_k$ *denote RMD estimators obtained using data* $(W_i)_{i \in I_k^c}$, *where* $I_k^c = \{1, ..., n\} \setminus I_k$, *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:*

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

3.3. **Properties of DML: Main Result.** We provide a single finite-sample result that allows us to cover both global and local functionals, implying uniformly valid rates of concentration and normal approximations over large classes 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 $\sigma$ is bounded, giving $1/\sqrt{n}$ concentration around $\theta_0$, and for non-regular functionals $\sigma \propto L \to \infty$ requring $L/\sqrt{n} \to 0$ to get concentration. Our normal approximation is accurate if kurtosis of $\psi_0$ does not grow to 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 non-regular 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:

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

where positive numbers $r_1$, $r_2$, and $r_3$ measure the quality of guarantee. Define $\mu$ 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 $\mu$ 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.[5]

Consider $P$ that satisfies the following conditions.

[5]If $b(X) = X$ are sub-Gaussian, then this assumption is also easily satisfied; however, this case is not of central interest to us.

R($\delta$) *With probability $1-\varepsilon$, the estimation errors $\{(\hat\beta_k - \beta_0, \hat\rho_k - \rho_0)\}_{k=1}^K$ take values in $\mathsf{S}$, with performance guarantees obeying*

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

A sufficient condition for $R(\delta)$ is given in the next section.

**Theorem 1** (**Adaptive Estimation and Approximate Gaussian Inference**). *Suppose $K$ divides $n$ for simplicity. Under condition R, 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 $\theta_0$, 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 constants can be chosen to yield an asymptotic result.

**Corollary 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 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 $\psi_0$ 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 BLP function $\gamma_0$ enjoys good properties under the stated regularity conditions. This result does not distinguish between inference on global functions 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 2** (**Inference on the Ultimate Parameter $\theta_0^\star$**). *Suppose that, in addition, $P$ satisfies the small approximation error condition:*

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

*Then conclusions of Theorem 1 hold with $\theta_0^\star$ replacing $\theta_0$, with $\sqrt{K}4\delta/\Delta$ increased by $\delta$, and the same probability. Conclusions of Corollary 1 continue to hold with $\theta_0^\star$ replacing $\theta_0$ for a class of probability laws $\mathcal{P}_n$, provided each $P \in \mathcal{P}_n$ satisfies the conditions above for the given $\delta = \delta_n \to 0$.*

The approximation bias for the ultimate target is 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. [2014d] show small approximation bias using Fourier bases as dictionaries, and using Sobolev and rearranged Sobolev balls, respectively, as the function classes.

**Corollary 3** (**Inference on the Perfectly Localized Parameter**). *Suppose that, in addition, $P$ satisfies the small approximation error condition:*

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

*and the localization bias is small:*

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

*Then conclusions of Theorem 1 hold with $\theta_0(\gamma_0^\star;\ell_0)$ replacing $\theta_0$, with $\sqrt{K}4\delta/\Delta$ increased by $2\delta$, and the same probability. Conclusions of Corollary 1 continue to hold with $\theta_0^\star(\gamma^\star;\ell_0)$ replacing $\theta_0$ for a class of probability laws $\mathcal{P}_n$, provided each $P \in \mathcal{P}_n$ satisfies the conditions above for the given $\delta = \delta_n \to 0$.*

The next section gives primitive conditions for bounding the localization bias.

3.4. **Efficiency for Estimating Penultimate Parameter $\theta_0$.** The DML-RR estimator $\hat{\theta}$ will be asymptotically efficient for estimating the penultimate parameter $\theta_0$, which is defined in terms of $\gamma_0$, the mean-square projection of $Y$ on $\bar{\Gamma}$ under $P$. 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. Our formal result only covers the regular case under fixed $P$ with the operator norm $L$ bounded, but we expect that a similar result continues to hold with $L \to \infty$, through the use of an appropriate formalization that handles $P$ changing with $n$ and rules out super-efficiency phenomena.

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 **?**. To confirm this intuition we give a precise result that constructs a class of regular parametric submodels for which the estimator here is locally regular and for which the closure of the set of scores is all measureable functions with zero mean and finite variance.

**Theorem 2** (**Efficiency Theorem**). *If $\mathrm{E}[Y^2] < \infty$, $\mathrm{E}[\psi_0(W)^2] < \infty$, , and $m(W,\gamma)$ is mean square continuous in $\gamma$, then $\psi_0(W)$ is the efficient (and only) influence function, so that the conclusion of Theorem 25.20 of Van der Vaart [2000] holds.*

3.5. **Properties of RMD Estimators.** Our goal is to verify that the guarantee $R(\delta)$ holds. In particular we have to bound the population prediction norm

$$\sqrt{\delta' G \delta}.$$

This is a more nuanced problem than bounding the empirical prediction norm $\sqrt{\delta' \hat{G} \delta}$, 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 $\ell_1$ 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* $\bar{\lambda} \geq \lambda$

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

The bounds on $\ell_1$ norm of coefficients are naturally motivated, for example, by working in Sobolev or rearranged Sobolev spaces, (see, Tsybakov [2012] and Belloni et al. [2014d], 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 behaviors than in the original Sobolev spaces.[6]

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 $\nu$ is the noise level. As demonstrated in the proof, the RMD estimator belongs to this set with high probability $1 - \epsilon$ for the noise level:

$$\nu = 4B\bar{\lambda},$$

where $\lambda$ is the penalty level of RMD ($\nu$ scales like $\sqrt{\log(p \vee n)}/\sqrt{n}$ in our problems).

**Definition 11** (**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) Gram matrix $G$, which makes it easy to verify regularity conditions. Note that

$$\text{if } G = I \text{ and } \|t_0\|_0 = s, \text{ then } s(t_0) \leq s.$$

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 $\ell_1-$ penalization has been

[6]The complexity of these function classes are also different. Sobolev spaces are Donsker sets under some conditions, whereas rearranged Sobolev spaces have the covering entropy bounded below by $\log p$ and are not Donsker if $p \to \infty$.

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 $\delta_A$ denote a vector with the $j$-th component set to $\delta_j$ if $j \in A$ and 0 if $j \notin A$.

**Lemma 5** (**A Bound on Effective Dimension in Approximately Sparse Model**). *Suppose that $t_0$ is approximately sparse, namely*

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

*for some finite positive constants $A$ and $a > 1$, where $(|t_0|_j^*)_{j=1}^p$ is the non-increasing re-arrangement 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 $\nu$ 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$(t_0^{\mathcal{M}})$, $\mathcal{M}^c = \{1, ..., p\} \setminus \mathcal{M}$, 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].

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 $\ell_1$-regularized estimators of approximately sparse $t_0$. The effective dimension is simply stated in terms of $t_0$ alone.

**Lemma 6** (**Finite-Sample Bound for RMD 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 and Jafarov [2015] 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" $s(t_0)\nu^2$ 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 RMD estimators of the Riesz representer and the Dantzing selector. 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 $\infty$ as $n \to \infty$ in the asymptotic results).

SC *(a) The $\ell_1$ norms of coefficients are bounded as $\|D_\rho^{-1}\rho_0\|_1 \le B$ and $\|D_\beta^{-1}\beta_0\|_1 \le B$, for $B \ge 1$,and the scaling matrices obey $\|D_\rho v\| \le \mu_D \sigma \|v\|$ for $D_\rho^{-1} v \in S(D^{-1}\rho_0, \nu)$ and $\|D_\beta u\| \le \mu_D \|u\|$ 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 $\{1, ..., n\}$ of size $m \ge n - \lfloor n/K \rfloor$, dictionary $b$ obeys with probability at least $1-\epsilon$, $\|\mathbb{G}_A bb'\|_\infty \le \tilde{\ell}$. (c) The penalty levels $\lambda_\alpha$ and $\lambda_\beta$ are chosen such that with probability at least $1-\epsilon$ $\|D_\beta^{-1}(\mathbb{G}_A bb'\beta_0 - \mathbb{G}_A Y b(X))\|_\infty/\sqrt{m} \le \lambda_\rho$, $\|D_\rho^{-1}(\mathbb{G}_A bb'\beta_0 - \mathbb{G}_A m(W, b))\|_\infty/\sqrt{m} \le \lambda_\beta$, and are not overly large, $\lambda^\beta \vee \lambda^\rho \le \tilde{\ell}/\sqrt{m}$.*

SC(a) records a restriction on the $\ell_1$ norm of $\beta_0$ and $\rho_0$. For instance, in Examples 1-3, $D_\rho \asymp \sigma I \asymp LI$, which requires the $\ell_1$-norm of $\rho_0$ to increase at most at the speed $L \asymp \sigma$.

SC(b) is a weak assumption: the bound $\bar{\lambda}$ and the penalty level $\lambda$ can be chosen proprtionally 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., 2014d] or high-dimensional central limit theorems [Chernozhukov et al., 2017], under mild moment conditions, without requiring sub-Gaussianity. For instance, Belloni et al. [2014d] employ these tools to show that, for the bounded design case $\|b\|_\infty \le C$, $\lambda$ 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 3** (**RMD 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 $C$,*

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

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

*where $s_\beta$ and $s_\rho$ are the effective dimensions for parameters $D_\beta^{-1}\beta_0$ and $D_\rho^{-1}\rho_0$ for the noise level $\nu = 4B\tilde{\ell}/\sqrt{n}$. Hence the guarantee $R(\delta)$ holds with $\varepsilon = 1 - K4\epsilon$, provided*

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

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

**Remark 2** (**Sharpness of Conditions**). *This gives sufficient conditions such that (ignoring slowly growing term $\tilde{\ell}$) the condition $R(o(1))$ holds if*

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

*where $s_\beta$ and $s_\rho$ 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 be "dense", having 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.*

## 4. Structure of Functionals and Their Scores in Leading Examples

4.1. **Structure of Global Functionals and Scores.** Here we develop bounds on the key quantities: the standard deviation $\sigma$ of the score, its kurtosis $\kappa/\sigma$, and the modulus of continuity $L$. In the regular case, these quantities bounded. Here we would like to study how the bounds depend on $L$, and we analyze the non-regular cases arising from taking sequence of models with $L \to \infty$.

To make key points, we focus on the case where either $\bar{\Gamma} = L^2(F)$ or $\bar{\Gamma} \subset L^2(F)$ with the additive model AM holding. Furthermore, we develop these bounds in the context of Examples 1-3, though the proofs are useful to characterize bounds in other contexts. Our goal is to fix a weighting function $\ell$, and to consider how a non-regularity $L \to \infty$ can arise from modeling quantities like[7]

$$1/\mathrm{P}(D = d \mid Z), \quad (d(F_1 - F_0)/dF) \circ X, \quad (d(F_1 - F)/dF) \circ X, \tag{15}$$

taking high values due to the denominator taking values close to zero. We may characterize such cases as the weakening of overlap of supports of relevant distributions (e.g., $F$ puts small mass on points where $F_1$ puts a lot of mass).

In the sequel, we say that $a \lesssim b$ under the asymptotics with an index $n \to \infty$ if $a \leq Cb$ for all $n$ sufficiently large, and $a \asymp b$ if both $a \lesssim Cb$ and $b \lesssim Ca$ for all $n$ sufficiently large, where $C \geq 1$ is a positive constant that does not depend on $n$.

**Lemma 7** (**Structure of Global Average Effects Functionals in Examples 1,2,3**). *Suppose that either (a) $\bar{\Gamma} = L^2(F)$ or (b) that $\bar{\Gamma} \subset L^2(F)$ with the additive model AM holding. Suppose that the universal Riesz representers $\alpha_0(X) = \alpha_0(X, \ell)$ given in formulae (10), (11), (12) for Examples 1-3 exist and are in $L^2(F)$. Suppose that $\alpha_0^\star(X) = \alpha_0(X)$ in the case (a) and $\alpha_0^\star(X_1) = \mathrm{E}[\alpha_0^\star(X) \mid X_1]$ in the case (b) obey:*

$$\|\alpha_0^\star\|_{P,3} \leq c(\|\alpha_0^\star\|_{P,2}^2 \vee 1), \tag{16}$$

*for some finite constant $c$ and that*

$$U_1 = m(W, \gamma_0^\star(X)) - \mathrm{E}m(W, \gamma_0^\star(X)) \text{ and } U_2 = Y - \gamma_0^\star(X)$$

*obey the bounded moment and bounded heteroscedasticity conditions:*

$$(\mathrm{E}[|U_1|^q])^{1/q} \leq \bar{c}, \quad 0 < \underline{c} \leq (\mathrm{E}[|U_2|^q|X])^{1/q} \leq \bar{c} \text{ a.s., for } q \in \{2, 3\},$$

*for some finite positive constants $\underline{c}$ and $\bar{c}$. Then*

$$\underline{c}L \leq \sigma \leq \bar{c}\sqrt{1 + L^2}, \quad \kappa \leq \bar{c}(1 + c(L^2 \vee 1)).$$

*If, as $n \to \infty$, we have that $L \to \infty$ and the constants $(c, \underline{c}, \bar{c})$ are bounded away from zero and above, then*

$$(\kappa/\sigma) \lesssim \sigma \asymp L \to \infty.$$

[7] In Example 4, a similar issue could arise due to $1/f(D|Z)$ taking high values; for brevity, we don't analyze this source of non-regularity for Example 4 and focus on localization as the source.

Condition (16) allows the $L_3(F)$ norm of the representer to be much larger than the $L_2(F)$ norm, but limits how much larger. For instance, consider Example 1. Suppose $\bar{\Gamma} = L^2(F)$ so that $\alpha^\star = \alpha_0$ and that the propensity score $P[D = 1 \mid Z]$ is uniformly distributed on $[\pi, 1/2]$. Then $\|\alpha_0\|_{P,2} \asymp (1/\pi)^{1/2}$ and $\|\alpha_0\|_{P,3} \asymp (1/\pi^2)^{1/3} \ll \|\alpha_0\|^2_{P,2}$ when $\pi \searrow 0$, so the condition is easily met.

4.2. **Structure of Local and Localized Functionals and Scores.** Here we focus on local functionals and develop bounds that relate key quantities: the standard deviation $\sigma$ of the score, its kurtosis $\kappa/\sigma$, and the modulus of continuity $L$. Our first goal is examine how the localization of the weighting function $\ell$ creates the non-regularity $L \to \infty$. Our inference theory outlined above covers the local functional provided $L/\sqrt{n}$ is small, and it also covers perfectly localized functional provided the scaled localization bias is small:

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

We provide the bound on the localization bias in terms of the smoothness and the kernel order. The latter additional requirement means that the inference on perfectly localized functionals is less robust than the inference on the local functionals (analogously, to the point that was made by Genovese and Wasserman [2008]).

**Lemma 8** (**Structure of Local Average Effects Functionals and Scores in Examples 1, 2, 3**). *Suppose that either (a) $\bar{\Gamma} = L^2(F)$ or (b) $\bar{\Gamma} \subset L^2(F)$ with the additive model AM holding. Suppose the universal Riesz representer $\alpha_0(X, 1)$, corresponding to the flat weighting function $\ell = 1$, given in formulae (10), (11), and (12), corresponding to Examples 1,2, and 3, exists and obeys*

$$0 < \underline{\alpha} \leq \alpha_0(X, 1) \leq \bar{\alpha}, \qquad a.s. \tag{17}$$

*Suppose for some $h_0 > 0$, we have that $N_{h_0}(d_0) = \{d : \|d - d_0\|_\infty \leq h\} \subset \mathcal{D}$. Suppose that for $\ell = \ell_h$ with $h \leq h_0$:*

$$U_1 = m(W; \gamma_0^\star(X), \ell) - \mathrm{E}m(W; \gamma_0^\star(X), \ell) \text{ and } U_2 = Y - \gamma_0^\star(X),$$

*obey the bounded heteroscedastic moment conditions:*

$$(\mathrm{E}[|U_1|^q])^{1/q} \leq \bar{c}\|\ell\|_{P,q}, \quad 0 < \underline{c} \leq (\mathrm{E}[|U_2|^q|X])^{1/q} \leq \bar{c} \text{ a.s., for } q \in \{2, 3\}.$$

*Suppose that the pdf $f_D$ of $D$ obeys the bounds:*

$$0 < \underline{f} \leq f_D(d) \leq \bar{f} \text{ and } \|\partial f_D(d)\|_1 \leq \bar{f}', \text{ for all } d \in N_{h_0}(d_0).$$

*Then the finite-sample bounds stated in the proof of this lemma hold. In particular, if $h \searrow 0$ and $(\underline{\alpha}, \bar{\alpha}, \underline{c}, \bar{c}, \underline{f}, \bar{f}, \bar{f}', h_0)$ are bounded away from zero and bounded above, then*

$$(\kappa/\sigma) \lesssim h^{-p_1/6} \lesssim \sigma \asymp L \asymp \|\ell\|_{P,2} \asymp h^{-p_1/2} \to \infty.$$

The lemma shows that the main source of non-regularity is the bandwidth going to zero. The condition (17) shuts down the previous source of non-regularity, and says that the quantities in (15) are now bounded from below and above. It is possible to analyze the case where both sources of non-regularity are present and to bound behavior of $\sigma, \kappa/\sigma$, and $L$. Our general inference theory allows for such complicated sources of nonregularity as long as these parameters are much smaller than $\sqrt{n}$.

We now turn to characterization of the local average derivatives.

**Lemma 9** (**Structure of Local Average Derivative Functionals and Scores in Example 4**). *Suppose that either (a) $\bar{\Gamma} = L^2(F)$ or that (b) $\bar{\Gamma} \subset L^2(F)$ with the additive model AM holding. Suppose the universal Riesz representer $\alpha_0(X, \ell_h)$ given in formula (13) exists for all $0 < h < h_0$, where $h_0$ is a constant. Suppose that the errors*

$$U_1 = m_0(W; \gamma_0^\star(X))\ell_h(X) - \mathrm{E}m_0(W; \gamma_0^\star(X))\ell_h(X) \text{ and } U_2 = Y - \gamma_0^\star(X)$$

*obey the bounded heteroscedastic moment conditions:*

$$(\mathrm{E}[|U_1|^q])^{1/q} \leq \bar{c}\|\ell_h\|_{P,q}, \quad 0 < \underline{c} \leq (\mathrm{E}[|U_2|^q|X])^{1/q} \leq \bar{c}, \ a.s., \quad q \in \{2, 3\}.$$

*Suppose that $N_h(d_0) = \{d : \|d - d_0\|_\infty \leq h\} \subset \mathcal{D}$ and that for all $d \in N_h(d_0)$:*

$$0 < \underline{f} \leq f_D(d \mid Z) \leq \bar{f}, \quad \|\partial f_D(d \mid Z)\|_1 \leq \bar{f}', \quad t(d, Z) \leq \bar{t}, \quad |\mathrm{div_d} t(d, Z)| \leq \bar{t}' \ a.s.,$$

$$\mathrm{E}(t^2(d, X)|D = d) \geq \underline{t}^2 \text{ for the case (a), } \mathrm{E}((\mathrm{E}[t(X) \mid X_1])^2|D = d) \geq \underline{t}^2 \text{ for the case (b).}$$

*Then the finite-sample bounds stated in the proof of this lemma hold. In particular, if $h \searrow 0$ and $(\underline{c}, \bar{c}, \underline{t}, \bar{t}, \bar{t}', \underline{f}, \overline{f}, \bar{f}')$ are bounded away from zero and bounded above, then*

$$\kappa/\sigma \lesssim h^{-p_1/6} \lesssim \sigma \asymp L \asymp h^{-p_1/2-1} \to \infty.$$

We next characterize the bias of approximating the perfectly localized parameter.

**Lemma 10** (**Structure of Bias in Perfect Localization**). *Suppose that for some $h_0 > 0$, $d \mapsto m(d) = \mathrm{E}[m(W, \gamma_0^\star) \mid D = d]$ and $d \mapsto f_D(d)$ are continuously differentiable on $N_{h_0}(d_0)$ to the integer order $\mathsf{sm}$, and for $\mathsf{v} := \mathsf{sm} \wedge \mathsf{o}$ and $\partial_d^{\mathsf{v}}$ denoting the tensor $\partial^{\mathsf{v}}/(\partial d)^{\mathsf{v}}$ we have*

$$\sup_{d \in N_{h_0}(d_0)} \|\partial_d^{\mathsf{v}}(m(d)f_D(d))\|_{op} \leq \bar{g}_{\mathsf{v}}, \quad \sup_{d \in N_{h_0}(d_0)} \|\partial_d^{\mathsf{v}} f_D(d)\|_{op} \leq \bar{f}_{\mathsf{v}}, \quad \inf_{d \in N_{h_0}(d_0)} f_D(d) \geq \underline{f}.$$

*We have that for all $h < h_1 \leq h_0$,*

$$|\theta(\gamma_0^\star; \ell_h) - \theta(\gamma_0^\star; \ell_0)| \leq Ch^{\mathsf{v}},$$

*where the constant $C$ and $h_1$ depend only on $K, \mathsf{v}, \bar{g}_{\mathsf{v}}, \bar{f}_{\mathsf{v}}, \underline{f}$. If the latter constants are bounded away from above and zero, as $h \searrow 0$, we have $|\theta(\gamma_0^\star; \ell_h) - \theta(\gamma_0^\star; \ell_0)| \lesssim h^{\mathsf{v}}$.*

## 5. Practical Implementation Details and Examples

5.1. **Practical Implementation Details.** In practice we use the following generic algorithm for computing RMD estimators over subsamples $A$. In particular, for regression we set $m(W, b) = Yb(X)$.

(1) Obtain initial estimate $\hat{t}$ using a low-dimensional sub-dictionary $b_0$ of $b$:

$$\hat{t} \leftarrow (\hat{t}_0', 0')'; \ \hat{t}_0 = \hat{G}^{-1}\hat{M}_0; \ \hat{M}_0 \leftarrow \mathbb{E}_A m(W, b_0); \ \hat{G}_0 \leftarrow \mathbb{E}_A b_0 b_0;$$

Compute the empirical moments for the full dictionary:

$$\hat{M} \leftarrow \mathbb{E}_A m(W, b); \quad \hat{G} \leftarrow \mathbb{E}_A bb'.$$

(2) Update the diagonal normalization matrix:

$$\hat{D}^2 \leftarrow \mathrm{diag}\left(\mathbb{E}_A[\{b(X)b(X)'\hat{t} - m(W, b)\}_j^2]; \quad j = 1, ..., p\right).$$

| Income Quintile | N treated | N untreated | ATE | SE |
|---|---|---|---|---|
| All | 3682 | 6193 | 7897.41 | 1228.72 |
| 1 | 272 | 1703 | 4270.77 | 910.28 |
| 2 | 527 | 1448 | 1405.03 | 1578.36 |
| 3 | 753 | 1222 | 5085.46 | 1186.21 |
| 4 | 962 | 1013 | 8904.84 | 2061.85 |
| 5 | 1168 | 807 | 18987.16 | 5790.07 |

TABLE 1. Average Treatment Effect of 401K Eligibility on Net Financial Assets. Local Average Treatment Effects are Reported by Income Quintile Groups.

(3) Update the RMD estimate, using current estimate as the starting point in the algorithm:

$$\hat{t} \leftarrow \arg\min \|t\|_1 : \|\hat{D}^{-1}(\hat{M} - \hat{G}t)\|_\infty \leq \lambda; \quad \lambda = c\Phi^{-1}(1 - \mathsf{a}/2p)/\sqrt{n},$$

(4) Iterate on steps 2 and 3 several times. Return the final estimate $\hat{t}$.

We note the following. First, theoretical arguments similar to Belloni et al suggest that the data-driven algorithm behaves as the algorithm that knows the ideal $D$, since iterations yield $\|D\hat{D}^{-1} - I\|_\infty \to_{\mathrm{P}} 0$. The argument works provided we can set $c > 1.1$ . In practice, however, $c = 1$ works just fine from the outset. We set $\mathsf{a}$ small, e.g. $\mathsf{a} = .05$.

Second, Chernozhukov et al. [2013] discuss finer data-driven choices of penalty levels based on the Gaussian or empirical bootstraps:

$$\lambda = (1 - \alpha) - \text{quantile}(\|\hat{D}^{-1}(\hat{M}^* + \hat{G}^* t)\|_\infty \mid (W_i)_{i \in I_k^c}),$$

where $\hat{M}^*$ and $\hat{G}^*$ are bootstrap copies of $\hat{M}$ and $\hat{G}$. This method yields an even lower theoretically valid penalty levels, because they adapt to the correlation structure much better. For instance, for highly-correlated empirical moments, the penalty level produced by this method can be substantially lower than the simple plug-in choice made above (in the extreme case, where the moments are perfectly correlated, the penalty level of Chernozhukov et al. [2013] approximates $\Phi^{-1}(1 - \mathsf{a}/2)$).

5.2. **Global and Local Effects of 401(k) Eligibility on Net Financial Assets.** First, we use DMLR 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)? We follow the identification strategy of Poterba and Venti [1994], 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.

We use data from the 1991 US Survey of Income and Program Participation, using sample selection and variable construction as in Abadie [2003], Chernozhukov and Hansen [2004]. The outcome $Y$ is net financial assets defined as the sum of IRA balances, 401(k) balances,

| Income Quintile | N | APE | SE |
|---|---|---|---|
| all | 5001 | -0.21 | 0.25 |
| 1 | 1001 | -0.12 | 0.06 |
| 2 | 1000 | 0.03 | 0.07 |
| 3 | 1000 | 0.11 | 0.08 |
| 4 | 1000 | -0.03 | 0.09 |
| 5 | 1000 | 0.11 | 0.25 |

TABLE 2. Estimated Average Derivative (Price Elasticity) of Gasoline demand. Local Average Derivatives Reported by Income Quintile

checking accounts, US saving bonds, other interest-earning accounts, stocks, mutual funds, and other interest-earning assets minus non-mortgage 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 $n = 9915$ 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.

Table 1 summarizes results for the entire population and for each quintile of the income distribution. We use $L = 5$ folds in cross-fitting. We find ATE of 7897 (1229) by DMLR, which directly estimates the RR in stage 1. 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, the lower standard error of DMLR reflects numerical stability conferred by avoiding the estimated probability in the denominator. We find that ATE is not statistically significant for the second quintile, and it is statistically significant, positive, and heterogenous for the other quintiles. The results are broadly consistent with Poterba and Venti [1994], Poterba et al. [1995], who use a simpler specification motivated by economic reasoning.

5.3. **Global and Local Price Elasticity of Gasoline Demand.** Second, we use DMLR to estimate the average price elasticity (APE) of household gasoline 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], 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.

We use data from the 1994-1996 Canadian National Private Vehicle Use Survey, using sample selection and variable construction as in Yatchew and No [2001], Belloni et al. [2019]. The outcome $Y$ is log gasoline consumption. The variable $D$ with respect to which we differentiate is log price per liter. 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 $b(D, X)$ of Chernozhukov and Semenova [2017] with $p = 91$ including polynomials of continuous covariates and interactions of log price (and its square) with time and household composition dummies.

Table 2 summarizes results for the entire population and for each quintile of the income distribution. We use $L = 5$ folds in cross-fitting. We find average price elasticity of $-0.21$ (0.25) by DMLR. For comparison, OLS regression of log demand on log age, log income, and log distance as well as geographical, time, and household composition dummies with yields an estimate of 0.14 (0.06). The linear specification leads to a positive elasticity estimate, contradicting economic intuition. We find that average price elasticity is statistically significant and negative for the first quintile, and it is otherwise statistically insignificant. The results are broadly consistent with Chernozhukov and Semenova [2017], who more explicitly consider the relationship between average price elasticity and income.

# Appendix A.

A.1. **Notation Glossary.** Let $W = (Y, X')'$ be a random vector with law $P$ on the sample space $\mathcal{W}$, and $W_1^n = (Y_i, X_i)_{i=1}^n$ denote i.i.d. copies of $W$. The law of $X$ is denoted by $F$. All models and probability measure $P$ can be indexed by $n$, the sample size, so that the models and their dimensions and parameters determined by $P$ change with $n$. We use notation from the empirical process theory, see Van Der Vaart and Wellner [1996]. Let $\mathbb{E}_{I_k} f$ denote the empirical average of $f(W_i)$ over $i \in I \subset \{1, ..., n\}$: $\mathbb{E}_{I_k} f := \mathbb{E}_{I_k} f(W) = |I|^{-1} \sum_{i \in I} f(W_i)$. Let $\mathbb{G}_I$ denote the empirical process over $f \in \mathcal{F} : \mathcal{W} \to \mathbb{R}^p$ and $i \in I$, namely $\mathbb{G}_I f := \mathbb{G}_I f(W) := |I|^{-1/2} \sum_{i \in I} (f(W_i) - Pf)$, where $Pf := Pf(W) := \int f(w) dP(w)$. 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 $\|f\|_{P,q} = \|f(W)\|_{P,q}$. We use $\|\cdot\|_q$ to denote $\ell_q$ norm on $\mathbb{R}^d$. For a differentiable map $x \mapsto f(x)$, from $\mathbb{R}^d$ to $\mathbb{R}^k$, we use $\partial_{x'} f(x)$ to abbreviate the partial derivatives $(\partial/\partial x') f(x)$, and we use $\partial_{x'} f(x_0)$ to mean $\partial_{x'} f(x) \mid_{x=x_0}$, etc. We use $x'$ to denote the transpose of a column vector $x$. We use $\text{div}_d$ to denote the divergence of scalar function: $\text{div}_d\ g = \sum_{j=1}^{\dim(d)} \partial_{d_j} g(d)$. We say that $a \lesssim b$ under the asymptotics with an index $n \to \infty$ if $a \leq Cb$ for all $n$ sufficiently large, and $a \asymp b$ if both $a \lesssim Cb$ and $b \lesssim Ca$ for all $n$ sufficiently large, where $C \geq 1$ is a positive constant that does not depend on $n$.

A.2. **Few Preliminaries.** To prove the first couple of lemmas we recall the following definitions and results. Given two normed vector spaces $V$ and $W$ over the field of real numbers $\mathbb{R}$, a linear map $A : V \to W$ is continuous if and only if

$$\|A\|_{op} := \inf\{c \geq 0 : \|Av\| \leq c\|v\| \text{ for all } v \in V\} < \infty,$$

where $\|\cdot\|_{op}$ is the operator norm. The operator norm depends on the choice of norms for the normed vector spaces $V$ and $W$. A Hilbert space is a complete linear space equipped with an inner product $\langle f, g\rangle$ and the norm $|\langle f, f\rangle|^{1/2}$. The space $L^2(P)$ is the Hilbert space with the inner product $\langle f, g\rangle = \int fg dP$ and norm $\|f\|_{P,2}$. The closed linear subspaces of $L^2(P)$ equipped with the same inner product and norm are Hilbert spaces.

**Hahn-Banach Extension for Normed Vector Spaces.** If $V$ is a normed vector space with linear subspace $U$ (not necessarily closed) and if $\phi : U \mapsto K$ is continuous and linear, then there exists an extension $\psi : V \mapsto K$ of $\phi$ which is also continuous and linear and which has the same operator norm as $\phi$.

**Riesz-Frechet Representation Theorem.** Let $H$ be a Hilbert space over $\mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle$, and $T$ a bounded linear functional mapping $H$ to $\mathbb{R}$. If $T$ is bounded then there exists a unique $g \in H$ such that for every $f \in H$ we have $T(f) = \langle f, g\rangle$. It is given by $g = z(Tz)$, where $z$ is unit-norm element of the orthogonal complement of the kernel subspace $K = \{a \in H : Ta = 0\}$. Moreover, $\|T\| = \|g\|$, where $\|T\|$ denotes the operator norm of $T$, while $\|g\|$ denotes the Hilbert space norm of $g$.

**Radon-Nykodym Derivative.** Consider a measure space $(\mathcal{X}, \Sigma)$ on which two $\sigma$-finite measure are defined, $\mu$ and $\nu$. If $\nu \ll \mu$ (i.e. $\nu$ is absolutely continuous with respect to $\mu$), then there is a measurable function $f : \mathcal{X} \to [0, \infty)$, such that for any measurable set $A \subseteq \mathcal{X}$, $\nu(A) = \int_A f\, d\mu$. The function $f$ is conventionally denoted by $d\nu/d\mu$.

**Integration by Parts.** Consider a closed measurable subset $\mathcal{X}$ of $\mathbb{R}^k$ equipped with Lebesgue measure $V$ and piecewise smooth boundary $\partial\mathcal{X}$, and suppose that $v : \mathcal{X} \to \mathbb{R}^k$ and $\phi : \mathcal{X} \to \mathbb{R}$ are both $C^1(\mathcal{X})$, then

$$\int_{\mathcal{X}} \varphi \operatorname{div} v \, dV = \int_{\partial\mathcal{X}} \varphi \, v' dS - \int_{\mathcal{X}} v' \operatorname{grad} \varphi \, dV,$$

where $S$ is the surface measure induced by $V$.

## Appendix B. Proofs for Section 2

B.1. **Proof of Lemma 1.** We note that $\Gamma = \operatorname{span}(\Gamma_0)$ is a linear subspace of $L^2(F)$, and $\bar{\Gamma}$ is a closed subspace by definition. Therefore, $\bar{\Gamma}$ is a Hilbert space with norm $g \mapsto \|g\|_{P,2}$ and inner product $(f, g) \mapsto \langle f, g\rangle = \int f g dF$.

To show claim (i), we note that by the Hahn-Banach extension theorem, the operator $\theta : \Gamma \to \mathbb{R}$ can be extended to $\tilde{\theta} : \bar{\Gamma} \to \mathbb{R}$ such that $\|\tilde{\theta}\|_{op} = \|\theta\|_{op}$. By the Riesz-Frechet theorem there exists a unique representer $\alpha_0^\star$ such that $\tilde{\theta}(\gamma) = \langle \gamma, \alpha_0^\star\rangle$ on $\gamma \in \bar{\Gamma}$ and $\|\tilde{\theta}\|_{op} = \|\alpha_0^\star\|_{P,2}$.

To show claim (ii), we are given a linear representer $\alpha_0$. Denote by $\alpha_0^\star$ the projection of $\alpha_0$ onto $\bar{\Gamma}$. Then $\gamma \mapsto \varphi(\gamma) := \langle \gamma, \alpha_0\rangle = \langle \gamma, \alpha_0^\star\rangle$ agrees with $\gamma \mapsto \theta(\gamma)$ on $\gamma \in \Gamma$. Extend $\varphi$ to $\bar{\Gamma}$ by defining $\varphi(\gamma) = \langle \gamma, \alpha_0^\star\rangle$ for $\gamma \in \bar{\Gamma} \setminus \Gamma$, which is well-defined by Cauchy-Schwarz inequality. Then $\|\varphi\|_{op} = \|\alpha_0^\star\|_{P,2} \le \|\alpha_0\|_{P,2} < \infty$, since projection reduces norm. Further,

$$\infty > \|\alpha_0^\star\|_{P,2} \;=\; \sup_{\gamma \in \bar{\Gamma}\setminus\{0\}} |\langle \gamma, \alpha_0^\star\rangle| / \|\gamma\|_{P,2} = \sup_{\gamma \in \bar{\Gamma}\setminus\{0\}} |\tilde{\theta}(\gamma)| / \|\gamma\|_{P,2} = \|\tilde{\theta}\|_{op}.$$

Hence $\alpha_0^\star$ is a representer for the extension $\tilde{\theta}$, and the Riesz-Frechet theorem implies that $\alpha_0^\star$ is unique. $\square$

B.2. **Proof of Lemma 2.** Use the same notation as in the proof of the previous lemma. In all examples, $\alpha_0 \in L^2(F)$ and $\gamma \in L^2(F)$ imply that $|\langle \alpha_0, \gamma\rangle| < \|\alpha_0\|_{P,2}\|\gamma\|_{P,2} < \infty$.

Proof of claim (i). In Example 1, since $dF(x) = \sum_{k=0}^1 P[D = k|Z = z] 1(k = d) dF(z)$ by the Bayes rule, we have

$$\langle \alpha_0, \gamma\rangle = \int \gamma(d, z)\ell(x) \frac{1(d = 1) - 1(d = 0)}{P[D = d|Z = z]} dF(x) = \theta(\gamma).$$

In Example 2, $\ell\alpha_0 \in L^2(F)$ means that the Radon-Nykodym derivatives $\frac{dF_1}{dF}$ and $\frac{dF_0}{dF}$ exist on the support of $\ell$, so that

$$\langle \alpha_0, \gamma\rangle = \int \gamma \ell \left( \frac{dF_1}{dF} - \frac{dF_0}{dF} \right) dF = \int \gamma \ell (dF_1 - dF_0) = \theta(\gamma).$$

We can demonstrate the claim for Example 3 similarly to Example 2.

In Example 4, we can write

$$\langle \alpha_0, \gamma\rangle = -\int\int \gamma(x) \frac{\operatorname{div}_d(\ell(x)t(x)f(d|z))}{f(d|z)} f(d|z) \mathrm{d}d dF(z)$$

$$= \int \int \partial_d \gamma(x)' t(x) \ell(x) f(d|z) \mathrm{d}d \mathrm{d}F(z) = \theta(\gamma),$$

where we used the integration by parts and that $\gamma(x)\ell(x)t(x)f(d|z)$ vanishes on the boundary of $\mathcal{D}_z$. The rest of the claim is immediate from Lemma 1.

Proof of claim (ii). We can refer to the case of linear regression discussed in Section 2.

In what follows consider the case of $G > 0$ and $\ell = 1$.

In Example 1, $M = \mathrm{E}(b(1, Z) - b(0, Z))$. Suppose $P[D = 0|Z] \in \{0, 1\}$ with probability in $[\pi, 1 - \pi]$ for $\pi > 0$, but such that $G > 0$ (this puts restrictions on $b$). This is known as the case of failing overlap assumption in causal inference. Then $\alpha_0(X)$ is $\mathsf{na}$ with probability $\pi$.

In Example 2 and 3, $M = \int b(dF_1 - dF_0)$ is well defined, but $\alpha_0(X) = \mathsf{na}$ whenever $dF_1/dF$ and $dF_1/dF$ do not exist. For instance, $F_1$ and $F_0$ can have point masses, where $F$ does not, while retaining the same support as $F$.

In Example 4, take basis functions $b$ and a constant direction $t(X) = 1$, such that $M = \mathrm{E}\partial_d b(D, Z)$ is well defined. Consider the case where $f(d|Z) = 0$ with positive probability so that $\alpha_0(X) = \mathsf{na}$ with this probability. □

B.3. **Proof of Lemma 3.** The projection operator onto $\bar{\Gamma}_1 = L^2(F_1)$ is the conditional expectation with conditioning on $X_1$. The contractive property follows from Jensen's inequality. □

## Appendix C. Proofs for Section 3

C.1. **Proof of Theorem 1.** The proof uses empirical process notation: $\mathbb{G}_I$ denotes the empirical process over $f \in \mathcal{F} : \mathcal{W} \to \mathbb{R}^p$ and $I \subset \{1, ..., n\}$, namely

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

**Step 1.** We have a random partition $(I_k, I_k^c)$ of $\{1, ..., n\}$ into sets of size $m = n/K$ and $n - n/K$. Let

$$\bar{\theta}_k = \theta_0 - \mathbb{E}_{I_k} \psi_0(W).$$

Observe that in Lemma 4, derivatives don't depend on $\theta$. Hence for all $\theta$,

$$\partial_\beta \psi(W, \theta; \beta_0, \rho_0) = -m(W, b) + {\rho_0}' b(X) b(X) =: \partial_\beta \psi_0(W)$$

$$\partial_\rho \psi(W, \theta; \beta_0, \rho_0) = -b(X)(Y - b(X)'\beta_0) =: \partial_\rho \psi_0(W)$$

$$\partial^2_{\beta\rho'} \psi(X, \theta; \beta_0, \rho_0) = b(X) b(X)' =: \partial^2_{\beta\rho'} \psi_0(W),$$

where $\psi_0(W) := \psi(W, \theta_0; \beta_0, \rho_0)$ as before.

Define the estimation errors $u := \hat{\beta}_k - \beta_0$ and $v := \hat{\rho}_k - \rho_0$. Using Lemma 4, we have by the exact Taylor expansion around $(\beta_0, \rho_0)$

$$\hat{\theta}_k = \bar{\theta}_k - (\mathbb{E}_{I_k}\partial_\beta\psi_0(W))'u - (\mathbb{E}_{I_k}\partial_\rho\psi_0(W))'v - u'(\mathbb{E}_{I_k}\partial^2_{\beta\rho'}\psi_0(W))v.$$

Consider the event $\mathcal{E}$ that Condition R holds. On this event:

$$\begin{aligned}(\sqrt{m}/\sigma)(\hat{\theta}_k - \bar{\theta}_k) \quad &= \quad \mathrm{rem}_k := \sum_{j=1}^4 \mathrm{rem}_{jk} := -\sigma^{-1}[\mathbb{G}_n\partial_\beta\psi_0(W)]'u \\ &\quad -\sigma^{-1}[\mathbb{G}_n\partial_\rho\psi_0(W)]'v - \sigma^{-1}u'[\mathbb{G}_n\partial^2_{\beta\rho'}]v - \sigma^{-1}\sqrt{m}u'[P\partial^2_{\beta\rho'}\psi_0(W)]v,\end{aligned}$$

where we have used that by Lemma 4

$$P\partial_\beta\psi_0(W)'u = 0, \quad P\partial_\rho\psi_0(W)'v = 0.$$

We now bound $\mathrm{E}[\mathrm{rem}_k^2 1(\mathcal{E})]$ by analyzing each of its terms. By the law of iterated expectations

$$\mathrm{E}[\mathrm{rem}_k^2 1(\mathcal{E})] = \mathrm{E}[\mathrm{E}[\mathrm{rem}_k^2 1(\mathcal{E})|(W_i)_{i\in I_k^c}]] \le 4\sum_{j=1}^4 \mathrm{E}[\mathrm{E}[\mathrm{rem}_{jk}^2 1(\mathcal{E})|(W_i)_{i\in I_k^c}]]$$

using the fact that $\mathrm{E}\left(\sum_{j=1}^J V_j\right)^2 \le J\sum_{j=1}^J \mathrm{E}V_j^2$ for arbitrary random variables $(V_j)_{j=1}^J$.

Note that $u$ and $v$ are fixed once we condition on the observations $(W_i)_{i\in I_k^c}$. On the event $\mathcal{E}$, by condition R, $\mathrm{rem}_{1k}, \mathrm{rem}_{2k}$ and $\mathrm{rem}_{3k}$ have conditional mean 0 and conditional variance given by

$$\begin{aligned}\sigma^{-1}\sqrt{Var}[\mathrm{rem}_{1k} \mid (W_i)_{i\in I_k^c}] \quad &= \quad \sigma^{-1}\sqrt{Var}[(\partial_\beta\psi_0(W)'u) \mid (W_i)_{i\in I_k^c}] \\ &\le \quad \sigma^{-1}\mu\sigma\sqrt{u'Gu} = \sigma^{-1}\mu\sigma r_1 \le \delta, \\ \sigma^{-1}\sqrt{Var}[\mathrm{rem}_{2k} \mid (W_i)_{i\in I_k^c}] \quad &= \quad \sigma^{-1}\sqrt{Var}[(\partial_\rho\psi_0(W)'v) \mid (W_i)_{i\in I_k^c}] \\ &\le \quad \sigma^{-1}\mu\sqrt{v'Gv} = \sigma^{-1}\mu\sigma r_2 \le \delta, \\ \sigma^{-1}\sqrt{Var}[\mathrm{rem}_{3k} \mid (W_i)_{i\in I_k^c}] \quad &= \quad \sigma^{-1}\sqrt{Var}[u'b(X)b(X)'v \mid (W_i)_{i\in I_k^c}] \\ &\le \quad \sigma^{-1}\mu(\sqrt{v'Gv} + \sqrt{u'Gu}) \\ &\le \quad \sigma^{-1}\mu(\sigma r_2 + r_1) \le \delta.\end{aligned}$$

On the event $\mathcal{E}$, $\mathrm{rem}_{4k}$ has conditional mean and conditional variance given by

$$|\sigma^{-1}\sqrt{m}u'[P\partial^2_{\beta\rho'}\psi_0(W)]v| \le \sigma^{-1}\sqrt{m}\sigma r_3 \le \delta, \sqrt{Var}[\mathrm{rem}_{4k} \mid (W_i)_{i\in I_k^c}] = 0.$$

In summary,

$$\mathrm{E}[\mathrm{rem}_k^2 1(\mathcal{E})] \le 4[\delta^2 + \delta^2 + \delta^2 + \delta^2] = 16\delta^2.$$

**Step 2.** Here we bound the difference between $\hat{\theta} = K^{-1}\sum_{k=1}^K \hat{\theta}_k$ and $\bar{\theta} = K^{-1}\sum_{k=1}^K \bar{\theta}_k$:

$$\sqrt{n}/\sigma|\hat{\theta} - \bar{\theta}| \le \frac{\sqrt{n}}{\sqrt{m}}\frac{1}{K}\sum_{k=1}^K \sqrt{m/\sigma}|\hat{\theta}_k - \bar{\theta}_k| \le \frac{\sqrt{n}}{\sqrt{m}}\frac{1}{K}\sum_{k=1}^K \mathrm{rem}_k.$$

By Markov inequality we have

$$\begin{aligned}
&\mathrm{P}\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{rem}_k > 4\delta/\Delta\right) \le \mathrm{P}\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{rem}_k > 4\delta/\Delta \cap \mathcal{E}\right) + \mathrm{P}\left(\mathcal{E}^c\right) \\
&\le K^{-2}\mathrm{E}\left(\left(\sum_{k=1}^{K}\mathrm{rem}_k\right)^2 1(\mathcal{E})\right)\Delta^2/(16\delta^2) + \epsilon \\
&\le K^{-2}K^2 \max_k \mathrm{E}(\mathrm{rem}_k^2 1(\mathcal{E}))\Delta^2/(16\delta^2) + \epsilon \le \Delta^2 + \epsilon.
\end{aligned}$$

And we have that $\sqrt{n/m} = \sqrt{K}$. So it follows that

$$|\sqrt{n}(\hat{\theta} - \bar{\theta})/\sigma| \le \mathsf{err} = 4\sqrt{K}\delta/\Delta$$

with probability at least $1 - \Pi$ for $\Pi := \Delta^2 + \epsilon$.

**Step 3**. To show the second claim, let $Z := \sqrt{n}(\bar{\theta} - \theta_0)/\sigma$. By the Berry-Esseen bound, for some absolute constant $A$,

$$\sup_{z\in\mathbb{R}} |\mathrm{P}(Z \le z) - \Phi(z)| \le A\|\psi_0/\sigma\|_{P,3}^3 n^{-1/2} = A(\kappa/\sigma)^3 n^{-1/2}.$$

The current best estimate of $A$ is 0.4748, due to Shevtsova [2011]. Hence, using Step 2, for any $z \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathrm{P}(\sqrt{n}(\hat{\theta} - \theta_0)/\sigma \le z) - \Phi(z) = \mathrm{P}(\sqrt{n}(\hat{\theta} - \bar{\theta})/\sigma + Z \le z) - \Phi(z) \\
&= \mathrm{P}(Z \le z + \sqrt{n}(\bar{\theta} - \hat{\theta})/\sigma) - \Phi(z) \le \mathrm{P}(Z \le z + \mathsf{err}) + \Pi - \Phi(z) \\
&= \mathrm{P}(Z \le z + \mathsf{err}) - \Phi(z + \mathsf{err}) + \Phi(z + \mathsf{err}) - \Phi(z) + \Pi \\
&\le A(\kappa/\sigma)^3 n^{-1/2} + \mathsf{err}/\sqrt{2\pi} + \Pi,
\end{aligned}$$

where $1/\sqrt{2\pi}$ is the upper bound on the derivative of $\Phi$. Similarly, conclude that

$$\mathrm{P}(\sqrt{n}\sigma^{-1}(\hat{\theta} - \theta_0) \le z) - \Phi(z) \ge A(\kappa/\sigma)^3 n^{-1/2} - \mathsf{err}/\sqrt{2\pi} - \Pi.$$

The result follows by noting that $4/\sqrt{2\pi} = 1.5957... < 2$. $\square$

**Proof of Theorem 2.** From Van der Vaart [2000], Theorem 25.20, it suffices to exhibit a class of one dimensional parametric submodels that are regular and for whom the tangent set is all random variables with mean zero and finite variance. Suppose that $W_i$ had pdf $f_0$ under $P$ with respect to some measure $\mu$. Consider a parametric submodel (i.e. path) of the form

$$f_\tau(w) = f_0(w)\left[1 + \tau\delta(w)\right], \ \mathrm{E}\left[\delta(W)\right] = 0, \ \delta(W) \text{ bounded.}$$

Note that the score of $f_\tau(w)$ is $\delta(W)$. Because $\delta(W)$ is bounded $f_\tau(w)$ and $f_0(w)$ dominate each other so that $\bar{\Gamma}$ does not depend on $\tau$. Let $\gamma_\tau$ denote least squares projection of $Y$ on $\bar{\Gamma}$ under $f_\tau$. Then

$$\mathrm{E}\left[\gamma_\tau(X)^2\right] \le C\mathrm{E}_\tau\left[\gamma_\tau(X)^2\right] \le C\mathrm{E}_\tau\left[Y^2\right] \le C\mathrm{E}\left[Y^2\right] = C.$$

Note that by $\gamma_\tau, \gamma_0 \in \bar{\Gamma}$ and the previous inequality, as $\tau \longrightarrow 0$

$$\begin{aligned}
&\mathrm{E}\left[\gamma_\tau(X)\gamma_0(X)\right] = \mathrm{E}_\tau\left[\gamma_\tau(X)\gamma_0(X)\right] + o(1) \\
&= \mathrm{E}_\tau\left[Y\gamma_0(X)\right] + o(1) = \mathrm{E}\left[Y\gamma_0(X)\right] + o(1) = \mathrm{E}\left[\gamma_0(X)^2\right] + o(1).
\end{aligned}$$

Similarly we have

$$\mathrm{E}\left[\gamma_\tau(X)^2\right] = \mathrm{E}_\tau\left[\gamma_\tau(X)^2\right] + o(1) = \mathrm{E}_\tau\left[Y\gamma_\tau(X)\right] + o(1)$$

$$= \mathrm{E}\left[Y\gamma_\tau(X)\right] + o(1) = \mathrm{E}\left[\gamma_0(X)\gamma_\tau(X)\right] + o(1) \longrightarrow \mathrm{E}[\gamma_0(X)^2].$$

Therefore it follows that

$$\mathrm{E}\left[\{\gamma_\tau(X) - \gamma_0(X)\}^2\right] = \mathrm{E}\left[\gamma_\tau(X)^2\right] + \mathrm{E}\left[\gamma_0(X)^2\right] - 2\mathrm{E}\left[\gamma_\tau(X)\gamma_0(X)\right] \longrightarrow 0.$$

Note that $|\mathrm{E}\left[\alpha_0(X)\{\gamma_\tau(X) - \gamma(X)\}\delta(W)\right]| \leq C\mathrm{E}\left[|\alpha_0(X)|\,|\gamma_\tau(X) - \gamma_0(X)|\right] \longrightarrow 0$ so that

$$\begin{aligned}
\mathrm{E}\left[m(W,\gamma_\tau)\right] - \mathrm{E}\left[m(W,\gamma_0)\right] &= \mathrm{E}\left[\alpha_0(X)\{\gamma_\tau(X) - \gamma_0(X)\}\right] \\
&= \mathrm{E}_\tau\left[\alpha_0(X)\{\gamma_\tau(X) - \gamma_0(X)\}\right] \\
&\quad - \tau\mathrm{E}\left[\alpha_0(X)\{\gamma_\tau(X) - \gamma_0(X)\}\delta(W)\right] \\
&= \mathrm{E}_\tau\left[\alpha_0(X)\{Y - \gamma_0(X)\}\right] + o(\tau) \\
&= \mathrm{E}_\tau\left[\alpha_0(X)\{Y - \gamma_0(X)\}\right] - \mathrm{E}\left[\alpha_0(X)\{Y - \gamma_0(X)\}\right] + o(\tau) \\
&= \tau\mathrm{E}\left[\alpha_0(X)\{Y - \gamma_0(X)\}\delta(W)\right] + o(\tau).
\end{aligned}$$

Therefore $\mathrm{E}\left[m(W,\gamma_\tau)\right]$ is differentiable at $\tau = 0$ with

$$\partial\mathrm{E}\left[m(W,\gamma_\tau)\right]/\partial\tau = \mathrm{E}\left[\alpha_0(X)\{Y - \gamma_0(X)\}\delta(W)\right].$$

In addition, by mean-square continuity of $m(W,\gamma)$,

$$\begin{aligned}
\mathrm{E}_\tau\left[m(W,\gamma_\tau)\right] - \mathrm{E}[m(W,\gamma_\tau)] &= \tau\mathrm{E}\left[m(W,\gamma_\tau)\delta(W)\right] \\
&= \tau\mathrm{E}\left[m(W,\gamma_0)\delta(W)\right] + \tau\mathrm{E}[\{m(W,\gamma_\tau) - m(W,\gamma_0\}\delta(W)] \\
&= \tau\mathrm{E}\left[m(W,\gamma_0)\delta(W)\right] + o(\tau).
\end{aligned}$$

It follows that $\mathrm{E}_\tau\left[m(W,\gamma_\tau)\right] - \mathrm{E}[m(W,\gamma_\tau)]$ is differentiable with

$$\frac{\partial\{\mathrm{E}_\tau\left[m(W,\gamma_\tau)\right] - \mathrm{E}[m(W,\gamma_\tau)]\}}{\partial\tau} = \mathrm{E}\left[m(W,\gamma_0)\delta(W)\right] = \mathrm{E}[\{m(W,\gamma_0) - \theta_0\}\delta(W)].$$

It then follows by the derivative of the sum being the sum of the derivatives that $\theta_\tau = \mathrm{E}_\tau[m(W,\gamma_\tau)]$ is differentiable at $\tau = 0$ and

$$\frac{\partial\theta_\tau}{\partial\tau} = \mathrm{E}[\psi_0(W)\delta(W)].$$

Next let $\mathcal{S} = \{s(W) : \mathrm{E}[s(W)] = 0,\ \mathrm{E}[s(W)^2] < \infty\}$. It is straightforward to show that $\mathcal{S}$ is the closed linear span of scores $\delta(W)$ that are bounded with mean zero. Then since $\psi_0(W) \in \mathcal{S}$ it follows that $\psi_0(W)$ is the projection of $\psi_0(W)$ on $\mathcal{S}$ and so is the efficient and only influence function, and the hypotheses of Theorem 25.20 of Van der Vaart [2000] are satsified. $\square$

**Proof of Lemma 5.** First, we note that

$$\|t_0^{\mathcal{M}}\|_0 = |\mathcal{M}| \leq s := \max\{x : Ax^{-a} \geq \nu\} = (A/\nu)^{1/a}.$$

Define

$$t^r := t_0 - t_0^{\mathcal{M}} = t_0 1(|t_0| \leq \nu).$$

Note that

$$\|t^r\|_1 \leq \nu s + \int_s^\infty Ax^{-a}dx = \nu s + \frac{1}{1-a}As^{-a+1} = \nu s + \frac{1}{1-a}\nu s = \frac{a}{a-1}\nu s.$$

Then $\delta \in S(t_0, \nu)$ implies that, by the repeated use of the triangle inequality:

$$\begin{aligned}
&\|t_0 + \delta\|_1 \leq \|t_0\|_1 \iff \|t_0^{\mathcal{M}} + \delta_{\mathcal{M}}\|_1 + \|t_0^r + \delta_{\mathcal{M}^c}\|_1 \leq \|t_0^{\mathcal{M}}\|_1 + \|t_0^r\|_1 \\
&\implies \|\delta_{\mathcal{M}^c}\|_1 - \|t_0^r\|_1 \leq \|t_0^r + \delta_{\mathcal{M}^c}\|_1 \leq \|t_0^{\mathcal{M}}\|_1 - \|t_0^{\mathcal{M}} + \delta_{\mathcal{M}}\|_1 + \|t_0^r\|_1 \\
&\implies \|\delta_{\mathcal{M}^c}\|_1 - \|t_0^r\|_1 \leq \|\delta_{\mathcal{M}}\|_1 + \|t_0^r\|_1 \implies \|\delta_{\mathcal{M}^c}\|_1 \leq \|\delta_{\mathcal{M}}\|_1 + 2\|t_0^r\|_1.
\end{aligned}$$

If $2\|t^r\|_1 \leq \|\delta_{\mathcal{M}}\|_1$, we have that $\|\delta_{\mathcal{M}^c}\|_1 \leq 2\|\delta_{\mathcal{M}}\|_1$, so using the definition of the cone invertibility factor we obtain

$$(k/s)\|\delta\|_1 \leq \|G\delta\|_\infty \leq \nu \implies \delta' G\delta \leq \|\delta\|_1 \|G\delta\|_\infty \leq (s/k)\nu^2.$$

If $2\|t^r\|_1 \geq \|\delta_{\mathcal{M}}\|_1$, then $\|\delta\|_1 \leq 6\|t^r\|_1$

$$\delta' G\delta \leq \|\delta\|_1 \|G\delta\|_\infty \leq 6\|t^r\|_1 \nu \leq 6\frac{a}{a-1}s\nu^2. \qquad \square$$

C.2. **Proof of Lemma 6.** Consider the event $\mathcal{R}$ such that

$$\|\hat{g}(t_0)\|_\infty \leq \lambda, \qquad \|\hat{g}(\hat{t})\|_\infty \leq \lambda, \tag{18}$$

holds. This event holds with probability at least $1-\epsilon$. The event $\mathcal{R}$ implies that $\|\hat{t}\|_1 \leq \|t_0\|_1$ by definition of $\hat{t}$, which further implies that for $\delta = \hat{t} - t_0$

$$\begin{aligned}
\|G\delta\|_\infty &\leq \|(G - \hat{G})\delta\|_\infty + \|\hat{G}\delta\|_\infty \\
&= \|(G - \hat{G})\delta\|_\infty + \|\hat{g}(\hat{t}) - \hat{g}(t_0)\|_\infty \\
&\leq \|(G - \hat{G})\|_\infty \|\delta\|_1 + \|\hat{g}(\hat{t})\|_\infty + \|\hat{g}(t_0)\|_\infty \\
&\leq \bar{\lambda} 2B + 2\lambda \leq \bar{\nu}.
\end{aligned}$$

Hence $\delta \in S(t_0, \nu)$ with probability $1 - \epsilon$.

The first inequality now in the bound follows from the definition of $s(t_0)$: $\sup_{\delta \in S(t_0,\nu)} \delta' G\delta \leq s(t_0)\nu^2$. The second bound follows by $\|\delta\|_1 \leq 2B$, $\delta' G\delta \leq \|G\delta\|_\infty \|\delta\|_1 \leq \nu 2B$. $\square$

**Proof of Theorem 3.** Application of Lemma 6 implies that with probability at least $1 - 4\epsilon$, estimation errors $\tilde{u} = D_\beta^{-1}(\hat{\beta}_A - \beta_0)$ and $\tilde{v} = D_\rho^{-1}(\hat{\rho}_A - \rho_0)$ obey

$$\tilde{u}' G\tilde{u} \leq C[(B^2\tilde{\ell}^2 s(D_\beta^{-1}\beta_0; \nu)/n) \wedge (B\tilde{\ell}/\sqrt{n})],$$

$$\tilde{v}' G\tilde{v} \leq C[(B^2\tilde{\ell}^2 s(D_\rho^{-1}\rho_0; \nu)/n) \wedge (B\tilde{\ell}/\sqrt{n})],$$

where $C$ is an absolute constant. Then

$$|u'Gu| \leq \mu_D^2 \tilde{u}' G \tilde{u}, \quad |v'Gv| \leq \mu_D^2 \sigma^2 \tilde{v}' G \tilde{v}.$$

The stated bounds then follow. Hence the guarantee $R(\delta)$ holds for $\varepsilon = 1 - K4\epsilon$ provided that for some large enough absolute $C$:

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

for $r_1$, $r_2$, and $r_3$ given in the theorem. □

## Appendix D. Proofs for Section 4

D.1. **Proof of Lemma 7.** The proof uses the fact that $m(W,\gamma) = m(X,\gamma)$, and that

$$\psi_0(W) = U_1 + \alpha_0^\star(X) U_2.$$

Since $\mathrm{E} U_1 U_2 \alpha_0^\star(X) = 0$ by the LIE, using the bounded moments assumption we have:

$$\sigma^2 = \mathrm{E}U_1^2 + \mathrm{E}U_2^2 \alpha_0^{\star 2} \geq \mathrm{E}[\mathrm{E}(U_2^2 \mid X)\alpha_0^{\star 2}(X)] \geq \underline{c}^2 L^2.$$

The bound from above follows similarly:

$$\sigma^2 = \mathrm{E}U_1^2 + \mathrm{E}U_2^2 \alpha_0^{\star 2} \leq \bar{c}^2 + \mathrm{E}[\mathrm{E}(U_2^2 \mid X)\alpha_0^{\star 2}(X)] \leq \bar{c}^2 + \bar{c}^2 L^2.$$

Using the triangle inequality and bounded moments assumptions, we have:

$$\begin{aligned}
\kappa \leq \|U_1\|_{P,3} + \|U_2 \alpha_0^\star\|_{P,3} &\leq \bar{c} + (\mathrm{E}(\mathrm{E}[|U_2|^3 \mid X]|\alpha_0^\star(X)|^3))^{1/3}, \\
&\leq \bar{c} + \bar{c}\|\alpha_0^\star\|_{P,3} \leq \bar{c}(1 + c(L^2 \vee 1)),
\end{aligned}$$

where the last line follows by assumption. □

D.2. **Proof of Lemma 8.** We shall use that $m(W,\gamma) = m(X,\gamma)$, and

$$\psi_0(W) = U_1 + \alpha_0^\star(X) U_2.$$

Then by $\mathrm{E} U_1 U_2 \alpha_0^\star(X) = 0$, holding by the LIE, we have

$$\sigma^2 = \mathrm{E}U_1^2 + \mathrm{E}U_2^2 \alpha_0^{\star 2} = \mathrm{E}U_1^2 + \mathrm{E}(\mathrm{E}[U_2^2 \mid X]\alpha_0^{\star 2}(X)).$$

Then using the moment assumptions, we have

$$\underline{c}^2 \|\alpha_0^\star\|_{P,2}^2 \leq \sigma^2 \leq \bar{c}^2(\|\ell\|_{P,2}^2 + \|\alpha_0^\star\|_{P,2}^2).$$

Using the triangle inequality, the LIE, and the bounded heteroscedasticity assumption, conclude

$$\bar{\kappa} \leq \|U_1\|_{P,3} + \|U_2\alpha_0^\star\|_{P,3} \leq \bar{c}(\|\ell\|_{P,3} + \|\alpha_0^\star\|_{P,3}).$$

For the case (a), $\alpha_0^\star(X) = \alpha_0(X,1)\ell(X)$, using the assumed bound $\underline{\alpha} \leq \alpha_0(X,1) \leq \bar{\alpha}$ conclude that

$$\underline{\alpha}\|\ell\|_{P,2} \leq L = \|\alpha_0^\star\|_{P,2} \leq \bar{\alpha}\|\ell\|_{P,2}, \quad \|\alpha_0^\star\|_{P,3} \leq \bar{\alpha}\|\ell\|_{P,3}.$$

For the case (b), $\alpha_0^\star(X_1) = \mathrm{E}[\alpha_0(X,1) \mid X_1]\ell(X_1)$, so that by Jensen's inequality

$$\|\alpha_0^\star\|_{P,q} \leq \|\alpha_0(X,1)\ell(X_1)\|_{P,q} \leq \bar{\alpha}\|\ell\|_{P,q}$$

and using

$$\underline{\alpha} \leq \mathrm{E}[\alpha_0(X,1) \mid X_1],$$

holding because conditional expectation preserves order, conclude that

$$\|\alpha_0^\star\|_{P,2}^2 = \mathrm{E}(\mathrm{E}[\alpha_0(X,1) \mid X_1]^2 \ell(X_1)^2) \geq \underline{\alpha}^2 \|\ell\|_{P,2}^2.$$

Further, by change of variables in $\mathbb{R}^{p_1}$: $u = (d_0 - d)/h,$ so that $\mathsf{d}u = h^{-p_1}\mathsf{d}d$, we have that

$$\|\ell\|_{P,q}^q w^q = \int_{\mathbb{R}^{p_1}} h^{-p_1 q} |K^q((d_0-d)/h)| f_D(d)\mathsf{d}d = \int_{\mathbb{R}^{p_1}} h^{-p_1(q-1)} |K^q(u)| f_D(d_0+uh)\mathsf{d}u$$

so that

$$h^{-p_1(q-1)/q} \underline{f}^{1/q} \left( \int |K|^q \right)^{1/q} \leq \|\ell\|_{P,q} w \leq h^{-p_1(q-1)/q} \bar{f}^{1/q} \left( \int |K|^q \right)^{1/q}.$$

Further, we have that

$$w = \int h^{-p_1} K((d_0-d)/h) f_D(d)\mathsf{d}d = \int K(u) f_D(d_0+uh)\mathsf{d}u.$$

Using the Taylor expansion in $h$ around $h=0$ and the Holder inequality:

$$|w - f_D(d_0)| = \left| \int K(u) h \partial f_D(d_0 + u\tilde{h})' u \mathsf{d}u \right| \leq h\bar{f}' \int \|u\|_\infty |K(u)| du,$$

for some $0 \leq \tilde{h} \leq h$. Hence for all $h < h_1 < h_0$, with $h_1$ depending only on $(K, \bar{f}', \underline{f}, \bar{f})$:

$$\underline{f}/2 \leq w \leq 2\bar{f}.$$

In summary, we have the following finite-sample bounds for all $0 < h < h_1$:

$$\underline{c\alpha}\|\ell\|_{P,2} \leq \sigma \leq \bar{c}\sqrt{1+\bar{\alpha}}\|\ell\|_{P,2}, \quad \underline{\alpha}\|\ell\|_{P,2} \leq L \leq \bar{\alpha}\|\ell\|_{P,2}, \quad \kappa \leq \bar{c}(1+\bar{\alpha})\|\ell\|_{P,3},$$

where

$$h^{-p_1(q-1)/q} \underline{f}^{1/q} \left( \int |K|^q \right)^{1/q} /(2\bar{f}) \leq \|\ell\|_{P,q} \leq h^{-p_1(q-1)/q} \bar{f}^{1/q} \left( \int |K|^q \right)^{1/q} 2/\underline{f}.$$

As $h \to 0$, we have that

$$\sigma \asymp L \asymp \|\ell\|_{P,2} \asymp h^{-p_1/2}, \quad \kappa \lesssim h^{-2p_1/3}, \quad \kappa/\sigma \lesssim h^{-p_1/6}.$$

$\square$

D.3. **Proof of Lemma 9.** Similarly to the proof of Lemma 8, using the LIE and bounded heteroscedasticity, we obtain

$$\|\alpha_0^\star\|_{P,2}^2 \underline{c}^2 \leq \sigma^2 \leq \|\ell\|_{P,2}^2 \bar{c}^2 + \|\alpha_0^\star\|_{P,2}^2 \bar{c}^2,$$

and by the triangle inequality

$$\kappa \leq \|\ell\|_{P,3}\bar{c} + \|\alpha_0^\star\|_{P,3}\bar{c}.$$

It remains to bound $\|\alpha_0^\star\|_{P,q}$. To help this, introduce notation

$$v(X) := f(D \mid Z).$$

Case (a). We have that

$$\alpha_0^\star = \alpha_0 = \mathrm{div}_d(\ell)t + \mathrm{div}_d(t)\ell + \mathrm{div}_d(v)\ell t/v.$$

By the triangle inequality,

$$\|\alpha^\star\|_{P,q} \le \|\mathrm{div}_d(\ell)t\|_{P,q} + \|\mathrm{div}_d(t)\ell\|_{P,q} + \|\mathrm{div}_d(v)\ell t/v\|_{P,q},$$
$$\|\alpha^\star\|_{P,2} \ge \|\mathrm{div}_d(\ell)t\|_{P,2} - \|\mathrm{div}_d(t)\ell\|_{P,2} - \|\mathrm{div}_d(v)\ell t/v\|_{P,2}.$$

Using the bounds assumed in the Lemma, we have

$$\|\mathrm{div}_d(\ell)t\|_{P,q} \le \|\mathrm{div}_d(\ell)\|_{P,q}\bar t; \quad \|\mathrm{div}_d(t)\ell\|_{P,q} \le \bar t'\|\ell\|_{P,q}; \quad \|\mathrm{div}_d(v)\ell t/v\|_{P,q} \le \|\ell\|_{P,q}(\bar f'\bar t/\underline{f}).$$

By the proof of Lemma 8, for all $h < h_1 < h_0$, with $h_1$ depending only on $(K, \bar f', \underline{f}, \bar f)$:

$$\underline{f}/2 \le w \le 2\bar f,$$

and

$$h^{-p_1(q-1)/q}\underline{f}^{1/q}\left(\int |K|^q\right)^{1/q}/(2\bar f) \le \|\ell\|_{P,q} \le h^{-p_1(q-1)/q}\bar f^{1/q}\left(\int |K|^q\right)^{1/q} 2/\underline{f}.$$

Furthermore, by the LIE and the assumed lower bounds in the statement:

$$\begin{aligned}
\|\mathrm{div}_d(\ell)t\|_{P,2}^2 &= w^{-2}\mathrm{E}[\mathrm{div}(\ell)^2\mathrm{E}(t^2|D)] \\
&= w^{-2}h^{-2}h^{-p_1 2}\int (\mathrm{div}K((d_0-d)/h)^2\mathrm{E}(t^2|D=d)f(d)\mathsf{d}d \\
&= w^{-2}h^{-2}h^{-p_1}\int (\mathrm{div}K(u))^2\mathrm{E}(t^2|D=d_0+hu)f(d_0+hu)du \\
&\ge (2\bar f)^{-2}h^{-2}h^{-p_1}\underline{t}^2\underline{f}\int (\mathrm{div}K)^2,
\end{aligned}$$

and similarly

$$\|\mathrm{div}_d(\ell)\|_{P,q}^q \le w^{-q}h^{-q}h^{-p_1(q-1)}\bar f\int |\mathrm{div}K|^q \le (\underline{f}/2)^{-q}h^{-q}h^{-p_1(q-1)}\bar f\int |\mathrm{div}K|^q$$

Case (b). Here we have, using the notation as above

$$\begin{aligned}
\alpha_0^\star(X_1) = \mathrm{E}[\alpha_0 \mid X_1] &= \mathrm{div}_d(\ell(X_1))\mathrm{E}[t(X_1) \mid X_1] \\
&+ \mathrm{E}[\mathrm{div}_d(t(X) \mid X_1]\ell(X_1) + \mathrm{E}[\mathrm{div}_d(v(X))t(X)/v(X) \mid X_1]\ell(X_1).
\end{aligned}$$

Then by contractive property of the conditional expectation $\|\alpha_0^\star\|_{P,q} \le \|\alpha_0\|_{P,q}$, so the upper bounds apply from case (a).

We only need to establish lower bound on $\|\alpha_0^\star\|_{P,2}$. By the triangle inequality,

$$\|\alpha^\star\|_{P,2} \ge \|\mathrm{div}_d(\ell)\mathrm{E}[t \mid X_1]\|_{P,2} - \|\mathrm{E}[\mathrm{div}_d(t) \mid X_1]\ell\|_{P,2} - \|\mathrm{E}[\mathrm{div}_d(t) \mid X_1]\ell\|_{P,2}.$$

By Jensen's inequality, and using the same calculations as in case (a):

$$\|\mathrm{div}_d(\ell(X_1))\mathrm{E}[t(X_1) \mid X_1]\|_{P,2} \le \|\mathrm{div}_d(\ell(X_1))t(X_1)\|_{P,2} \le \bar t\|\mathrm{div}_d(\ell)\|_{P,q};$$
$$\|\mathrm{E}[\mathrm{div}_d(t) \mid X_1]\ell\|_{P,2} \le \|\mathrm{div}_d(t)\ell\|_{P,q} \le \bar t'\|\ell\|_{P,q};$$
$$\|\mathrm{E}[\mathrm{div}_d(v)t/v \mid X_1]\ell\|_{P,2} \le \|\mathrm{div}_d(v)\ell t/v\|_{P,q} \le \|\ell\|_{P,q}(\bar f'\bar t/\underline{f}).$$

And, similarly to the calculation above

$$\begin{aligned}
\|\mathrm{div}_d(\ell)\mathrm{E}[t \mid X_1]\|_{P,2}^2 &= \mathrm{E}[\mathrm{div}_d(\ell)^2\mathrm{E}((\mathrm{E}[t \mid X_1])^2|D)] \\
&= w^{-2}h^{-2}h^{-p_1 2}\int(\mathrm{div}K((d_0-d)/h)^2\mathrm{E}((\mathrm{E}[t \mid X_1])^2|D=d)f(d)\mathsf{d}d \\
&= w^{-2}h^{-2}h^{-p_1}\int(\mathrm{div}K(u)^2\mathrm{E}((\mathrm{E}[t \mid X_1])^2|D=d_0+hu)f(d_0+hu)du \\
&\geq w^{-2}h^{-2}h^{-p_1}\underline{t}^2\underline{f}\int(\mathrm{div}K)^2 \\
&\geq (2\bar{f})^{-2}h^{-2}h^{-p_1}\underline{t}^2\underline{f}\int(\mathrm{div}K)^2,
\end{aligned}$$

using the assumed bound $\mathrm{E}((\mathrm{E}[t \mid X_1])^2|D=d) \geq \underline{t}^2$ for $d \in N_h(d_0)$.

In either case (a) or (b), we now summarize the bounds asymptotically by letting $h \searrow 0$:

$$L \lesssim \sigma \lesssim h^{-p_1/2}(1+h^{-1}), \quad h^{-p_1/2}(h^{-1}-1) \lesssim L \lesssim h^{-p_1/2}(h^{-1}+1),$$
$$\kappa \lesssim h^{-2p_1/3}(h^{-1}+1), \quad \kappa/\sigma \lesssim h^{-p_1/6}.$$

□

**Proof of Lemma 10.** Introduce $m(d) := \mathrm{E}[m(W,\gamma_0^\star) \mid D=d]$ and note

$$\vartheta_1(h) = \int m(d)h^{-p_1}K((d_0-d)/h)f_D(d)\mathsf{d}d = \int m(d_0+hu)K(u)f_D(d_0+hu)\mathsf{d}u,$$

$$\vartheta_2(h) = \int h^{-p_1}K((d_0-d)/h)f_D(d)\mathsf{d}d = \int K(u)f_D(d_0+uh)\mathsf{d}u.$$

Note that by $\int K = 1$,

$$\vartheta_1(0) = m(d_0)f_D(d_0), \quad \vartheta_2(0) = f_D(d_0).$$

Hence

$$\theta(\gamma_0^\star;\ell_h) = \frac{\vartheta_1(h)}{\vartheta_2(h)}, \quad \theta(\gamma_0^\star,\ell_0) := \frac{\vartheta_1(0)}{\vartheta_2(0)} = m(d_0).$$

By the standard argument to control the bias of the higher-order kernel smoothers, e.g. by Lemma B2 in Newey Newey [1994b], which employs the Taylor expansion of order $\mathsf{v}$ in $h$ around $h=0$, for some constants $A_\mathsf{v}$ that depend only on $\mathsf{v}$:

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

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

where $\mathsf{v} = \mathsf{o} \wedge \mathrm{sm}$. Then using the relation

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

we deduce the following bound that applies for all $h < h_1 \leq h_0$,

$$|\theta(\gamma_0^\star;\ell_h)-\theta(\gamma_0^\star,\ell_0)| \leq \left|\frac{\vartheta_1(h)}{\vartheta_2(h)} - \frac{\vartheta_1(0)}{\vartheta_2(0)}\right| \leq Ch^\mathsf{v},$$

where the constant $C$ and $h_1$ depend only on $K, \mathsf{v}, \bar{g}_{\mathsf{v}}, \bar{f}_{\mathsf{v}}, \underline{f}$. $\square$

## References


Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv:1201.0220*, 2011.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28 (2):29–50, 2014a.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014b.

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

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014d.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 2019.

Peter J Bickel and Yaacov Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

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

Jelena Bradic and Mladen Kolar. Uniform inference for high-dimensional quantile regression: Linear functionals and regression rank scores. *arXiv:1702.06209*, 2017.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Sourav Chatterjee and Jafar Jafarov. Prediction error of cross-validated lasso. *arXiv:1502.06291*, 2015.

Victor Chernozhukov and Christian Hansen. The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *Review of Economics and Statistics*, 86(3):735–751, 2004.

Victor Chernozhukov and Vira Semenova. Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions. *arXiv:1702.06240*, 2017.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688, 2015.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv:1608.00033*, 2016.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

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

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Learning L2 continuous regression functionals via regularized Riesz representers. *arXiv:1809.05224*, 2018b.

David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *arXiv preprint arXiv:1908.02399*, 2019.

Christopher Genovese and Larry Wasserman. Adaptive confidence bands. *The Annals of Statistics*, 36(2):875–905, 2008.

Zijian Guo and Cun-Hui Zhang. Local inference in additive models with decorrelated local linear estimator. *arXiv preprint arXiv:1907.12732*, 2019.

RZ Hasminskii and IA Ibragimov. On the nonparametric estimation of functionals. In *Mandl P. and M. Huskova, Proceedings of the Second Prague Symposium on Asymptotic Statistics*, 1979.

Jerry A Hausman and Whitney K Newey. Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, pages 1445–1476, 1995.

Miguel A Hernan and James M Robins. *Causal Inference*. CRC, 2019.

David A Hirshberg and Stefan Wager. Balancing out regression error: Efficient treatment effect estimation without smooth propensities. *arXiv:1712.00038v1*, 2017.

David A Hirshberg and Stefan Wager. Debiased inference of average partial effects in single-index models. *arXiv preprint arXiv:1811.02547*, 2018.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv:1712.00038v5*, 2019.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Jana Jankova and Sara Van De Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

Jana Jankova and Sara Van De Geer. Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv:1610.01353*, 2016.

Jana Jankova and Sara Van De Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

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

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014b.

Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

Bing-Yi Jing, Qi-Man Shao, Qiying Wang, et al. Self-normalized cramér-type large deviations for independent random variables. *The Annals of probability*, 31(4):2167–2215, 2003.

Ying-Ying Lee. Double machine learning nonparametric inference on continuous treatment effects. *University of California Irvine Preprint*, 2019.

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

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

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

Whitney K Newey, Fushing Hsieh, and James M Robins. Undersmoothing and bias corrected functional estimation. Technical report, MIT Department of Economics, 1998.

Whitney K Newey, Fushing Hsieh, and James M Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72(3):947–962, 2004.

Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443, 2018.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT press, 2017.

James M Poterba and Steven F Venti. 401(k) plans and tax-deferred saving. In *Studies in the Economics of Aging*, pages 105–142. University of Chicago Press, 1994.

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

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

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

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

Dominik Rothenhäusler and Bin Yu. Incremental causal effects. *arXiv:1907.13258*, 2019.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

Richard Schmalensee and Thomas M Stoker. Household gasoline demand in the United States. *Econometrica*, 67(3):645–662, 1999.

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

B Toth and MJ van der Laan. TMLE for marginal structural models based on an instrument. Technical report, UC Berkeley Division of Biostatistics, 2016.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2012.

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

Mark J Van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media, 2011.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Aad Van Der Vaart et al. On differentiable functionals. *The Annals of Statistics*, 19(1): 178–204, 1991.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

Adonis Yatchew and Joungyeo Angela No. Household gasoline demand in Canada. *Econometrica*, 69(6):1697–1709, 2001.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540, 2010.

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

Yinchu Zhu and Jelena Bradic. Breaking the curse of dimensionality in regression. *arXiv:1708.00430.*, 2017.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

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