Manuscript submitted to *The Econometrics Journal*, pp. S1–S26.

De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers

Victor Chernozhukov[†], Whitney K. Newey[†], and Rahul Singh[†]

[†]*MIT Economics, 50 Memorial Drive, Cambridge MA 02142, USA.*
E-mail: vchern@mit.edu, wnewey@mit.edu, rahul.singh@mit.edu

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

Keywords: *Neyman orthogonality, Gaussian approximation, sparsity*

S1. RELATED WORK

S1.1. Previous Learning Problems

The paper 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, 2014, 2015); Javanmard and Montanari (2014a,b, 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 and learn the RR automatically from the data.

S1.2. De-biased Estimation

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. Newey (1994a) derives such scores as a part of the computation of the semi-parametric efficiency bound for regular functionals. 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 a low-dimensional nonparametric setting. In the regular case, Chernozhukov et al. (2016, 2018) analyze the double robust/debiased 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—see Section S3 for the explicit definition of the additive model and structure of representers in that case). 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.

Various papers have considered direct estimation of the Riesz representer. Among these papers, ours is the first to present a framework for direct estimation of the Riesz representer of a *broad class of linear functionals, in a high-dimensional setting, without requiring strong Donsker class assumptions*. The earliest reference of which we know is Robins et al. (2007), a comment on another paper, which consider only the global average treatment effect (ATE). Zhu and Bradic (2017) show that it is possible to attain $\sqrt{n}$-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. Newey and Robins (2018) present and analyze estimators based on regression splines, while we present and analyze sparse estimators methods for the high-dimensional setting. The Athey et al. (2018) estimator of the ATE is based on sparse linear regression 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.

Since the first version of this paper was posted online, subsequent work has built upon its insights. 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. Hirshberg and Wager (2018) consider the average derivative functional in a single index model, analyzing a variant of the estimator proposed here, adapted to the single-index regression structure. Rothenhäusler and Yu (2019) 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 follow-up work, Chernozhukov, Newey, and Singh (2018) consider 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.

S1.3. Localized Functionals

A new development incorporated in this version of the 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. Various papers have studied debiased moment equations for certain localized functionals: conditional average treatment effect (CATE), continuous treatment effect (CTE), and regression derivative at a point. We instead present a unified analysis for the general class of localized functionals. 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).

The debiased CATE and CTE literature is vast. Prominent examples of the debiased CATE literature include Wang et al. (2010), van der Laan and Luedtke (2014), Luedtke and Van Der Laan (2016), Nie and Wager (2017), Lee et al. (2017), and most recently Kennedy (2020). 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. Prominent examples of the debiased CTE literature include Rubin and van der Laan (2006), Díaz and van der Laan (2013), Galvao and Wang (2015), Kennedy et al. (2017), Kallus and Zhou (2018), and Colangelo and Lee (2020). These works develop inference on perfectly localized average potential outcomes with continuous treatment effects, using a different approach than what we develop here. Our development is complementary as it covers a much broader collection of functionals.

The debiased literature on regression derivative at a point is more recent. 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. Semenova and Chernozhukov (2021) 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 (without reliance on close-form solutions for Riesz representers), and kernels are used for localization instead of series.

Our work complements existing work that considers the problem of estimating general nonpathwise differentiable functionals like the localized ones here. Early contributions include Robins and Rotnitzky (2001), Van Der Laan and Dudoit (2003), and Rubin and van der Laan (2005). More recently, Athey et al. (2019) consider this issue in the context of generalized random forests. Foster and Syrgkanis (2019) present a general theory, but without inference guarantees. Unlike previous work, we analyze finite sample Gaussian approximation.

S2. NOTATION AND PRELIMINARIES

S2.1. 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 $\mathbb{E}_I f$ denote the empirical average of $f(W_i)$ over $i \in I \subset \{1, ..., n\}$: $\mathbb{E}_I f := \mathbb{E}_I f(W) = |I|^{-1} \sum_{i \in I} f(W_i)$. 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 ℓ_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 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 $a \le Cb$ for all n sufficiently large, and $a \asymp b$ if both $a \lesssim Cb$ and $b \lesssim Ca$ for all n sufficiently large, where $C \ge 1$ is a positive constant that does not depend on n.

S2.2. 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 it has a bounded operator norm:

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

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 ϕ which is also continuous and linear and which has the same operator norm as ϕ.

Riesz–Frechet representation theorem. Let H be a Hilbert space over $\mathbb{R}$ with an 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, $\|T\|_{op} = \|g\|$, where $\|T\|_{op}$ 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 σ-finite measure are defined, μ and ν. If $\nu \ll \mu$ (i.e. ν is absolutely continuous with respect to μ), 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.

S3. STRUCTURE OF FUNCTIONALS AND THEIR SCORES IN LEADING EXAMPLES

We see that the key quantities in the main inference results are the operator norm L of the linear functional and the standard deviation σ and kurtosis κ/σ of the score ψ_0. In this section we establish bounds on these quantities in the key Examples 2.1, 2.2, 2.3, and 2.4, focusing on either unrestricted or additive nonparametric models.

S3.1. Structure of Riesz representers for unrestricted and additive models

Below we derive linear representers through change of measure and integration by parts. These representers are universal since they apply to the *unrestricted model*, where $\bar{\Gamma} = L^2(F)$. We remark here that these representers are universal, since they can represent θ_0 even when $\bar{\Gamma} \neq L^2(F)$, if they exist. These universal representers are not minimal unless $\bar{\Gamma} = L^2(F)$. Theorem 4.2 implies that it is better to use the minimal representer than the universal representer to attain full semi-parametric efficiency (unless $\Gamma = L^2(F)$).

Consider the following (some well-known) candidates for universal linear representers in Examples 2.1, 2.2, 2.3, and 2.4:

$$\alpha_0(x;\ell) = [(1(d=1) - 1(d=0))/\mathrm{P}(D=d \mid Z=z)]\ell(x); \tag{S.1}$$

$$\alpha_0(x;\ell) = [d(F_1(x) - F_0(x))/dF(x)]\ell(x); \tag{S.2}$$

$$\alpha_0(x;\ell) = [d(F_1(x) - F(x))/dF(x)]\ell(x), \ F_1 = \mathrm{Law}(T(X)); \tag{S.3}$$

$$\alpha_0(x;\ell) = -(\operatorname{div}_d(\ell(x)t(x)f(d|z))/f(d|z), \ f(d|z) = \text{ pdf of } D \text{ given } Z=z; \tag{S.4}$$

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(ℓ), div_d denotes the divergence of scalar function:

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

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

LEMMA 3.1. (UNIVERSAL REPRESENTERS FOR KEY EXAMPLES) *In Examples 2.1, 2.2, 2.3, and 2.4, (i) If $\alpha_0(X;\ell)$ is real-valued a.s. and $\alpha_0(\cdot;\ell) \in L^2(F)$, then it is the universal representer for the corresponding linear functional $\gamma \mapsto \theta(\gamma)$, and the latter is continuous. In Example 2.4, 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, the minimal representer $\alpha_0^\star$ is obtained by projecting α_0 onto $\bar{\Gamma}$. (ii) There are examples of P, exhibited in the proof of this lemma, such that linear functionals* *in Examples 2.1, 2.2, 2.3, and 2.4 can be continuous on* $\bar{\Gamma} \neq L^2(F)$*, but* $\alpha_0(X;\ell) = \text{na}$ *with positive probability.*

Part of the lemma is well known (for example, the $\alpha_0(X;\ell)$ representer for ATE is the Horvitz-Thompson transformation), while a part of lemma appears to be new. 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 on some subsets of $L^2(F)$ 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 Γ is not dense in $L^2(F)$ and where minimal representers are not equal to the universal representers.

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

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

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* ℓ *depends only on the first component, namely* $m(w,\gamma;\ell) = m(w,\gamma_1;\ell)$ *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. We can immediately see that the minimal representers can be generated as conditional expectations of the universal representers.

Lemma 3.2. (Order-preserving, contractive representers for additive models) *Work with AM and assume* $\alpha_0(\cdot;\ell) \in L^2(F)$*. Then on* $\gamma \in \Gamma$*,*

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

where α_0 *is any linear representer for* $\gamma \mapsto \theta(\gamma)$ *on* Γ*. 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 under condition AM.

S3.2. Structure of global functionals and scores in key examples

Here we develop bounds on the key quantities: the standard deviation σ of the score, the kurtosis κ/σ, and the modulus of continuity L. In the regular case, these quantities are bounded. Here we would like to study how the bounds depend on L, and we analyze the non-regular cases arising from taking a 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 2.1, 2.2, and 2.3, though the proofs are useful to characterize bounds in other contexts. Our goal is to fix a weighting function ℓ, and to consider how a non-regularity $L \to \infty$ can arise from modeling quantities like

$$1/\mathrm{P}(D = d \mid Z), \quad (d(F_1 - F_0)/dF) \circ X, \quad (d(F_1 - F)/dF) \circ X, \tag{S.5}$$

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 Example 2.4, 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 2.4 and focus on localization as the source.

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 3.3. (Structure of global average effects functionals and scores) *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 (S.1), (S.2), (S.3) for Examples 2.1, 2.2, and 2.3 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:*

$$\|\alpha_0^\star\|_{P,3} \leq c(\|\alpha_0^\star\|_{P,2}^2 \vee 1), \tag{S.6}$$

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

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

Condition (S.6) 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 2.1. 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 $\|\alpha_0\|_{P,3} \asymp (1/\pi^2)^{1/3} \ll \|\alpha_0\|_{P,2}^2$ when $\pi \searrow 0$, so the condition is easily met.

S3.3. Structure of local and localized functionals and scores in key examples

Here we focus on local functionals and develop bounds that relate key quantities: the standard deviation σ of the score, the kurtosis κ/σ, and the modulus of continuity L.

Our first goal is examine how the localization of the weighting function ℓ creates the non-regularity $L \to \infty$. Our inference theory outlined above covers local functionals provided $L/\sqrt{n}$ is small, and it also covers perfectly localized functionals 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 a 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 3.4. (Structure of local average effects functionals and scores) *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 (S.1), (S.2), and (S.3), corresponding to Examples 2.1, 2.2, and 2.3, exists and obeys*

$$0 < \underline{\alpha} \le \alpha_0(X;1) \le \bar{\alpha}, \qquad a.s. \tag{S.7}$$

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

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

obey the bounded heteroscedastic moment conditions:

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

Suppose that the pdf f_D of D obeys the bounds:

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

Then the non-asymptotic 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 h going to zero. The condition (S.7) shuts down the previous source of non-regularity, and says that the quantities in (S.5) 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 3.5. (Structure of local average derivative functionals and scores) *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 (S.4) 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:

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

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

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

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

Then the non-asymptotic 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. In what follows the norm of a tensor $T = \partial^{\mathsf{v}}/(\partial d)^{\mathsf{v}}$ is defined as the injective norm

$$|T|_{op} = \sup_{\|u_1\|_2 \le 1, \ldots, \|u_{\mathsf{v}}\|_2 \le 1} |\langle T, u_1 \otimes .... \otimes u_{\mathsf{v}} \rangle|.$$

Lemma 3.6. (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* sm*, and for* $\mathsf{v} := \mathsf{sm} \wedge \mathsf{o}$ *and* ∂_d^{v} *denoting the tensor* $\partial^{\mathsf{v}}/(\partial d)^{\mathsf{v}}$ *we have*

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

In addition, assume

$$m(d_0) f_D(d_0) \le \bar{g}.$$

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

$$|\theta(\gamma_0^\star; \ell_h) - \theta(\gamma_0^\star; \ell_0)| \le C h^{\mathsf{v}},$$

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

S4. PROOFS FOR SECTION 2

S4.1. Proof of Lemma 2.1

We note that $\Gamma = \mathrm{span}(\Gamma_0)$ 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 α_0. Denote by $\alpha_0^\star$ the projection of α_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 θ to $\bar{\Gamma}$ by defining $\tilde{\theta}(\gamma) = \varphi(\gamma) = \langle \gamma, \alpha_0^\star \rangle$ 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 the orthogonal projection reduces the 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.□

S5. DETAILS FOR SECTION 3

S5.1. Practical implementation details

In practice we use the following generic algorithm for computing GDS 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:

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

3 Update the GDS estimate, using the current estimate as the starting point in the algorithm:

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

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. (2012) 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 a small, e.g. $\mathsf{a} = 0.1$.

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

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

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 $c\Phi^{-1}(1 - \mathsf{a}/2))/\sqrt{n}$).

S5.2. Partial difference

Consider a simplification of Example 2.4, average derivative:

$$\theta_0^\star = \int \partial_d \gamma_0^\star(d, z)\ell(x) dF(x).$$

For nonparametric regression estimators that are linear in a dictionary $b(d, z)$, e.g. GDS and Lasso, the average derivative is straightforward to compute: apply the learned coefficients $\hat{\beta}$ to the derivative of the dictionary $\partial_d b(d,z)$, and average across observations using weighting $\ell(x)=\ell(d,z)$.

Random forest is an example of a nonparametric regression estimator that is not differentiable. A neural network is differentiable, but its derivative at each observation may be difficult to access when using a black-box implementation. For this reason, when using random forest or neural network, we use an average partial difference approximation of the average derivative.

Specifically, consider the average partial difference functional

$$\theta_0^* = \int \left[\gamma_0^\star(d+\Delta/2,z) - \gamma_0^\star(d-\Delta/2,z)\right] \frac{1}{\Delta}\ell(x)dF(x).$$

The theory developed for Example 2.3, policy effect from transporting X, directly applies to average partial difference. In practice, we take Δ to be one fourth of the standard deviation of D.

There is an important connection between average derivative and average partial difference when using a nonparametric regression estimator that is linear in a dictionary $b(d,z)$, e.g. GDS and Lasso. If the dictionary $b(d,z)$ is quadratic in d, then the average derivative estimate must be *numerically identical* to the average partial difference estimate. The specification from Semenova and Chernozhukov (2021) that we use when estimating average price elasticity of gasoline is quadratic in log price. Therefore Table 3 presents average partial difference estimates that perfectly coincide with average derivative estimates for GDS and Lasso, and that approximate average derivative estimates for random forest and neural network.

S5.3. Empirical results without debiasing

We present tables analogous to those in Section 3 without debiasing. Tables 1, 2, and 3 in the supplement correspond to Tables 1, 2, and 3 in the main text, respectively.

Table 1. Average treatment effect of 401(k) eligibility on net financial assets without debiasing. Localized average treatment effects are reported by income quintile groups. The regression is estimated by GDS or Lasso. Standard errors are reported in parentheses.

Income quintile	N treated	N untreated	GDS		Lasso	
All	3682	6187	3763.35	(31.01)	4526.42	(42.33)
1	272	1702	2604.14	(8.05)	2581.88	(26.53)
2	527	1447	126.69	(5.92)	298.56	(23.29)
3	755	1219	2819.64	(13.94)	2536.49	(28.56)
4	962	1012	5996.15	(57.05)	3287.30	(84.56)
5	1166	807	4528.12	(103.84)	6905.36	(159.28)

Table 2. Average treatment effect of 401(k) eligibility on net financial assets without debiasing. Localized average treatment effects are reported by income quintile groups. The regression is estimated by random forest or neural network. Standard errors are reported in parentheses.

Income quintile	N treated	N untreated	Random forest		Neural network	
All	3682	6187	10543.48	(178.37)	7807.97	(336.42)
1	272	1702	4378.26	(134.08)	4266.68	(308.06)
2	527	1447	1477.09	(329.52)	1281.15	(537.07)
3	755	1219	6997.80	(158.49)	5331.58	(336.25)
4	962	1012	12854.02	(467.54)	10234.88	(807.86)
5	1166	807	26845.23	(749.52)	21426.42	(1615.20)

Table 3. Estimated average derivative (price elasticity) of gasoline demand without debiasing. Localized average derivatives are reported by income quintile groups. The regression is estimated by GDS, Lasso, random forest, or neural network. Standard errors are reported in parentheses.

Income quintile	N	GDS		Lasso		Random forest		Neural network	
All	5001	-0.53	(0.00)	-0.06	(0.00)	-0.09	(0.02)	0.17	(0.01)
1	1001	-0.55	(0.01)	0.00	(0.00)	-0.26	(0.07)	0.18	(0.03)
2	1000	-0.34	(0.01)	0.00	(0.00)	-0.15	(0.07)	0.41	(0.03)
3	1000	-0.44	(0.01)	0.00	(0.00)	-0.30	(0.06)	-0.21	(0.03)
4	1000	-0.22	(0.01)	0.00	(0.00)	-0.15	(0.07)	0.23	(0.04)
5	1000	-0.05	(0.00)	0.00	(0.00)	0.00	(0.07)	0.61	(0.02)

S6. PROOFS FOR SECTION 4

S6.1. Proof of Theorem 4.1

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 4.1, derivatives don't depend on θ. Hence for all θ,

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

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

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

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 4.1, we have by the exact Taylor expansion around (β_0, ρ_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:

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

where we have used that by Lemma 4.1

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

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

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

using the fact that $\mathrm{E}\left(\sum_{j=1}^J V_j\right)^2 \leq J\sum_{j=1}^J \mathrm{E}V_j^2$ 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, $\text{rem}_{1k}, \text{rem}_{2k}$ and rem_{3k} have conditional mean 0 and conditional variance given by

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

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

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

In summary,

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

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

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

By Markov inequality we have

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

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

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

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,

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

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

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

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

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

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

S6.2. Proof of Theorem 4.2

We shall verify the hypotheses of Van der Vaart (2000), Theorem 25.20.

Step 1. Suppose that W had Radon–Nykodym derivative dP under P with respect to some measure μ. Consider the set for some $\varepsilon > 0$:

$$\mathcal{S}_\varepsilon = \{\delta \text{ measurable } : \mathcal{W} \to \mathbb{R}, \int \delta dP = 0, \|\delta\|_\infty \leq 1/(2\varepsilon)\}.$$

Consider a parametric submodel (i.e. path) of the form

$$\mathcal{P} = \Big\{ dP_\tau(w) = dP(w)[1 + \tau\delta(w)] : \quad \delta \in \mathcal{S}_\varepsilon \}_{\tau \in (0,\varepsilon)}.$$

It is standard to verify that δ is the score of dP_τ, namely $\delta(w) = \partial_\tau \log dP_\tau(w)$, and that quadratic mean differentiability holds:

$$\int [(\sqrt{dP_\tau} - \sqrt{dP})/\tau - (\delta/2)d\sqrt{dP}]^2 d\mu \to 0,$$

which implies that deviations from P are locally asymptotically normal. The collection of scores $\mathcal{S}_\varepsilon$ therefore form the tangent set of $\mathcal{P}$ at P.

Consider the parameter of interest:

$$\theta_\tau = \int m(w, \gamma_\tau) dP_\tau,$$

where $\gamma_\tau^\star$ abbreviates the heavy notation $\gamma_{0,P_\tau}^\star$, denoting the projection of Y on $\bar{\Gamma}$ under P_τ. We will also use $\gamma_0^\star$ to denote $\gamma_{0,P}^\star$.

Step 2 below shows the differentiability of the parameter with respect to τ:

$$\frac{\theta_\tau - \theta_0}{\tau} \to \int \psi_0 \delta dP, \text{ for each } \delta \in \mathcal{S}_\varepsilon,$$

where ψ_0 is a score function. This is done in Step 2 below.

This score function belongs to the $L^2(P)$ closure of the linear span of $\mathcal{S}_\varepsilon$:

$$\overline{\text{span}}(\mathcal{S}_\varepsilon) = \Big\{\delta \in L^2(P) : \int \delta dP = 0\Big\}.$$

so it follows that ψ_0 is the projection of itself on the $\mathcal{S}_\varepsilon$ and is therefore the only influence function.

Step 2. Because δ is bounded by $1/(2\varepsilon)$, the dP_τ and dP dominate each other so that $\bar{\Gamma}$ does not depend on τ. Let E_τ denote expectation under P_τ and E under P.

Then for some generic positive finite constant C

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

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

$$\mathrm{E}\left[\gamma_\tau^\star(X)\gamma_0^\star(X)\right] = \mathrm{E}_\tau\left[\gamma_\tau^\star(X)\gamma_0^\star(X)\right] + o(1)$$

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

Similarly we have

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

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

Therefore it follows that

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

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

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

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

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

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

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

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

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

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

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

□

S6.3. Proof of Lemma 4.2

First, we note that

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

Define

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

Note that

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

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

$$\|t_0 + \delta\|_1 \le \|t_0\|_1 \iff \|t_0^{\mathcal{M}} + \delta_{\mathcal{M}}\|_1 + \|t_0^r + \delta_{\mathcal{M}^c}\|_1 \le \|t_0^{\mathcal{M}}\|_1 + \|t_0^r\|_1$$

$$\implies \|\delta_{\mathcal{M}^c}\|_1 - \|t_0^r\|_1 \le \|t_0^r + \delta_{\mathcal{M}^c}\|_1 \le \|t_0^{\mathcal{M}}\|_1 - \|t_0^{\mathcal{M}} + \delta_{\mathcal{M}}\|_1 + \|t_0^r\|_1$$

$$\implies \|\delta_{\mathcal{M}^c}\|_1 - \|t_0^r\|_1 \le \|\delta_{\mathcal{M}}\|_1 + \|t_0^r\|_1 \implies \|\delta_{\mathcal{M}^c}\|_1 \le \|\delta_{\mathcal{M}}\|_1 + 2\|t_0^r\|_1.$$

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

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

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

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

S6.4. Proof of Lemma 4.3

Consider the event $\mathcal{R}$ such that

$$\|\hat{g}(t_0)\|_\infty \le \lambda, \qquad \|\hat{g}(\hat{t})\|_\infty \le \lambda, \tag{S.8}$$

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

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

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

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

S6.5. Proof of Theorem 4.3 and Corollary 4.4

Application of Lemma 4.3 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

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

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

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 corollary. □

S7. PROOFS FOR SECTION 5

S7.1. Proof of Theorem 5.1

Let $\phi(w,\gamma,\alpha) = \alpha(x)[y-\gamma(x)]$, $\psi(w,\gamma,\alpha,\theta) = \theta - m(w,\gamma) - \phi(w,\gamma,\alpha)$, $\bar{\phi}(\gamma,\alpha) = \int \phi(w,\gamma,\alpha)F_0(dw)$, and $\bar{m}(\gamma) = \int m(w,\gamma)F_0(dw)$. Note that

$$\bar{\phi}(\gamma_0^\star,\alpha_0^\star) = 0,\ \bar{\phi}(\gamma_0^\star,\hat{\alpha}_k) = 0,\ \bar{m}(\hat{\gamma}_k - \gamma_0^\star) = -\bar{\phi}(\hat{\gamma}_k,\alpha_0^\star). \tag{S.9}$$

Then we have

$$\hat{\theta}_k - \theta_0 + \frac{1}{n_k}\sum_{i\in I_k}\psi_0^\star(W_i) = \frac{1}{n_k}\sum_{i\in I_k}\{\psi(W_i,\gamma_0^\star,\alpha_0^\star,\theta_0) - \psi(W_i,\hat{\gamma}_k,\hat{\alpha}_k,\theta_0)\}$$

$$= \frac{1}{n_k}\sum_{i\in I_k}\{m(W_i,\hat{\gamma}_k) + \phi(W_i,\hat{\gamma}_k,\hat{\alpha}_k) - m(W_i,\gamma_0^\star) - \phi(W_i,\gamma_0^\star,\alpha_0^\star)\} = \hat{R}_1 + \hat{R}_2,$$

where

$$\begin{aligned}
\hat{R}_1 &= \frac{1}{n_k}\sum_{i\in I_k}[m(W_i,\hat{\gamma}_k - \gamma_0^\star) - \bar{m}(\hat{\gamma}_k - \gamma_0^\star)] \qquad \text{(S.10)}\\
&\quad + \frac{1}{n_k}\sum_{i\in I_k}[\phi(W_i,\hat{\gamma}_k,\alpha_0^\star) - \phi(W_i,\gamma_0^\star,\alpha_0^\star) - \bar{\phi}(\hat{\gamma}_k,\alpha_0^\star)]\\
&\quad + \frac{1}{n_k}\sum_{i\in I_k}[\phi(W_i,\gamma_0^\star,\hat{\alpha}_k) - \phi(W_i,\gamma_0^\star,\alpha_0^\star) - \bar{\phi}(\gamma_0^\star,\hat{\alpha}_k)],\\
\hat{R}_2 &= \frac{1}{n_k}\sum_{i\in I_k}[\phi(W_i,\hat{\gamma}_k,\hat{\alpha}_k) - \phi(W_i,\hat{\gamma}_k,\alpha_0^\star) - \phi(W_i,\gamma_0^\star,\hat{\alpha}_k) + \phi(W_i,\gamma_0^\star,\alpha_0^\star)]\\
&= -\frac{1}{n_k}\sum_{i\in I_k}[\hat{\alpha}_k(X_i) - \alpha_0^\star(X_i)][\hat{\gamma}_k(X_i) - \gamma_0^\star(X_i)]. \qquad \text{(S.11)}
\end{aligned}$$

Define $\hat{\Delta}_{ik} = m(W_i,\hat{\gamma}_k - \gamma_0^\star) - \bar{m}(\hat{\gamma}_k - \gamma_0^\star)$ for $i \in I_k$ and let $\mathcal{W}_k^c$ denote the observations W_i for $i \notin I_k$. Note that $\hat{\gamma}_k$ depends only on $\mathcal{W}_k^c$ by construction. Then by independence of

$\mathcal{W}_k^c$ and $\{W_i, i \in I_k\}$ we have $\mathrm{E}[\hat{\Delta}_{ik}|\mathcal{W}_k^c] = 0$. Also by independence of the observations, $\mathrm{E}[\hat{\Delta}_{ik}\hat{\Delta}_{jk}|\mathcal{W}_k^c] = 0$ for $i, j \in I_k$. Furthermore, for $i \in I_k$ $\mathrm{E}[\hat{\Delta}_{ik}^2|\mathcal{W}_k^c] \leq \int[m(w, \hat{\gamma}_k - \gamma_0^\star)]^2 F_0(dw)$. Then by equation (5.20) we have

$$\mathrm{E}\left[\left(\frac{1}{n_k}\sum_{i\in I_k}\hat{\Delta}_{ik}\right)^2|\mathcal{W}_k^c\right] = \frac{1}{n_k^2}\mathrm{E}\left[\left(\sum_{i\in I_k}\hat{\Delta}_{ik}\right)|\mathcal{W}_k^c\right] = \frac{1}{n_k^2}\sum_{i\in I_k}\mathrm{E}[\hat{\Delta}_{ik}^2|\mathcal{W}_k^c]$$
$$\leq \frac{1}{n_k}\int[m(w, \hat{\gamma}_k - \gamma_0^\star)]^2 F_0(dw) = o_p(\sigma^2/n_k) = o_p(\sigma^2/n).$$

The conditional Markov inequality then implies that $\sum_{i\in I_k}\hat{\Delta}_{ik}/n = o_p(\sigma/\sqrt{n})$. The analogous results also hold for $\hat{\Delta}_{ik} = \phi(W, \hat{\gamma}_k, \alpha_0^\star) - \phi(W, \gamma_0^\star, \alpha_0^\star) - \bar{\phi}(\hat{\gamma}_k, \alpha_0^\star)$ and $\hat{\Delta}_{ik} = \phi(W, \gamma_0^\star, \hat{\alpha}_k) - \phi(W, \gamma_0^\star, \alpha_0^\star) - \bar{\phi}(\gamma_0^\star, \hat{\alpha}_k)$ by $\bar{\phi}(\gamma_0^\star, \alpha_0^\star) = 0$. Summing across the three terms in $\hat{R}_1$ gives $\hat{R}_1 = o_p(\sigma/\sqrt{n})$.

Next let $\hat{\Delta}_k(x) = -[\hat{\alpha}_k(x) - \alpha_0^\star(x)][\hat{\gamma}_k(x) - \gamma_0^\star(x)]$. Then by the triangle and Cauchy-Schwartz inequalities,

$$\mathrm{E}[|R_2|\,|\mathcal{W}_k^c] \leq \int\left|\hat{\Delta}_k(x)\right|F(dx) \leq \|\hat{\alpha}_k - \alpha_0^\star\|_{P,2}\,\|\hat{\gamma}_k - \gamma_0^\star\|_{P,2} = \sigma\sigma^{-1}\,\|\hat{\alpha}_k - \alpha_0^\star\|_{P,2}\,\|\hat{\gamma}_k - \gamma_0^\star\|_{P,2}$$
$$\leq \sigma\sigma^{-1}(\|\hat{\alpha}_k - \alpha_0\|_{P,2} + \|\alpha_0 - \alpha_0^\star\|_{P,2})\,\|\hat{\gamma}_k - \gamma_0^\star\|_{P,2}\,.$$

By hypothesis $r_2^\star r_1^\star = o\left(1/\sqrt{n}\right)$, so that by the conditional Markov inequality and the definition of $r_2^\star$,

$$\hat{R}_2 = O_p(\sigma r_2^\star r_1^\star) = o_p(\sigma/\sqrt{n}).$$

The conclusion then follows by the triangle inequality. □

S8. PROOFS FOR SECTION S3

S8.1. Proof of Lemma 3.1

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 2.1, 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 2.2, $\ell\alpha_0 \in L^2(F)$ means that the Radon–Nykodym derivatives $\frac{dF_1}{dF}$ and $\frac{dF_0}{dF}$ exist on the support of ℓ, 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 2.3 similarly to Example 2.2.

In Example 2.4, we can write

$$\langle\alpha_0, \gamma\rangle = -\int\int\gamma(x)\frac{\mathrm{div}_d(\ell(x)t(x)f(d|z))}{f(d|z)}f(d|z)\mathrm{d}d\,dF(z)$$
$$= \int\int\partial_d\gamma(x)'t(x)\ell(x)f(d|z)\mathrm{d}d\,dF(z) = \theta(\gamma),$$

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 2.1.

Proof of claim (ii). We can refer to the case of linear regression discussed in Section 2.3. In what follows consider the case of $G > 0$ and $\ell = 1$.

In Example 2.1, $M = \mathrm{E}(b(1,Z) - b(0,Z))$. 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 na with probability π.

In Example 2.2 and 2.3, $M = \int b(dF_1 - dF_0)$ 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 2.4, 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. $\square$

S8.2. Proof of Lemma 3.2

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. $\square$

S8.3. Proof of Lemma 3.3

The proof uses the fact that $m(W,\gamma) = m(X,\gamma)$, and that

$$\psi^\star(X)_0(W) = -U_1 - \alpha_0^\star(X)U_2.$$

Since $\mathrm{E}U_1U_2\alpha_0^\star(X) = 0$ 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. $\square$

S8.4. Proof of Lemma 3.4

We shall use that $m(W,\gamma) = m(X,\gamma)$, and

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

Then by $\mathrm{E}U_1U_2\alpha_0^\star(X) = 0$, 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

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

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

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

so that

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

Further, we have that

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

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

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

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

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

In summary, we have the following non-asymptotic 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

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

$\square$

S8.5. Proof of Lemma 3.5

Similarly to the proof of Lemma 3.4, 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,

$$\|\alpha_0^\star\|_{P,q} \leq \|\mathrm{div}_d(\ell)t\|_{P,q} + \|\mathrm{div}_d(t)\ell\|_{P,q} + \|\mathrm{div}_d(v)\ell t/v\|_{P,q},$$

$$\|\alpha_0^\star\|_{P,2} \geq \|\mathrm{div}_d(\ell)t\|_{P,2} - \|\mathrm{div}_d(t)\ell\|_{P,2} - \|\mathrm{div}_d(v)\ell t/v\|_{P,2}.$$

Using the bounds assumed in the Lemma, we have

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

By the proof of Lemma 3.4, for all $h < h_1 < h_0$, with h_1 depending only on $(K, \bar{f}', \underline{f}, \bar{f})$:

$$\underline{f}/2 \leq \omega \leq 2\bar{f},$$

and

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

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

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

and similarly

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

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

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

Then by contractive property of the conditional expectation $\|\alpha_0^\star\|_{P,q} \leq \|\alpha_0\|_{P,q}$, 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,

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

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

$$\|\mathrm{div}_d(\ell(X_1))\mathrm{E}[t(X_1) \mid X_1]\|_{P,2} \leq \|\mathrm{div}_d(\ell(X_1))t(X_1)\|_{P,2} \leq \bar{t}\|\mathrm{div}_d(\ell)\|_{P,q};$$

$$\|\mathrm{E}[\mathrm{div}_d(t) \mid X_1]\ell\|_{P,2} \leq \|\mathrm{div}_d(t)\ell\|_{P,q} \leq \bar{t}'\|\ell\|_{P,q};$$

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

And, similarly to the calculation above

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

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 $h \searrow 0$:

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

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

□

S8.6. Proof of Lemma 3.6

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 (1994b), which employs the Taylor expansion of order v in h around $h=0$, for some constants A_v that depend only on v:

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

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

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 $h < h_1 \le h_0$,

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

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

Athey, S., G. W. Imbens, and S. Wager (2018). 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.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2011). Inference for high-dimensional sparse econometric models. *arXiv:1201.0220*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

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

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Volume 4. Johns Hopkins University Press.

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

Bradic, J. and M. Kolar (2017). Uniform inference for high-dimensional quantile regression: Linear functionals and regression rank scores. *arXiv:1702.06209*.

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

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

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

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

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

Chernozhukov, V., W. K. Newey, and R. Singh (2018). Learning L2 continuous regression functionals via regularized Riesz representers. *arXiv:1809.05224*.

Colangelo, K. and Y.-Y. Lee (2020). Double debiased machine learning nonparametric inference with continuous treatments. *arXiv:2004.03036*.

Díaz, I. and M. J. van der Laan (2013). Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference 1*(2), 171–192.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2019). Estimation of conditional average treatment effects with high-dimensional data. *arXiv:1908.02399*.

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

Galvao, A. F. and L. Wang (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association 110*(512), 1528–1542.

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

Guo, Z. and C.-H. Zhang (2019). Local inference in additive models with decorrelated local linear estimator. *arXiv:1907.12732*.

Hasminskii, R. Z. and I. A. Ibragimov (1979). On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*.

Hirshberg, D. A. and S. Wager (2017). Balancing out regression error: Efficient treatment effect estimation without smooth propensities. *arXiv:1712.00038v1*.

Hirshberg, D. A. and S. Wager (2018). Debiased inference of average partial effects in single-index models. *arXiv:1811.02547*.

Hirshberg, D. A. and S. Wager (2019). Augmented minimax linear estimation. *arXiv:1712.00038v5*.

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

Jankova, J. and S. Van De Geer (2016). Confidence regions for high-dimensional generalized linear models under sparsity. *arXiv:1610.01353*.

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

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

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

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

Kallus, N. and A. Zhou (2018). Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pp. 1243–1251.

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

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B, Statistical Methodology 79*(4), 1229.

Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*(7), 1207–1225.

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

Newey, W. K. (1994a). The asymptotic variance of semiparametric estimators. *Econometrica*, 1349–1382.

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

Newey, W. K., F. Hsieh, and J. M. Robins (1998). Undersmoothing and bias corrected functional estimation. Technical report, MIT Department of Economics.

Newey, W. K., F. Hsieh, and J. M. Robins (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica 72*(3), 947–962.

Newey, W. K. and J. R. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv:1801.09138*.

Neykov, M., Y. Ning, J. S. Liu, and H. Liu (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science 33*(3), 427–443.

Nie, X. and S. Wager (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv:1712.04912*.

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

Ren, Z., T. Sun, C.-H. Zhang, and H. H. Zhou (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics 43*(3), 991–1026.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment on "performance of double-robust estimators when inverse probability weights are highly variable". *Statistical Science 22*(4), 544–559.

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

Robins, J. M. and A. Rotnitzky (2001). Comment on "inference for semiparametric models: Some questions and an answer". *Statistica Sinica 11*(4), 920–936.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

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

Rubin, D. and M. J. van der Laan (2005). A general imputation methodology for nonparametric regression with censored data. Technical report, UC Berkeley Division of Biostatistics.

Rubin, D. and M. J. van der Laan (2006). Extending marginal structural models through local, penalized, and additive learning. Technical report, UC Berkeley Division of Biostatistics.

Semenova, V. and V. Chernozhukov (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal 24*(2), 264–289.

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

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

Van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Van Der Laan, M. J. and S. Dudoit (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, UC Berkeley Division of Biostatistics.

van der Laan, M. J. and A. R. Luedtke (2014). Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. Technical report, UC Berkeley Division of Biostatistics.

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

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

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

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Van Der Vaart, A. W. and J. A. Wellner (1996). Weak convergence. In *Weak Convergence and Empirical Processes*, pp. 16–28. Springer.

Wang, L., A. Rotnitzky, and X. Lin (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association 105*(491), 1135–1146.

Zhang, C.-H. and S. S. Zhang (2014). 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.

Zhu, Y. and J. Bradic (2017). Breaking the curse of dimensionality in regression. *arXiv:1708.00430.*.

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

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