The Annals of Statistics
2021, Vol. 49, No. 6, 3206–3227
https://doi.org/10.1214/21-AOS2080
© Institute of Mathematical Statistics, 2021

AUGMENTED MINIMAX LINEAR ESTIMATION

BY DAVID A. HIRSHBERG* AND STEFAN WAGER†

*Graduate School of Business, Stanford University, *davidahirshberg@stanford.edu; †swager@stanford.edu*

Many statistical estimands can expressed as continuous linear functionals of a conditional expectation function. This includes the average treatment effect under unconfoundedness and generalizations for continuous-valued and personalized treatments. In this paper, we discuss a general approach to estimating such quantities: we begin with a simple plug-in estimator based on an estimate of the conditional expectation function, and then correct the plug-in estimator by subtracting a minimax linear estimate of its error. We show that our method is semiparametrically efficient under weak conditions and observe promising performance on both real and simulated data.

1. Introduction. Suppose we observe n independent and identically distributed samples $(Z_i, Y_i) \sim P$ with support in $\mathcal{Z} \times \mathbb{R}$, and we want to estimate a continuous linear functional of the form

$$\psi(m) = \mathbb{E}\big[h(Z_i, m)\big] \quad \text{at } m(z) = \mathbb{E}[Y_i | Z_i = z]. \tag{1}$$

Our main result establishes that we can build efficient estimators for a wide variety of such problems simply by subtracting from a plugin estimator $\psi(\hat{m})$ a minimax linear estimate of its error $\psi(\hat{m}) - \psi(m)$.

The following estimands from the literature on causal inference and missing data are of this type and can be estimated efficiently by our approach.

EXAMPLE 1 (Mean with outcomes missing at random). We observe covariates X_i and some but not all of the corresponding outcomes $Y_i^\star$. We write $W_i \in \{0, 1\}$ to indicate whether the outcome $Y_i^\star$ was observed, and define $Z_i = (X_i, W_i)$ and $Y_i = W_i Y_i^\star$; we then estimate the linear functional $\psi(m) = \mathbb{E}[m(X_i, 1)]$ at $m(x, w) = \mathbb{E}[Y_i | X_i = x, W_i = w]$. This will be equal to the mean $\mathbb{E}[Y_i^\star]$ if, conditional on covariates X_i, each outcome $Y_i^\star$ is independent of its nonmissingness W_i (Rosenbaum and Rubin (1983)).

EXAMPLE 2 (Average partial effect). Letting $Z_i = (X_i, W_i) \in \mathcal{X} \times \mathbb{R}$, we estimate the average of the derivative of the response surface $m(x, w)$ with respect to w, $\psi(m) = \mathbb{E}[\frac{\partial}{\partial w}\{m(X_i, w)\}_{w=W_i}]$. This estimand, and weighted variants of it quantify the average effect of a continuous treatment W_i under exogeneity (Powell, Stock and Stoker (1989)).

EXAMPLE 3 (Average partial effect in the conditionally linear model). In the setting of the previous example, we make the additional assumption that the regression function m is conditionally linear in w, $m(x, w) = \mu(x) + w\tau(x)$. The average partial effect is then $\psi(m) = \mathbb{E}[\tau(X_i)]$.

EXAMPLE 4 (Distribution shift). We estimate the effect of a shift in the distribution of the conditioning variable Z from one known distribution, P_0, to another, P_1, that is,

Received July 2018; revised March 2021.
MSC2020 subject classifications. 62F12.
Key words and phrases. Causal inference, convex optimization, semiparametric efficiency.

$\psi(m) = \int m(z)(dP_1(z) - dP_0(z))$ for $m(z) = \mathbb{E}[Y_i \mid Z_i = z]$. Under exogeneity assumptions, this estimand can be used to compare policies for assigning personalized treatments, and estimators for it form a key building block in methods for estimation of optimal treatment policies.

Below, we first discuss our estimator in the simple case that $h(z, m)$ in (1) does not depend on z, that is, $h(z, m) = \psi(m)$. In this case, for example, in Example 4, we can evaluate $\psi(m)$ without knowledge of the distribution P of z, and we say that our functional of interest $\psi(\cdot)$ is *evaluable*. From Section 1.3 on, we will address the general case where h also depends on z and so, even if we knew m a priori, we could only approximate $\psi(m)$ with a sample average $n^{-1}\sum_{i=1}^n h(Z_i, m)$.

1.1. *Estimating evaluable linear functionals.* Consider the estimation of $\psi(m)$ where $\psi(\cdot)$ is an evaluable mean-square-continuous linear functional. The estimator we propose takes a plugin estimator $\psi(\hat{m})$, and then subtracts out an estimate of its error $\psi(\hat{m}) - \psi(m) = \psi(\hat{m} - m)$ obtained as a weighted average of regression residuals,

$$\hat{\psi} = \psi(\hat{m}) - \frac{1}{n}\sum_{i=1}^n \hat{\gamma}_i\big(\hat{m}(Z_i) - Y_i\big). \tag{2}$$

Our approach builds on a result of Chernozhukov et al. (2016) and Chernozhukov, Newey and Robins (2018), who show that we can use the Riesz representer for ψ to construct efficient estimators of this type.

To motivate this approach recall that, by the Riesz representation theorem, any continuous linear functional $\psi(\cdot)$ on the square integrable functions from $\mathcal{Z}$ to $\mathbb{R}$ has a Riesz representer $\gamma_\psi(\cdot)$, that is, a function satisfying $\int \gamma_\psi(z) f(z)\, dP(z) = \psi(f)$ for all square-integrable functions f (e.g., Peypouquet (2015), Theorem 1.41). Then, if we set $\hat{\gamma}_i = \gamma_\psi(Z_i)$ in (2), the second term in the estimator acts as a correction for the error of $\psi(\hat{m})$ because

$$\begin{aligned}\psi(\hat{m}) - \psi(m) &= \int \gamma_\psi(z)(\hat{m} - m)(z)\, dP(z) \approx \frac{1}{n}\sum_{i=1}^n \gamma_\psi(Z_i)\big(\hat{m}(Z_i) - m(Z_i)\big) \\ &= \frac{1}{n}\sum_{i=1}^n \gamma_\psi(Z_i)\big(\hat{m}(Z_i) - Y_i\big) + \frac{1}{n}\sum_{i=1}^n \gamma_\psi(Z_i)\big(Y_i - m(Z_i)\big).\end{aligned} \tag{3}$$

Thus, plugging the above expression into (2), we see that if we could compute our estimator with the oracle Riesz representer weights $\gamma_\psi(Z_i)$, its error would very nearly be a weighted sum of mean-zero noise $n^{-1}\sum_{i=1}^n \gamma_\psi(Z_i)\varepsilon_i$ where $\varepsilon_i = Y_i - m(Z_i)$. This behavior is asymptotically optimal with a great deal of generality (e.g., Newey (1994), Proposition 4).

Our goal will be to imitate the behavior of this oracle estimator without a priori knowledge of the Riesz representer. One possible approach is to determine the form of the Riesz representer $\gamma_\psi(\cdot)$ by solving analytically the set of equations that define it,

$$\int \gamma_\psi(z) f(z)\, dP(z) = \psi(f) \quad \text{for all } f \text{ satisfying } \int f(z)^2\, dP(z) < \infty, \tag{4}$$

then estimate it and plug the resulting weights $\hat{\gamma}_i = \hat{\gamma}_\psi(Z_i)$ into (2). In the context of our first example, the estimation of a mean with outcomes missing, the Riesz representer is the inverse probability weight $\gamma_\psi(w, x) = w/e(x)$ where $e(x) = P[W_i = 1 \mid X_i = x]$, and this plug-in approach involves first obtaining an estimate $\hat{e}(x)$ of treatment probabilities and then weighting by its inverse. This is the well-known Augmented Inverse Probability Weighting (AIPW) estimator of Robins, Rotnitzky and Zhao (1994). Chernozhukov et al. (2018) provide

general results on the efficiency of such estimators, provided $\hat{\gamma}_\psi(Z_i) - \gamma_\psi(Z_i)$ goes to zero fast enough in squared-error loss.

We take another approach. Considering our regression estimator $\hat{m}$ and the design $Z_1 \dots Z_n$ to be fixed,[1] we simply choose the weights $\hat{\gamma} \in \mathbb{R}^n$ that make our correction term $n^{-1}\sum_{i=1}^n \hat{\gamma}_i(\hat{m}(Z_i) - Y_i)$ a minimax linear estimator of what it is intended to correct for, $\psi(\hat{m} - m)$. To be precise, we first choose an absolutely convex set of functions $\mathcal{F}$ which we believe should contain the regression error $\hat{m} - m$. We then choose weights $\hat{\gamma}_i$ that perform best in terms of worst case mean squared error over possible regression errors $\hat{m} - m \in \mathcal{F}$ and conditional variances satisfying $\mathrm{Var}[Y_i \mid Z_i] \le \sigma^2$. This specifies the weights $\hat{\gamma}$ as the solution to a convex optimization problem,

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma \in \mathbb{R}^n} \left\{ I^2_{\psi,\mathcal{F}}(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2 \right\}, \qquad I_{\psi,\mathcal{F}}(\gamma) = \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^n \gamma_i f(Z_i) - \psi(f) \right\}.$$

The good properties of minimax linear estimators like this one are well known. Donoho (1994) and related papers (Armstrong and Kolesár (2018), Cai and Low (2003), Donoho and Liu (1991), Ibragimov and Khas'minskiĭ (1984), Johnstone (2015), Juditsky and Nemirovski (2009)) show that when a regression function m is in a convex set $\mathcal{F}$ and $Y_i | Z_i \sim N(0, \sigma_i^2)$, a minimax linear estimator of a linear functional $\psi(m)$ will come within a factor 1.25 of the minimax risk over all estimators. In addition to strong conceptual support, estimators of the type have been found to perform well in practice across several application areas (Armstrong and Kolesár (2018), Imbens and Wager (2019), Zubizarreta (2015)).

Methodologically, the main difference between our proposal and the references cited above is that we use the minimax linear approach to debias a plugin estimator $\psi(\hat{m})$ rather than as a stand-alone estimator. Because we "augment" the minimax linear estimator by applying it after regression adjustment in the same way that the AIPW estimator augments the inverse probability weighting estimator, we refer to our approach as the Augmented Minimax Linear (AML) estimator. Our main result establishes semiparametric efficiency of the AML estimator under considerable generality.

We note that the weights $\hat{\gamma}$ that underlie minimax linear estimation can be interpreted as a penalized least-squares solution to a set of estimating equations suggested by the definition (4) of the Riesz representer γ_ψ,

$$\frac{1}{n}\sum_{i=1}^n \gamma_i f(Z_i) \approx \psi(f) \quad \text{for all } f \in \mathcal{F}. \tag{5}$$

These estimating equations generalize covariate balance conditions from the literature on the estimation of average treatment effects, and when analyzing our estimator we build on approaches used to study treatment effect estimators that use balancing weights (e.g., Athey, Imbens and Wager (2018), Graham, De Xavier Pinto and Egel (2012), Imai and Ratkovic (2014), Kallus (2020), Zubizarreta (2015)); see Section 1.5 for further discussion.

The restriction of f to a strict subset $\mathcal{F}$ of the square-integrable functions is necessary, as there are infinitely many square-integrable functions f that agree on our sample $Z_1 \dots Z_n$ and they need not even approximately agree in terms of $\psi(f)$. Our choice of this subset $\mathcal{F}$, a set that characterizes our uncertainty about the regression error function $\hat{m} - m$, focuses our estimated weights $\hat{\gamma}$ on the role they play in ensuring that (5) is satisfied for this function $f = \hat{m} - m$. The size of this subset $\mathcal{F}$, measured by, for example, its Rademacher complexity, determines the accuracy with which these equations (5) can be simultaneously satisfied. The

[1] If we estimate $\hat{m}$ on an auxiliary sample, this is the case when we condition on both that sample and on $Z_1 \dots Z_n$. However, our results do not require $\hat{m}$ to be estimated on an auxiliary sample.

smaller we can make $\mathcal{F}$, that is, the better the consistency guarantees we have for $\hat{m}$, the more accurately we can solve (5). In practice, we may take $\mathcal{F}$ to be a set of smooth functions, functions that are approximately sparse in some basis, functions of bounded variation, etc.

That our weights $\hat{\gamma}_i$ approximately solve the estimating equations (5) does not imply that they estimate the Riesz representer $\gamma_\psi(\cdot)$ well in the mean-square sense. However, to whatever degree the oracle weights $\gamma_i = \gamma_\psi(Z_i)$ also approximately solve (5), it will imply that $\hat{\gamma}$ and $\gamma_\psi(\cdot)$ are close in the sense that

$$\frac{1}{n}\sum_{i=1}^{n}\big[\hat{\gamma}_i - \gamma_\psi(Z_i)\big]f(Z_i) \approx 0 \quad \text{for all } f \in \mathcal{F}. \tag{6}$$

This property holds if and only if the vector with elements $\hat{\gamma}_i - \gamma_\psi(Z_i)$ is small or approximately orthogonal to every vector with elements $f(Z_i)$ for $f \in \mathcal{F}$. And it implies that when $\hat{m} - m \in \mathcal{F}$, our estimator (2) approximates the corresponding oracle estimator, as the difference between them is $n^{-1}\sum_{i=1}^{n}[\hat{\gamma}_i - \gamma_\psi(Z_i)][(\hat{m} - m)(Z_i) - \varepsilon_i]$.

We state below a simple version of our main result. In essence, if an estimator $\hat{m}$ converges to m in mean square and our regression error $\hat{m} - m$ is in a uniformly bounded Donsker class $\mathcal{F}$ or more generally satisfies $(\hat{m} - m)/O_P(1) \in \mathcal{F}$, then our approach can be used to define an efficient estimator.

1.2. *Definitions.* As a measure of the scale of a function f relative to an absolutely convex set $\mathcal{F}$, we define the *gauge* $\|f\|_{\mathcal{F}} = \inf\{\alpha > 0 : f \in \alpha\mathcal{F}\}$. We will write $\mathcal{F}_r$ to denote the localized class $\{f \in \mathcal{F} : \|f\|_{L_2(P)} \le r\}$, $g\mathcal{F}$ to denote the class of products $\{gf : f \in \mathcal{F}\}$, and $h(\cdot, \mathcal{F})$ to denote the image class $\{h(\cdot, f) : f \in \mathcal{F}\}$. We will write $\overline{\mathcal{S}}$ to denote the closure of a subspace $\mathcal{S}$ of the square-integrable functions and $\mathcal{S}_\perp$ to denote its orthogonal complement, and will write $\overline{\operatorname{span}}\mathcal{F}$ to denote the closure of $\operatorname{span}\mathcal{F}$. We will say that a set of functions $\mathcal{F}$ from $\mathcal{Z} \to \mathbb{R}$ is pointwise bounded if $\sup_{f\in\mathcal{F}}|f(z)| < \infty$ for all $z \in \mathcal{Z}$, uniformly bounded if $\sup_{f\in\mathcal{F}}\|f\|_\infty < \infty$ where $\|f\|_\infty = \sup_{z\in\mathcal{Z}}|f(z)|$, and pointwise closed if $f \in \mathcal{F}$ whenever it is the limit of a sequence $f_j \in \mathcal{F}$ in the sense that $\lim_{j\to\infty} f_j(z) = f(z)$ for all $z \in \mathcal{Z}$.

1.3. *Setting.* We observe $(Y_1, Z_1)\ldots(Y_n, Z_n) \overset{\text{i.i.d.}}{\sim} P$ with $Y_i \in \mathbb{R}$ and Z_i in an arbitrary set $\mathcal{Z}$. We assume that $m(z) = \mathbb{E}[Y_i \mid Z_i = z]$ is in a subspace $\mathcal{S}$ of the square integrable functions and that $v(z) = \operatorname{Var}[Y_i \mid Z_i = z]$ is bounded. And we let $\mathcal{F}$ be an absolutely convex set of square integrable functions.

Our estimand is $\psi(m)$ for a continuous linear functional $\psi(\cdot)$ on a subspace $\mathcal{S} \cup \operatorname{span}\mathcal{F}$ of the square integrable functions, which takes the form $\psi(m) = \mathbb{E}h(Z_i, m)$. The Riesz representation theorem guarantees the existence and uniqueness of a function $\gamma_\psi \in \overline{\operatorname{span}}\mathcal{F}$ satisfying the set of equations $\{\mathbb{E}\gamma_\psi(Z)f(Z) = \psi(f) : f \in \overline{\operatorname{span}}\mathcal{F}\}$.[2] We call this function the Riesz representer of ψ on the *tangent space* $\overline{\operatorname{span}}\mathcal{F}$. This generalizes our prior definition (4), coinciding when $\overline{\operatorname{span}}\mathcal{F}$ is the space of square integrable functions.

Our regularity and efficiency claims are relative to the set of all one-dimensional submodels P_t through $P_0 = P$ for which, letting $(Y_t, Z_t) \sim P_t$, the regression functions $m_{P_t}(z) = \mathbb{E}[Y_t \mid Z_t = z]$ are in $\mathcal{S}$ and satisfy $\lim_{t\to 0}\|m_{P_t} - m_P\|_{L_2(P)} = 0$ and the squares of $\epsilon_t = Y_t - m_{P_t}(Z_t)$ are uniformly integrable. For these claims, we use the additional assumptions that there is a regular conditional probability $P[Y_i \in \cdot \mid Z_i = z]$ and that $\mathcal{S}_\perp$ has a dense subset of bounded functions.

[2] In this statement, we implicitly work with the unique extension of the continuous functional $\psi(\cdot)$ defined on $\operatorname{span}\mathcal{F}$ to a functional defined on its closure $\overline{\operatorname{span}}\mathcal{F}$ (e.g., Lang (1993), Theorem IV.3.1).

THEOREM 1. *In the setting above, choose finite $\sigma > 0$ and consider the estimator*

$$\hat{\psi}_{\mathrm{AML}} = \frac{1}{n}\sum_{i=1}^{n}\left[h(Z_i, \hat{m}) - \hat{\gamma}_i\left(\hat{m}(Z_i) - Y_i\right)\right] \quad \textit{where} \tag{7}$$

$$\begin{aligned}\hat{\gamma} &= \operatorname*{argmin}_{\gamma\in\mathbb{R}^n}\left\{I^2_{h,\mathcal{F}}(\gamma) + \frac{\sigma^2}{n^2}\|\gamma\|^2\right\},\\ I_{h,\mathcal{F}}(\gamma) &= \sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\gamma_i f(Z_i) - h(Z_i, f)\right]\right\}.\end{aligned} \tag{8}$$

If $\mathcal{F}$ is uniformly bounded and pointwise closed; $\mathcal{F}$, $\gamma_\psi\mathcal{F}$ and $h(\cdot,\mathcal{F})$ are Donsker; and $h(Z,\cdot)$ is pointwise bounded and mean-square equicontinuous on $\mathcal{F}$ in the sense that $\sup_{f\in\mathcal{F}}|h(z,f)| < \infty$ for each $z\in\mathcal{Z}$ and $\lim_{r\to 0}\sup_{f\in\mathcal{F}_r}\|h(\cdot,f)\|_{L_2(P)} = 0$; then our weights converge to the Riesz representer of ψ on the tangent space $\overline{\operatorname{span}}\mathcal{F}$, that is,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\gamma}_i - \gamma_\psi(Z_i)\right)^2 \to_P 0. \tag{9}$$

If, in addition, $\hat{m}$ has the tightness and consistency properties

$$\|\hat{m} - m\|_{\mathcal{F}} = O_P(1) \quad \textit{and} \quad \|\hat{m} - m\|_{L_2(P_n)} = o_P(1) \tag{10}$$

then our estimator $\hat{\psi}_{\mathrm{AML}}$ is asymptotically linear, that is,

$$\begin{aligned}\hat{\psi}_{\mathrm{AML}} - \psi(m) &= \frac{1}{n}\sum_{i=1}^{n}\iota(Y_i, Z_i) + o_P\left(n^{-1/2}\right) \quad \textit{where}\\ \iota(y,z) &= h(z,m) - \gamma_\psi(z)\left(m(z) - y\right) - \psi(m)\end{aligned} \tag{11}$$

and, therefore, $\sqrt{n}(\hat{\psi}_{\mathrm{AML}} - \psi(m))/V^{1/2} \Rightarrow \mathcal{N}(0,1)$ with $V = \mathbb{E}[\iota(Y,Z)^2]$.

Furthermore, an estimator satisfying (11) is regular on the model class $\mathcal{S}$ if $\mathcal{S}\subseteq\overline{\operatorname{span}}\mathcal{F}$, and asymptotically efficient if, in addition, $v(\cdot)\gamma_\psi(\cdot)\in\overline{\mathcal{S}}$.[3]

Theorem 1 follows from a finite sample result, Theorem 2, that we will discuss in Section 2. We end this section with a few remarks on the statistical behavior of the estimator, focusing on the choices of $\hat{m}, \mathcal{F}, \sigma$ that define a specific estimator $\hat{\psi}$ of this type. We defer the discussion of computational issues to Appendix D (Hirshberg and Wager (2021)).

REMARK 1. Our approach does not require knowledge of the functional form of the Riesz representer $\gamma_\psi(\cdot)$, sparing us the trouble of solving (4) analytically.

REMARK 2. If $\|m\|_{\mathcal{F}} < \infty$, the tightness and consistency properties (10) are satisfied by the penalized least squares estimator $\hat{m} = \operatorname{argmin} n^{-1}\sum_{i=1}^{n}(Y_i - m(Z_i))^2 + \lambda\|m\|_{\mathcal{F}}$ for an appropriate choice of λ (see Appendix E). For example, we might choose $\mathcal{F}$ to be the absolutely convex hull $\{\sum_j \beta_j\phi_j : \|\beta\|_{\ell_1}\le 1\}$ of a sequence of basis functions satisfying $\sum_{j=1}^{\infty}\mathbb{E}\phi_j^2(Z_i) < \infty$. It is Donsker (van der Vaart and Wellner (1996), Section 2.13.2) and

[3]If an estimator satisfies (11), a combination of two simple conditions implies efficiency: $\overline{\operatorname{span}}\mathcal{F} = \overline{\mathcal{S}}$ and $v(\cdot)\overline{\mathcal{S}}\subseteq\overline{\mathcal{S}}$. The first says that we correct for all error functions $\hat{m} - m$ permitted by our assumption that $m\in\mathcal{S}$, and waste no effort on those (in $\mathcal{S}_\perp$) ruled out by it. The second holds when the conditional variance $v(z)$ is sufficiently simple relative to $\overline{\mathcal{S}}$, for example, when $v(z)$ is constant or when the model class $\mathcal{S}$ is fully nonparametric in the sense that it contains an approximation to every square integrable function.

the corresponding estimator $\hat{m}$ is ℓ_1-penalized regression in this basis. This approach is easy to implement and performs well in simulation when λ is chosen by cross-validation. In our simulations, we use a class of this type defined in terms of a basis of scaled Hermite polynomials. Note that the requirement of a square summable basis rules out the settings commonly used in high-dimensional statistics, in which sparsity and incoherence properties of the basis functions $\phi_1, \phi_2, \ldots$ play a crucial role (e.g., Candes and Tao (2007)).

REMARK 3. The choices we make for $\hat{m}$ and $\mathcal{F}$ reflect assumptions about the regression function m. In addition to nonparametric assumptions like smoothness, we may make parametric or semiparametric assumptions. A semiparametric assumption distinguishes Examples 2 and 3, which consider the average partial effect for arbitrary functions $m(x, w)$ and for functions of the form $m(x, w) = \mu(x) + w\tau(x)$, respectively.

In the latter case, which we discuss in detail in Section 3, the tangent space $\overline{\text{span}}\mathcal{F}$ is smaller than the space of all square integrable functions, and the Riesz representer $\gamma_{\mathcal{F}}$ for $\psi(\cdot)$ will be the orthogonal projection onto $\overline{\text{span}}\mathcal{F}$ of the Riesz representer γ_{L_2} for $\psi(\cdot)$ on the tangent space of all square-integrable functions. An important consequence is that, under our efficiency condition $v\gamma_\psi \in \overline{\mathcal{S}}$, the optimal asymptotic variance in Example 3 is smaller than that in Example 2.[4] This reflects the ease of estimating the average partial effect in the conditionally linear model relative to the general case.

Naturally, such an estimator will be considered superefficient if we entertain the possibility that $m(x, w)$ does not have the form $\mu(x) + w\tau(x)$, that is, if our regularity condition $\mathcal{S} \subseteq \overline{\text{span}}\mathcal{F}$ is not satisfied. In this case, our weights fail to adjust for the deviation $\hat{m} - m$ for some possible regression function $m \in \mathcal{S}$ in a neighborhood of $\hat{m}$, and any gain in efficiency possible by doing so is, in a local minimax sense, spurious. Characterization of the behavior of our estimator under this form of misspecification is important but beyond the scope of this paper.

This phenomenon is not unique to our approach; for additional discussion of the choice of tangent space when estimating a Riesz representer see, for example, Remark 2.5 of Chernozhukov et al. (2018) and Section 3 of Robins et al. (2007). It pervades the literature on inference in high dimensional statistics, which typically involves an estimate of the Riesz representer on an appropriate tangent space of high-dimensional parametric functions (e.g., Athey, Imbens and Wager (2018), Javanmard and Montanari (2014), Zhang and Zhang (2014)). For example, when estimating a mean with outcomes missing at random in a high-dimensional linear model $m(x, w) = wx^T\beta$, γ_ψ is the best linear-in-x approximation to the inverse propensity weights $w/e(x)$.

REMARK 4. Our assumption that $\psi(\cdot)$ has a square-integrable Riesz representer γ_ψ, equivalent to its mean-square continuity, is necessary in the sense that $\psi(m)$ does not have a regular estimator when it is violated (Theorem 2.1 van der Vaart (1991), see Section B.1.2 here for details). If $\mathcal{F}$ has a finite uniform entropy integral, it is also sufficient. Theorem 1 requires no additional conditions on γ_ψ because under this condition on $\mathcal{F}$, the square integrability of γ_ψ implies our condition that $\gamma_\psi\mathcal{F}$ is Donsker (van der Vaart and Wellner (1996), Example 2.10.23).

In the context of Example 1, in which $\gamma_\psi(x, w)$ is the inverse probability weight $w/e(x)$ for $e(x) = P[W_i = 1 \mid X_i = x]$, this means that all we require of $e(x)$ is that $\mathbb{E}\gamma_\psi^2(X_i, W_i) =$

[4]The difference in asymptotic variance between estimators using weights converging to γ_{L_2} (Example 2) and weights converging to $\gamma_{\mathcal{F}}$ (Example 3) is $\mathbb{E}v(Z)[\gamma_{L_2}^2(Z) - \gamma_{\mathcal{F}}^2(Z)] = \mathbb{E}v(Z)[\gamma_{L_2}(Z) - \gamma_{\mathcal{F}}(Z)]^2 + 2\mathbb{E}v(Z)\gamma_{\mathcal{F}}(Z)[\gamma_{L_2}(Z) - \gamma_{\mathcal{F}}(Z)]$. The first term in this decomposition is positive and the second term is zero *if* $v\gamma_{\mathcal{F}} \in \overline{\text{span}}\mathcal{F}$, as in this case $\mathbb{E}\gamma_{L_2}(Z)[v(Z)\gamma_{\mathcal{F}}(Z)] = \psi(v\gamma_{\mathcal{F}}) = \mathbb{E}\gamma_{\mathcal{F}}(Z)[v(Z)\gamma_{\mathcal{F}}(Z)]$.

$\mathbb{E}1/e(X_i) < \infty$. D'Amour et al. (2021) highlights the need for a weak condition like this, showing that the usual "strict overlap" condition that $e(x)$ is bounded away from zero implies strong constraints on the conditional distribution of $X_i \mid W_i$. Chen, Hong and Tarozzi (2008) discusses the estimation of parameters defined by nonlinear moment conditions using overlap assumptions comparable to what we use here.

In simulation settings in which $\gamma_\psi(Z_i)$ has a spiky distribution, our estimator sometimes outperforms a double robust oracle estimator that weights using the true Riesz representer γ_ψ, while a typical double robust estimator performs substantially worse than this oracle estimator. This suggests that common responses to limited overlap, like changing the estimand (e.g., Crump et al. (2009), Li, Morgan and Zaslavsky (2018)) or assuming a semiparametric model as in Remark 3, may not be needed as frequently with our approach.

REMARK 5. Although we assume no regularity conditions on the Riesz representer γ_ψ, our weights $\hat{\gamma}_i$ still estimate it consistently. This is a universal consistency result, in line with well-known results about k-nearest neighbors regression and related estimators (Lugosi and Zeger (1995), Stone (1977)). Heuristically, the reason for this phenomenon is that the Riesz representer γ_ψ is the unique[5] weighting function that sets a population-analogue of $I_{h,\mathcal{F}}$ to 0; because $\hat{\gamma}$ comes close to doing the same, it must also approximate γ_ψ. This universal consistency property is not what controls the bias of our estimator $\hat{\psi}$. In fact, the rate of convergence of $\hat{\gamma}_i$ to $\gamma_\psi(X_i)$ is in general too slow for standard arguments for plug-in estimators to apply. However, it plays a key role in understanding why we get efficiency under heteroskedasticity even though we choose our weights by solving an optimization problem (8) that is not calibrated to the conditional variance structure of Y_i.

To understand this phenomenon, observe that under the conditions of Theorem 1, the conditional bias term $n^{-1}\sum_{i=1}^n h(Z_i, \hat{m} - m) - \hat{\gamma}_i(\hat{m}(Z_i) - m(Z_i))$ in our error is $o_P(n^{-1/2})$. It is therefore unnecessary to make an optimal bias-variance tradeoff by this sort of calibration to get efficiency under heteroskedasticity and heteroskedasticity-robust confidence intervals; the asymptotic behavior of our estimator is determined by the asymptotic behavior of our noise term $n^{-1}\sum_{i=1}^n \hat{\gamma}_i \varepsilon_i$ and, therefore, by the limiting weights $\gamma_\psi(Z_i)$.

For the same reason, it is not necessary to know the error scale $\|\hat{m} - m\|_{\mathcal{F}}$ to form asymptotically valid confidence intervals. We stress that this is an asymptotic statement; in finite samples, there are strong impossibility results for uniform inference that is adaptive to the scale of an unknown signal (Armstrong and Kolesár (2018)). Furthermore, tuning approaches that estimate and incorporate individual variances σ_i into the minimax weighting problem (8) like those discussed in Armstrong and Kolesár (2017) may offer some finite-sample improvement.

1.4. *Comparison with double-robust estimation.* Perhaps the most popular existing paradigm for building asymptotically efficient estimators in our setting is via constructions that first compute stand-alone estimates $\hat{m}(\cdot)$ and $\hat{\gamma}_\psi(\cdot)$ for the regression function and the Riesz representer, and then plug them into the following functional form (Chernozhukov et al. (2016), Newey (1994), Robins and Rotnitzky (1995)):

$$\hat{\psi}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^n \big[h(Z_i, \hat{m}) - \hat{\gamma}_\psi(Z_i)\big(\hat{m}(Z_i) - Y_i\big)\big], \tag{12}$$

[5]This uniqueness is violated when the tangent space $\overline{\operatorname{span}}\mathcal{F}$ that ψ acts on is not the space of all square integrable functions. However, the dual characterization Lemma 2 shows that our weights must converge to a function in this tangent space, and it follows that they converge to the unique Riesz representer γ_ψ on this tangent space.

or an asymptotically equivalent expression (e.g., van der Laan and Rubin (2006)). This estimator has a long history in the context of many specific estimands, for example, the aforementioned AIPW estimator for the estimation of a mean with outcomes missing at random (Cassel, Särndal and Wretman (1976), Robins, Rotnitzky and Zhao (1994)). In recent work, Chernozhukov, Newey and Robins (2018) describe a general approach of this type, making use of a novel estimator for the Riesz representer of a functional γ_ψ in high dimensions motivated by the Dantzig selector of Candes and Tao (2007).

In considerable generality, this estimator $\hat{\psi}_{\mathrm{DR}}$ is efficient when we use sample splitting[6] to construct $\hat{m}$ and these estimators satisfy (Chernozhukov et al. (2018), Zheng and van der Laan (2011))

$$\frac{1}{n}\sum_{i=1}^{n}\big[\hat{\gamma}_\psi(Z_i)-\gamma_\psi(Z_i)\big]\big[\hat{m}(Z_i)-m(Z_i)\big]=o_P\big(n^{-1/2}\big). \tag{13}$$

Taking the Cauchy–Schwarz bound on this bilinear form results in a well-known sufficient condition on the product of errors, $\|\hat{\gamma}_\psi-\gamma_\psi\|_{L_2(P_n)}\|\hat{m}-m\|_{L_2(P_n)}=o_P(n^{-1/2})$. This phenomenon, that we can trade off accuracy in how well the two nuisance functions m and γ_ψ are estimated, is called *double-robustness*.

While the estimator $\hat{\psi}_{\mathrm{AML}}$ defined in (7) shares the form of $\hat{\psi}_{\mathrm{DR}}$, it is not designed to be double robust. The weights $\hat{\gamma}$ used in $\hat{\psi}_{\mathrm{AML}}$ are optimized for the task of correcting the error of the plugin estimator $\psi(\hat{m})$ when our assumptions on the regression error function $\hat{m}-m$ are correct. When this is the case and the class $\mathcal{F}$ characterizing our uncertainty about this function is sufficiently small (e.g., Donsker), this allows us to be completely robust to the difficulty of estimating the Riesz representer γ_ψ. Our estimator will be efficient essentially because the error $\hat{\gamma}-\gamma_\psi$ will be sufficiently orthogonal to all functions $f\in\mathcal{F}$ that (13) will be satisfied uniformly over the class of possible regression error functions $\hat{m}-m\in\mathcal{F}$. As the existence of an estimator $\hat{m}$ whose error $\hat{m}-m$ is tight in the gauge of some Donsker class $\mathcal{F}$ is equivalent to the existence of an $o_P(n^{-1/4})$-consistent estimator of m, relative to the aforementioned sufficient condition on the product of error rates, this characterization completely eliminates regularity requirements on the Riesz representer γ_ψ while requiring the same level of regularity on the regression function m.

This type of phenomenon is not unique to our approach. The higher order influence function estimator of Robins et al. (2017) is efficient under the minimal Hölder-type smoothness conditions on γ_ψ and m. This includes the case where either m or γ_ψ admits an $o_P(n^{-1/4})$-consistent estimator with no conditions on the other, as well as possibilities interpolating these in which neither does (Robins et al. (2009)). Furthermore, Newey and Robins (2018) show that, if $\hat{m}$ and $\hat{\gamma}_\psi$ are appropriately tuned series estimators fit using a three-way cross-fitting scheme, $\hat{\psi}_{\mathrm{DR}}$ is efficient under minimal or nearly minimal Hölder-type smoothness conditions. They also show that for this $\hat{m}$, a cross-fit plug-in estimator $n^{-1}\sum_{i=1}^{n}h(Z_i,\hat{m})$ will be efficient if m is Hölder-smooth enough to admit an $o_P(n^{-1/4})$-consistent estimator, and beyond this regime exhibits some double robustness—it is also efficient when m is less smooth and γ_ψ is smooth enough.

The use of undersmoothed, that is, less biased than variable, nuisance estimators seems to be an important ingredient in estimators that beat the error rate product bound (see also Kennedy (2020), van der Laan, Benkeser and Cai (2019)). Both here and in Newey and

[6]In particular, this result holds if we use the cross-fitting construction of Schick (1986), where separate data folds are used to estimate the nuisance components $\hat{m}$ and $\hat{\gamma}_\psi$ and to compute the expression (12) given those estimates. The three-way sample splitting scheme of Newey and Robins (2018), discussed below, refines this by using different folds to estimate the two nuisance functions, and the remaining ones to compute the expression (12).

Robins (2018), γ_ψ is estimated by solving a set of Riesz representer estimating equations (5) subject to weak regularization or constraints. Furthermore, when $\mathcal{F}$ is a ball in a reproducing kernel Hilbert space, the minimax linear estimator ($\hat{\psi}_{\mathrm{AML}}$ with $\hat{m} \equiv 0$) is equivalently described as a plug-in using a undersmoothed ridge regression estimator $\hat{m}$ (Kallus (2020), Theorem 22). Hirshberg, Maleki and Zubizarreta (2019) show that this estimator is efficient essentially whenever $\|m\|_{\mathcal{F}} < \infty$.

1.5. *Comparison with minimax linear and balancing estimators.* As discussed above, our approach is primarily motivated as a refinement of conditional-on-design minimax linear estimators as developed and studied by a large community over the past decades (e.g., Donoho (1994), Ibragimov and Khas'minskiĭ (1984), Juditsky and Nemirovski (2009)); however, our focus is on its behavior in a random-design setting, as in the literature on semiparametrically efficient inference and local asymptotic minimaxity, including results on doubly robust methods (e.g., Bickel et al. (1998), Robins and Rotnitzky (1995), van der Laan and Rubin (2006)). The conceptual distinction between these two settings is strong in causal inference and missing data problems, where in the former we consider an adversary that chooses $m(\cdot)$ having observed the realized covariates and pattern of missing data, and in the latter we consider an adversary that chooses $m(\cdot)$ having observed no part of the realized data.

We are aware of three estimators that can be understood as special cases of our augmented minimax linear estimator (7). In the case of parameter estimation in high-dimensional linear models, Javanmard and Montanari (2014) propose a type of debiased lasso that combines a lasso regression adjustment with weights that debias the ℓ_1-ball, a convex class known to capture the error of the lasso; Athey, Imbens and Wager (2018) develop a related idea for average treatment effect estimation with high-dimensional linear confounding; and Kallus (2018, 2020) proposes analogs for treatment effect estimation and policy evaluation, a special case of Example 4, that adjust for nonparametric confounding using weights that debias the unit ball of a reproducing kernel Hilbert space. The contribution of our paper relative to this line of work lies in the generality of our results, and also in characterizing the asymptotic variance of the estimator under heteroskedasticity and proving efficiency in the fixed-dimensional nonparametric setting. Given heteroskedasticity, the aforementioned papers prove $\sqrt{n}$-consistency but do not characterize the asymptotic variance directly in terms of the distribution of the data; instead, they express the variance in terms of the solution to an optimization problem analogous to (8).

In the special case of mean estimation with outcomes missing at random, the optimization problem (8) takes on a particularly intuitive form, with

$$I_{h,\mathcal{F}}(\gamma) = \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - W_i \gamma_i) f(X_i, 1) \right\} \tag{14}$$

measuring how well the γ-weighted average of $f(x, 1)$ over the units with observed outcomes matches its average over everyone. In other words, the minimax linear weights enforce "balance" between these subsamples, which has been emphasized as fundamental to this problem by several authors including Rosenbaum and Rubin (1983) and Hirano, Imbens and Ridder (2003). Recently, there has been considerable interest in the use of balancing weights, chosen to control $I_{h,\mathcal{F}}$ or a variant, in linear estimators and in augmented linear estimators (7) like those we consider here (Athey, Imbens and Wager (2018), Chan, Yam and Zhang (2016), Graham, De Xavier Pinto and Egel (2012), Graham, Pinto and Egel (2016), Hainmueller (2012), Imai and Ratkovic (2014), Kallus (2020), Ning, Peng and Imai (2017), Wang and Zubizarreta (2017), Wong and Chan (2018), Zhao (2019), Zubizarreta (2015)). In addition to generalizing beyond the missing-at-random problem, our Theorem 2 provides the sharpest results we are aware of for balancing-type estimators in this specific problem.

To do this, we bring together arguments from two strands of the balancing literature. The first focuses on balancing small finite-dimensional classes, and in several instances it has been shown that when tuned so that $I_{h,\mathcal{F}}(\hat{\gamma})$ is sufficiently small, the linear estimator is efficient under strong assumptions on both m and γ_ψ (Chan, Yam and Zhang (2016), Fan et al. (2016), Graham, De Xavier Pinto and Egel (2012), Wang and Zubizarreta (2017)). The arguments used to establish these results rely on the convergence of $\hat{\gamma}$ to γ_ψ at sufficient rate, much like those used with the estimators discussed in the previous section. The second focuses on balancing high or infinite-dimensional classes, and in several instances it has been shown that when tuned so that $I_{h,\mathcal{F}}(\hat{\gamma}) = O_P(n^{-1/2})$, a level of balance that is attainable under assumptions comparable to ours, the linear estimator is $\sqrt{n}$-consistent and the augmented linear estimator is $\sqrt{n}$-consistent and asymptotically unbiased (Athey, Imbens and Wager (2018), Kallus (2020), Wong and Chan (2018)). The arguments used to establish these results fundamentally rely on balance to bound the estimator's bias, and do not fully characterize the estimator's asymptotic distribution. Our argument is a refinement of this one, using balance to do the bulk of the work, but relying on the convergence of the balancing weights $\hat{\gamma}$ to γ_ψ to characterize the asymptotic distribution of our estimator and to establish asymptotic unbiasedness under weaker conditions.

2. Estimating linear functionals. In this section, we give a more general characterization of the behavior of our estimator. We begin by sketching our argument, which is based on a decomposition of our estimator's error into a bias-like term and a noise-like term. We consider error relative to a sample-average version of our estimand, $\tilde{\psi}(m) = n^{-1}\sum_{i=1}^n h(Z_i, m)$, as the difference $\psi(m) - \tilde{\psi}(m)$ is out of our hands:

$$\begin{aligned}\hat{\psi}_{\mathrm{AML}} - \tilde{\psi}(m) &= \frac{1}{n}\sum_{i=1}^n h(Z_i, \hat{m}) - \hat{\gamma}_i\big(\hat{m}(Z_i) - Y_i\big) - h(Z_i, m) \\ &= \frac{1}{n}\sum_{i=1}^n \underbrace{h(Z_i, \hat{m} - m) - \hat{\gamma}_i(\hat{m} - m)(Z_i)}_{\text{bias}} + \underbrace{\hat{\gamma}_i\big(Y_i - m(Z_i)\big)}_{\text{noise}}.\end{aligned} \tag{15}$$

In Appendix A, we prove finite sample bounds on the bias term and the difference between the noise term and that of the oracle estimator with weights $\gamma_\psi(Z_i)$. Our estimator will be asymptotically linear, with the influence function of the oracle estimator, if both of these quantities are $o_p(n^{-1/2})$. We establish these bounds in three steps.

Step 1. We bound $n^{-1}\sum_{i=1}^n(\hat{\gamma}_i - \gamma_i^\star)^2$ for $\gamma_i^\star = \gamma_\psi(Z_i)$. To do this, we work with a dual characterization of our weights $\hat{\gamma}_i$ as evaluations $\hat{\gamma}_\psi(Z_i)$ of a penalized least squares estimate of the Riesz representer γ_ψ:

$$\begin{aligned}\hat{\gamma}_\psi &= \operatorname*{argmin}_g\left\{\|g\|_{L_2(P_n)}^2 - \frac{2}{n}\sum_{i=1}^n h(Z_i, g) + \frac{\sigma^2}{n}\|g\|_{\mathcal{F}}^2\right\} \\ &= \operatorname*{argmin}_g\left\{\|g - \gamma_\psi\|_{L_2(P_n)}^2 - \frac{2}{n}\sum_{i=1}^n h_{\gamma_\psi}(Z_i, g) + \frac{\sigma^2}{n}\|g\|_{\mathcal{F}}^2\right\},\end{aligned} \tag{16}$$

where $h_\gamma(z, f) = h(z, f) - \gamma(z)f(z)$. Here, the term involving h_{γ_ψ} plays the role of "noise" in our least squares problem, as it has mean zero for any function $f \in \overline{\operatorname{span}}\mathcal{F}$. The first characterization is established using strong duality in Lemma 2 and the second is derived by completing the square.

Step 2. We bound the difference between our noise term and that of the oracle estimator, $n^{-1}\sum_{i=1}^n(\hat{\gamma}_i - \gamma_i^\star)(Y_i - m(Z_i))$, using the result of Step 1.

Step 3. We bound our bias term by $\|\hat{m} - m\|_{\mathcal{F}} I_{h,\mathcal{F}}(\hat{\gamma})$, where as a consequence of the definition of our weights $\hat{\gamma}$ in (8),

$$I_{h,\mathcal{F}}^2(\hat{\gamma}) \leq I_{h,\mathcal{F}}^2(\gamma^\star) + \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\left(\gamma_i^{\star 2} - \hat{\gamma}_i^2\right). \tag{17}$$

The first term on the right-hand side can be characterized using empirical process techniques, as $I_{h,\mathcal{F}}(\gamma^\star)$ is the supremum of the empirical measure indexed by the class of mean-zero functions $h_{\gamma_\psi}(\cdot, \mathcal{F})$. And the second term can be shown, using some simple arithmetic, to be $o_p(n^{-1})$ when $\hat{\gamma}$ is consistent. Thus, our bias term will be bounded by $\|\hat{m} - m\|_{\mathcal{F}}[I_{\mathcal{F}}(\gamma^\star) + o_p(n^{-1/2})]$.

Step 3'. We refine this bound to take advantage of the consistency of $\hat{m}$. To do this, we show that our estimator behaves essentially the same way as an oracle that knows a sharp bound $\|\hat{m} - m\|_{L_2(P_n)} \leq \rho$ on our regression error and uses a refined model class $\mathcal{F}'_\rho = \{f : \|f\|_{\mathcal{F}}^2 + \rho^{-2}\|f\|_{L_2(P_n)}^2 \leq 1\}$ in place of $\mathcal{F}$. The key insight is that this substitution changes the dual (16) and its solution $\hat{\gamma}$ very little, so replacing $\mathcal{F}$ with $\mathcal{F}'_\rho$ in our bound (17) yields an inequality that is approximately satisfied. Given the assumptions of Theorem 1, the resulting refined bias term bound will be $o_p(n^{-1/2})$, as $\|\hat{m} - m\|_{\mathcal{F}'_\rho} = O_p(1)$ for $\rho \to 0$ given our tightness and consistency assumptions (10) and $I_{h,\mathcal{F}'_\rho}(\gamma^\star) = o_p(n^{-1/2})$ when $\rho \to 0$ given our Donskerity and equicontinuity assumptions.

We will now state our main result. Due to space constraints, all proofs are in the Appendices.

Definitions. To characterize the size of a set $\mathcal{G}$, we will use its *Rademacher complexity*, $R_n(\mathcal{G}) = \mathbb{E}\sup_{g\in\mathcal{G}}|n^{-1}\sum_{i=1}^n \epsilon_i g(Z_i)|$ where $\epsilon_i = \pm 1$ each with probability 1/2 independently and independently of the sequence $Z_1 \ldots Z_n$, as well as the uniform bound $M_\infty(\mathcal{G}) = \sup_{g\in\mathcal{G}}\|g\|_\infty$. Letting $h_\gamma(z, f) = h(z, f) - \gamma(z)f(z)$, our bound depends on the Rademacher complexity of the classes $\mathcal{F}_r$, $h_{\gamma_\psi}(\cdot, \mathcal{F}_r)$, and $h_{\tilde{\gamma}}(\cdot, \mathcal{F}_r)$ for a regularized approximation $\tilde{\gamma}$ to γ_ψ. The regularity of that approximation and, therefore, the regularity of γ_ψ itself, will be a factor in a higher order term. Without loss of generality, we will write our weights as function evaluations $\hat{\gamma}_i = \hat{\gamma}(Z_i)$, and we will write $a \vee b$ and $a \wedge b$, respectively, for the maximum and minimum of a and b and $a \lesssim b$ and $a \ll b$ meaning $a = O(b)$ and $a = o(b)$.

THEOREM 2. *In the setting described in Section 1.3, consider the estimator $\hat{\psi}_{\mathrm{AML}}$ defined in (7) with $\sigma > 0$ and $\mathcal{F}$ a uniformly bounded absolutely convex set of functions for which $h(\cdot, \mathcal{F})$ is pointwise bounded. Let γ_ψ be the Riesz representer of ψ on the tangent space $\overline{\operatorname{span}}\mathcal{F}$ and $\tilde{\gamma}$ minimize $\|\gamma_\psi - \gamma\|_{L_2(Q)}^2 + (\sigma^2/n)\|\gamma\|_{\mathcal{F}}^2$ for $Q = P$ or $Q = P_n$. If $\mathcal{F}$ is $\|\cdot\|_{L_2(Q)}$-closed, this argmin exists and is unique, and for any positive δ, on the intersection of an event of probability $1 - 4\delta - 3\exp(-c_2 n r_Q^2/M_\infty^2(\mathcal{F}))$ and one on which $\|\hat{m} - m\|_{\mathcal{F}} \leq s_{\mathcal{F}}$ and $\|\hat{m} - m\|_{L_2(P_n)} \leq s_{L_2(P_n)}$,*

$$\begin{aligned}
&\|\hat{\gamma} - \tilde{\gamma}\|_{L_2(P_n)}^2 \leq 6\left(nr^4/\sigma^2 + \|\tilde{\gamma}\|_{\mathcal{F}} r^2\right) \vee 8r^2 \quad \text{for } r = r_Q \vee r_M,\\
&r_Q = \inf\left\{r > 0 : R_n(\mathcal{F}_{c_0 r}) \leq c_1 r^2/M_\infty(\mathcal{F})\right\},\\
&r_M = \begin{cases} \inf\left\{r > 0 : R_n\left(h_{\tilde{\gamma}}(\cdot, \mathcal{F}_r)\right) \leq \delta r^2/2\right\} & \text{for } Q = P,\\ \inf\left\{r > 0 : R_n\left(h_{\gamma_\psi}(\cdot, \mathcal{F}_r)\right) \leq \delta r^2/2\right\} & \text{for } Q = P_n, \end{cases}
\end{aligned} \tag{18}$$

and for $\iota_\gamma(y,z) = h(z,m) - \gamma(z)(m(z)-y) - \psi(m)$ *and any positive* $\epsilon \le 9/16$,

$$
\begin{aligned}
&\sqrt{n}\left|\hat{\psi}_{\mathrm{AML}} - \psi(m) - n^{-1}\sum_{i=1}^{n}\iota_{\tilde{\gamma}}(Y_i,Z_i)\right| \\
&\quad\le (1/\sqrt{\delta})\|v\|_\infty \|\hat{\gamma}-\tilde{\gamma}\|_{L_2(P_n)} \\
&\qquad + \sqrt{2n}s_{\mathcal{F}}\phi\left(\frac{s_{L_2(P_n)}}{s_{\mathcal{F}}} \vee c_0 r \vee \frac{6\sigma}{\epsilon\sqrt{n}}\right)\left(1 + 2\epsilon/\sqrt{1-\epsilon^2/36}\right) \\
&\qquad + \sqrt{2}\sigma s_{\mathcal{F}}\left(\|\gamma_\psi\|_{L_2(P_n)} \wedge \|\gamma_\psi\|_{L_2(P_n)}^{1/2}\|\hat{\gamma}-\gamma_\psi\|_{L_2(P_n)}^{1/2}\right)/\sqrt{1-\epsilon^2/36}.
\end{aligned} \tag{19}
$$

Here, $c_0 \dots c_2$ *are universal constants and*

$$
\phi(\rho) = \frac{2R_n(h_{\gamma_\psi}(\cdot,\mathcal{F}_{\sqrt{2}\rho}))}{\delta} \vee \frac{216}{\epsilon^2}\left(r^2 + \frac{\sigma^2\|\tilde{\gamma}\|_{\mathcal{F}}}{n}\right) \vee \frac{36\sigma^2\|\gamma_\psi\|_{L_2(P)}}{\epsilon^2\sqrt{\delta}c_0 n r} \vee \frac{288\sigma^2}{\epsilon^2 n}.
$$

Generalization to classes $\mathcal{F}$ that are not uniformly bounded is discussed in Appendix A.6.

In the asymptotic setting we considered in the Introduction, in which the distribution P, the class $\mathcal{F}$ and the tuning parameter σ are fixed, this result implies Theorem 1. The key steps of the proof follow. We use the bound above for $Q = P_n$ and the bound $\|\tilde{\gamma}\|_{\mathcal{F}} \le (\sqrt{n}/\sigma)\|\gamma_\psi\|_{L_2(P_n)}$, which holds because $\|\gamma-\gamma_\psi\|_{L_2(P_n)}^2 + (\sigma^2/n)\|\gamma\|_{\mathcal{F}}^2$ is smaller at its minimizer than at $\gamma = 0$.

1. As γ_ψ is fixed, the regularized approximation $\tilde{\gamma}$ converges to γ_ψ in $\|\cdot\|_{L_2(P_n)}$ as the weight of regularization $\sigma^2/n \to 0$, so our "influence function" $\iota_{\tilde{\gamma}}$ converges to the limit ι_{γ_ψ}.

2. Given our tightness and consistency assumptions (10), we can take $s_{\mathcal{F}} \ge \|\hat{m}-m\|_{\mathcal{F}}$ to be of constant order and $s_{L_2(P_n)} \ge \|\hat{m}-m\|_{L_2(P_n)}$ to be converging to zero on a high probability event. Thus, our remainder bound (19) goes to zero if $\sqrt{n}\phi(s_n) \to 0$ for any sequence s_n converging to zero and $r \ll n^{-1/4}$ and, therefore, $\|\hat{\gamma}-\tilde{\gamma}\|_{L_2(P_n)} \to 0$ (via (18)).

3. Both of these conditions hold if $\lim_{t\to 0}\sqrt{n}R_n(\mathcal{F}_t) = \lim_{t\to 0}\sqrt{n}R_n(h_{\gamma_\psi}(\cdot,\mathcal{F}_t)) = 0$. The first limit is zero because $\mathcal{F}$ is Donsker. And the second is zero for the same reason, as $h_{\gamma_\psi}(\cdot,\mathcal{F}_t) \subseteq \mathcal{H}_{\omega(t)}$ where $\mathcal{H} = h_{\gamma_\psi}(\cdot,\mathcal{F})$ is Donsker and $\omega(t) = \sup_{f\in\mathcal{F}_t}\|h_{\gamma_\psi}(\cdot,f)\|_{L_2(P)}$ satisfies $\lim_{t\to 0}\omega(t) = 0$ under our equicontinuity and uniform boundedness assumptions.

We generally recommend that the tuning parameter σ be chosen without consideration of sample size. The simple heuristic $\sigma^2 \approx \max_{i\le n}\mathrm{Var}[Y_i \mid Z_i]$ arises from the minimax interpretation of our estimator, in which σ^2 is a bound on the conditional variance.[7] However, $\hat{\psi}_{\mathrm{AML}}$ is fairly robust to our choice of σ, and Theorem 2 justifies a wide range of choices.

To consider the impact of σ, we look at the role it plays in the dual characterization (16) of our weights. As discussed above, this is a penalized least squares problem for estimating γ_ψ. From this perspective, taking σ to be of constant order is regularizing very weakly, and we can improve the rate of convergence of $\hat{\gamma}$ to our regularized approximation $\tilde{\gamma}$ by increasing σ. On the other hand, consideration of the primal (8) shows that this comes at a cost in terms of the maximal conditional bias $I_{h,\mathcal{F}}(\hat{\gamma})$, and if we have confidence that $\hat{m}-m$ is in a small class $\mathcal{F}$, we can decrease σ so that $I_{h,\mathcal{F}}(\hat{\gamma})$ and, therefore, our bias is zero or nearly zero. Recalling our discussion in Section 1.4, our choice of σ essentially trades off between two properties of the error $\hat{\gamma}_\psi - \gamma_\psi$: its degree of orthogonality to the specific functions in $\mathcal{F}$, and its degree of "orthogonality" to all square integrable functions, that is, its magnitude $\|\hat{\gamma}-\gamma_\psi\|_{L_2(P)}$.

[7] In our minimax framework in Section 1.1, we also assume that $\|\hat{m}-m\|_{\mathcal{F}} \le 1$. If we instead believe that $\|\hat{m}-m\|_{\mathcal{F}} \approx \alpha$, our heuristic suggests $\sigma^2 \approx \alpha^{-2}\max_{i\le n}\mathrm{Var}[Y_i \mid Z_i]$.

When we choose σ proportional to $\sqrt{n}r$, $\hat{\psi}_{\mathrm{AML}}$ is essentially a standard doubly robust estimator. Our estimate of γ_ψ is not undersmoothed as discussed in Section 1.4; with this tuning, if $\|\gamma_\psi\|_{\mathcal{F}} < \infty$, our weights converge to γ_ψ in empirical mean square at the rate r, typically the minimax rate for estimating γ_ψ satisfying $\|\gamma_\psi\|_{\mathcal{F}} < \infty$ (see Appendix B.2). The asymptotic linearity of $\hat{\psi}_{\mathrm{AML}}$ may then follow from the rate-product condition $\|\hat{\gamma}_\psi - \gamma_\psi\|_{L_2(P_n)}\|\hat{m} - m\|_{L_2(P_n)} = o_P(n^{-1/2})$, which is a sufficient condition when we use sample splitting to fit $\hat{m}$.[8] However, to improve our rate of convergence, we sacrifice orthogonality of $\hat{\gamma}_\psi - \gamma_\psi$ to possible realizations of $\hat{m} - m$ in $\mathcal{F}$. This makes our estimator sensitive to the rate of convergence of $\hat{m} - m$. We see this in our bound (19); the term proportional to σ will be large.

3. Estimating the average partial effect in a conditionally linear outcome model. As a concrete instance of our approach, we consider the problem of estimating an average partial effect, assuming a conditionally linear treatment effect model. A statistician observes features $X \in \mathcal{X}$, a treatment dose $W \in \mathbb{R}$, and an outcome $Y \in \mathbb{R}$ and wants to estimate ψ, where

$$\psi = \mathbb{E}\big[\tau(X)\big] \quad \text{assuming} \quad \mathbb{E}[Y|X = x, W = w] = \mu(x) + w\tau(x). \tag{20}$$

By Theorem 1, our AML estimator will be efficient for ψ under regularity conditions when $\mathrm{Var}[Y_i|X_i, W_i] = v(X_i)$ is only a function of X_i.

In the classical case of an unconfounded binary treatment, the model (20) is general and the estimand ψ corresponds to the average treatment effect (Imbens and Rubin (2015), Rosenbaum and Rubin (1983)). At the other extreme, if W is real valued but $\tau(x) = \tau$ is constrained not to depend on x, then (20) reduces to the partially linear model as studied by Robinson (1988). The specific model (20) has recently been studied by Athey, Tibshirani and Wager (2019), Graham and Pinto (2018) and Zhao, Small and Ertefaie (2017). We consider the motivation for (20) in Section 4 in the context a real-world application; here, we focus on estimating ψ in this model.

Both $\mu(\cdot)$ and $\tau(\cdot)$ in the model (20) are assumed to have finite gauge with respect to an absolutely convex class $\mathcal{H}$, and we define

$$\mathcal{F}_{\mathcal{H}} = \big\{m : m(x, w) = \mu(x) + w\tau(x), \|\mu\|_{\mathcal{H}}^2 + \|\tau\|_{\mathcal{H}}^2 \le 1\big\}. \tag{21}$$

We can simplify the definition (8) of the minimax weights for this class:

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma \in \mathbb{R}^n} \sup_{\mu \in \mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^n \gamma_i \mu(X_i)\right]^2 + \sup_{\tau \in \mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^n (W_i\gamma_i - 1)\tau(X_i)\right]^2 + \frac{\sigma^2\|\gamma\|^2}{n^2}. \tag{22}$$

Given these weights, the augmented minimax linear estimator is

$$\hat{\psi}_{\mathrm{AML}} = \frac{1}{n}\sum_{i=1}^n \big(\hat{\tau}(X_i) - \hat{\gamma}_i\big(\hat{\mu}(X_i) + W_i\hat{\tau}(X_i) - Y_i\big)\big). \tag{23}$$

Our formal results above give conditions under which it is asymptotically efficient. In this section, our goal is to explore the behavior of this estimator empirically. For comparison, we introduce some alternatives. The first is the minimax linear estimator $\hat{\psi}_{\mathrm{MLIN}} = n^{-1}\sum_{i=1}^n \hat{\gamma}_i Y_i$, that is, $\hat{\psi}_{\mathrm{AML}}$ with $\hat{m} \equiv 0$. The others are variants of the doubly robust estimator $\hat{\psi}_{\mathrm{DR}}$. In this setting, the Riesz representer has the form $\gamma_\psi(x, w) = (w - e(x))/v_w(x)$ with

[8] It is common to use sample splitting to fit $\hat{\gamma}_\psi$ as well. Our bound (18) does not justify this, as it concerns empirical mean squared error on the sample used to estimate $\hat{\gamma}_\psi$. However, in the course of our proof in Appendix A, we show that with this tuning, $\hat{\gamma}_\psi$ converges to γ_ψ in population mean square at the rate r, which is sufficient.

$e(x) = \mathbb{E}[W|X = x]$ and $v_w(x) = \mathrm{Var}[W|X = x]$, so we consider a natural doubly robust estimator based on plug-in estimates of these quantities,[9]

$$\hat{\psi}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) - \left(\frac{W_i - \hat{e}(X_i)}{\hat{v}_w(X_i)}\right)\left(\hat{\mu}(X_i) + W_i\hat{\tau}(X_i) - Y_i\right)\right). \tag{24}$$

Below, we numerically compare the relative merits of minimax linear, augmented minimax linear, and plug-in doubly robust estimation of the average partial effect.

3.1. *A simulation study.* To better understand the merits of different approaches to average partial effect estimation, we conduct a simulation study. As baselines, we consider the *plug-in doubly robust estimator* defined in (24), where $\hat{e}(\cdot)$ and $\hat{v}_w(\cdot)$ are fit separately, and an *oracle doubly robust estimator* that uses the same functional form (24) but with oracle values of $e(X_i)$ and $v_w(X_i)$. We compare these baselines to an *augmented minimax linear estimator* (AML) that uses minimax linear weights for a class $\mathcal{F}_{\mathcal{H}}$ as described in (23), as well as an *augmented minimax linear estimator over an extended class* (AML+), a variant that uses the same functional form but with the minimax linear weights for an extended class $\mathcal{F}_{\mathcal{H}_+}$ that includes a set of estimated functions. We also consider the simpler *minimax linear estimator* for each class. We provide further implementation details below.

3.1.1. *Construction of augmented minimax linear estimators.* We first describe how we implement our approach, an augmented minimax linear estimator for the class $\mathcal{F}_{\mathcal{H}}$ described in the section above (21). We take $\mathcal{H}$ to be the absolutely convex hull of a mean-square summable set of basis functions as described in Remark 2. Specifically, we use a basis sequence $\phi_j = a_j\phi'_j$, where ϕ'_j are d-dimensional interactions of Hermite polynomials that are orthonormal with respect to the standard normal distribution. The sequence of weights $\{a_j\}$ varies with order k of the polynomial ϕ_j; $a_j = 1/(k\sqrt{n_{k,d}})$ where $n_{k,d}$ is the number of terms of order k. Observe that $\sum_{j=1}^{\infty} a_j^2 = \sum_{k=1}^{\infty} 1/k^2 < \infty$ and therefore $\sum_{j=1}^{\infty}\mathbb{E}\phi_j^2(X) < \infty$ for standard normal X or X with bounded density with respect to the standard normal.

Following our discussion in Remark 2, we take an ℓ_1-penalized least squares approach to estimating the regression function m. Rather than using a fully nonparametric estimate $\hat{m}(x, w)$, which would not be in our class $\mathcal{F}_{\mathcal{H}}$, we fit a conditionally linear model $\hat{\mu}(x) + w\hat{\tau}(x)$ using the R-lasso method proposed by Nie and Wager (2017). To do this, we first estimate the marginal response function $r(x) = \mathbb{E}[Y_i|X_i = x]$ and $e(x)$ via a cross-validated ℓ_1-penalized regression (Tibshirani (1996)) on the basis $\phi(x)$. We then fit $\tau_\beta(x) = \phi(x)^T\beta$ by minimizing the ℓ_1-penalized R-loss $n^{-1}\sum_{i=1}^{n}[Y_i - \hat{r}(X_i) - (W - \hat{e}(X_i))\tau_\beta(X_i)]^2 + \lambda\|\beta\|_{\ell_1}$, with λ chosen by cross-validation. Finally, we set $\hat{\mu}(x) = \hat{r}(x) - \hat{\tau}(x)\hat{e}(x)$. As discussed in Nie and Wager (2017), this method is appropriate when the treatment effect function $\tau(x)$ is simpler than $r(x)$ and $e(x)$, and allows for faster rates of convergence on $\tau(x)$ than the other regression components whenever the nuisance components can be estimated at $o_p(n^{-1/4})$ rates in root-mean squared error.

We consider two options for the bias-correcting weights $\hat{\gamma}$. The simpler option is to use the minimax weights for the class $\mathcal{F}_{\mathcal{H}}$ described in (21). This choice is directly motivated by our formal results given in Theorem 1. As an alternative, motivated by popular idea of propensity-stratified estimation in the causal inference literature (Rosenbaum and Rubin (1984)), we use

[9]For example, a random forest version of this estimator is available in the `grf` package of Athey, Tibshirani and Wager (2019). In the binary treatment assignment case $W_i \in \{0, 1\}$, we know that $v_w(x) = e(x)(1 - e(x))$; and if we set $\hat{v}_w(x) = \hat{e}(x)(1 - \hat{e}(x))$, then the estimator in (24) is equivalent to the augmented inverse-propensity weighted estimator of Robins, Rotnitzky and Zhao (1994). For more general W_i, however, $v_w(x)$ is not necessarily determined by $e(x)$ and so we need to estimate it separately.

minimax weights for an extended class $\mathcal{F}_{\mathcal{H}_+}$ where $\mathcal{H}_+$ extends $\mathcal{H}$ by adding to our basis expansion $\phi(x)$ the following random basis functions:

- Multiscale strata of the estimated average treatment intensity $\hat{e}(X_i)$ (we balanced over histogram bins of width 0.05, 0.1 and 0.2),
- Basis elements obtained by depth-3 recursive dyadic partitioning (i.e., pick a feature, split along its median and recurse), and
- Leaves generated by a regression tree on the W_i (Breiman et al. (1984)).

The underlying idea is that we may be able to improve the practical performance of the method by opportunistically adding a small number of basis functions that help mitigate bias in case of misspecification (i.e., when μ and τ do not have finite gauge $\|\cdot\|_{\mathcal{H}}$). The motivation for focusing on transformations of $\hat{e}(X_i)$ is that accurately stratifying on $e(X_i)$ would suffice to eliminate all confounding in the model (20).[10] Because $\mathcal{F}_{\mathcal{H}+}$ is a function of $Z_1 \dots Z_n$ for $Z_i = (X_i, W_i)$, it is not necessary to cross-fit to avoid biasing the "noise term" in our error decomposition (15). With both $\mathcal{F}_{\mathcal{H}}$ and $\mathcal{F}_{\mathcal{H}+}$, we take $\sigma^2 = 1$ in (22).

3.1.2. *Baselines and software details.* The baselines we consider combine the aforementioned regression $\hat{\mu}(x) + w\hat{\tau}(x)$ with various weighting schemes. The weights used in the plug-in double robust estimator (24) involve $\hat{e}$ as estimated above and an estimate of $v_w(x) = \mathrm{Var}[W \mid X = x]$, which we fit by cross-validated ℓ_1-penalized regression of $(W_i - \hat{e}(X_i))^2$ on $\phi(X_i)$. The weights used in the double-robust oracle substitute the true values of $e(x)$ and $v_w(w)$ in our simulated design.

Tenfold cross-fitting is used throughout: where $\hat{\tau}(X_i)$ and $\hat{\mu}(X_i)$ appear in (23) and (24), we use estimators $\hat{\tau}^{(-i)}$ and $\hat{\mu}^{(-i)}$ trained on the folds that do not include unit i. This reduces dependence on (Y_i, X_i, W_i) and, therefore, mitigates potential own-observation bias in $\hat{\psi}_{\mathrm{DR}}$ (see, e.g., Chernozhukov et al. (2018)). However, we do get some dependence through the estimates of $\hat{r}$ and $\hat{e}$ used to train $\hat{\tau}$ and through ℓ_1-penalty tuning parameters, which are chosen once for all i by cross-validation. While this dependence could be eliminated using a computationally demanding nested sample splitting scheme, we here follow the approach taken in the `grf` package of Athey, Tibshirani and Wager (2019) and use a simplified scheme described in Appendix C. Our theoretical results for $\hat{\psi}_{\mathrm{AML}}$ do not formally justify the use of this cross-fitting scheme, as $\hat{m}^{(-i)}(x, w) = \hat{\mu}^{(-i)}(x) + w\hat{\tau}^{(-i)}(x)$ is a function of the fold indicator f_i as well as x, w, and for this reason $\|\hat{m} - m\|_{\mathcal{F}_{\mathcal{H}}} = \infty$; however, this does not seem to cause problems in our simulations.

All methods are implemented in the `R` package `amlinear`, and replication files are available at https://github.com/davidahirshberg/amlinear. We computed minimax linear weights via the cone solver `ECOS` (Domahidi, Chu and Boyd (2013)), available in `R` via the package `CVXR` (Fu et al. (2017)). When needed, we run penalized regression using the `R` package `glmnet` (Friedman, Hastie and Tibshirani (2010)).

3.1.3. *Simulation design.* We considered data-generating distributions of the form

$$X_i \sim \mathcal{N}(0, I_{d\times d}), \qquad W_i | X_i \sim \mathcal{L}_{X_i}, \qquad Y_i | X_i, \qquad W_i = \mathcal{N}\big(b(X_i) + W_i\tau(X_i), 1\big),$$

for different choices of dimension d, treatment assignment distribution $\mathcal{L}_{X_i}$, baseline main effect $\mu(\cdot)$ and treatment effect function $\tau(\cdot)$. We considered the following four setups, each of which depends on a sparsity level k that controls the complexity of the signal.

[10] In the case of binary treatments W_i, this corresponds to the classical result of Rosenbaum and Rubin (1983), who showed that the propensity score is a balancing score. With nonbinary treatments, $\mathbb{E}[W_i | X_i]$ is not in general a balancing score (Imbens (2000)); however, it is a balancing score for our specific model (20).

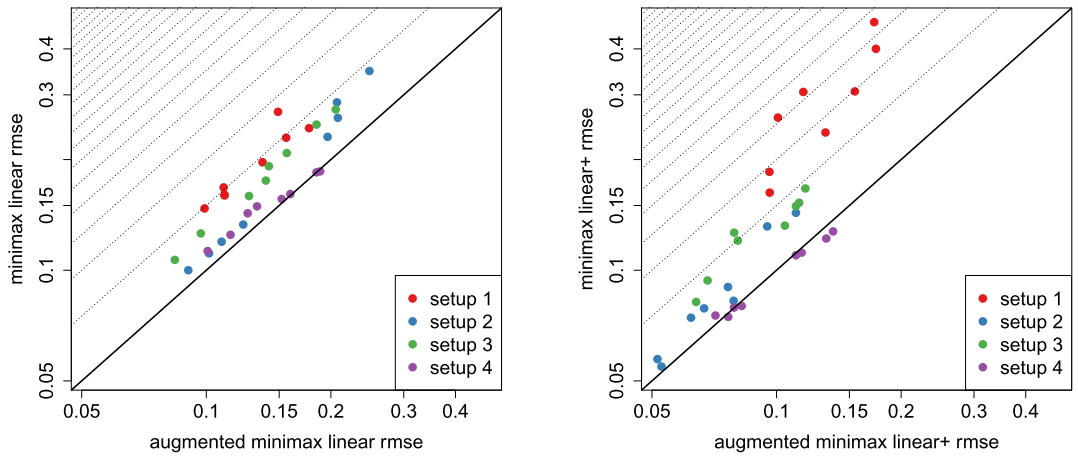

FIG. 1. *Comparing augmented minimax linear estimation with minimax linear estimation. The solid line $y = x$ indicates equivalent performance and the dotted lines indicate improvements of* 50%, 100%, 150%, *etc. in root mean squared error.*

1. Beta-distributed treatment, $W_i|X_i \sim B(\alpha(X_i), 1-\alpha(X_i))$, with $\zeta(x) = \sum_{j=1}^k x_j/\sqrt{k}$, $\eta(x) = \text{sign}(\zeta(x))\zeta^2(x)$, $\alpha(x) = \max\{0.05, \min\{0.95, 1/(1+\exp[-\eta(x)])\}\}$, $\mu(x) = \eta(x) + 0.2(\alpha(x) - 0.5)$, and $\tau(x) = -0.2$.
2. Scaled Gaussian treatment, $W_i|X_i \sim \mathcal{N}(\lambda(X_i), \lambda^2(X_i))$, with $\eta(x) = 2^{k-1}\prod_{j=1}^k x_j$, $\mu(x) = \text{sign}(\eta(x))\sqrt{|\eta(x)|}$, $\lambda(x) = 0.1\,\text{sign}(\mu(x)) + \mu(x)$, and $\tau(x) = \max\{x_1 + x_2, 0\}/2$.
3. Poisson treatment, $W_i|X_i \sim \text{Poisson}(\lambda(X_i))$, with $\tau(x) = k^{-1}\sum_{j=1}^k \cos(\pi x_j/3)$, $\lambda(x) = 0.2 + \tau^2(x)$, and $\mu(x) = 4d^{-1}\sum_{j=1}^d x_j + 2\lambda(x)$.
4. Log-normal treatment, $\log(W_i)|X_i \sim \mathcal{N}(\lambda(X_i), 1/3^2)$, with $\zeta(x) = \sum_{j=1}^k x_j/\sqrt{k}$, $\mu(x) = \max\{0, 2\zeta(x)\}$, $\lambda(x) = 1/(1+\exp[-\text{sign}(\zeta(x))\zeta^2(x)])$, and $\tau(x) = \sin(2\pi x_1)$.

3.2. *Results*. We first compare our augmented minimax linear estimators with the corresponding minimax linear estimators. Figure 1 compares the resulting mean-squared errors for ψ across several variants of the simulation design (the exact parameters used are the same as those used in Table 1). The left panel shows results where the weights are minimax over $\mathcal{F}_{\mathcal{H}}$, while the right panel has minimax weights over $\mathcal{F}_{\mathcal{H}_+}$.

Overall, we see that the augmented minimax linear estimator is sometimes comparable to the minimax linear one and sometimes substantially better. Thus, while results of Donoho (1994) and Armstrong and Kolesár (2018) imply that the augmented estimator can be little better than the minimax linear estimator for a convex signal class $\mathcal{F}$ in terms of its behavior at a few specific signals $m \in \mathcal{F}$, this does not appear representative of behavior in general. Furthermore, as the bias of our augmented estimator is bounded as a proportion of $\|\hat{m} - m\|_{\mathcal{F}}$ rather than $\|m\|_{\mathcal{F}}$, our approach offers a natural way to accommodate signals in some nonconvex signal classes: those for which, for some choice of $\hat{m}$, the regression error function $\hat{m} - m$ is well characterized in terms of some strong norm $\|\cdot\|_{\mathcal{F}}$. This can be the case, for example, when estimating a vector of regression coefficients β by ℓ_1-penalized regression: $\|\hat{\beta} - \beta\|_{\ell_1}$ will be small either if $\|\beta\|_{\ell_1}$ is small or, to a degree determined by incoherence properties of $\phi(X)$, if β is sparse (e.g., Lecué and Mendelson (2018)). This phenomenon offers some explanation for the good behavior we observe empirically, as the functions $\mu(x) = \phi(x)^T\beta_\mu$ and $\tau(x) = \phi(x)^T\beta_\tau$ defining our signal $m(x, w) = \mu(x) + w\tau(x)$ have some degree of sparsity and $\|\hat{m} - m\|^2_{\mathcal{F}_{\mathcal{H}}} = \|\hat{\beta}_\mu - \beta_\mu\|^2_{\ell_1} + \|\hat{\beta}_\tau - \beta_\tau\|^2_{\ell_1}$.

In Table 1, we compare augmented minimax linear estimation with doubly robust estimators, both using an estimated and an oracle Riesz representer. In terms of mean-squared

TABLE 1

Performance of four methods described in Section 3.1 on the simulation designs from Section 3.1.3. We report root-mean squared error, bias, and coverage of 95% confidence intervals averaged over 200 simulation replications

	Method			Double rob. plugin			Augm. minimax			Augm. minimax+			Double rob. oracle		
	n	p	k	rmse	bias	covg	rmse	bias	covg	rmse	bias	covg	rmse	bias	covg
setup 1	600	6	3	**0.13**	0.03	0.98	0.14	0.03	0.98	**0.13**	0.00	0.98	0.18	−0.01	0.96
	600	6	4	0.16	0.06	0.92	0.16	0.04	0.94	**0.15**	0.03	0.93	0.21	0.00	0.92
	600	12	3	0.22	0.09	0.78	0.18	−0.00	0.87	**0.17**	0.05	0.90	0.27	−0.04	0.90
	600	12	4	0.21	0.14	0.78	**0.15**	0.01	0.94	0.17	0.09	0.90	0.23	−0.03	0.93
	1200	6	3	**0.10**	0.03	0.94	0.11	0.06	0.92	**0.10**	0.02	0.96	0.12	0.00	0.98
	1200	6	4	0.11	0.03	0.94	0.11	0.05	0.92	**0.10**	0.02	0.96	0.13	0.00	0.94
	1200	12	3	0.11	0.02	0.90	**0.10**	0.01	0.95	**0.10**	0.02	0.94	0.14	0.00	0.94
	1200	12	4	0.15	0.06	0.86	**0.11**	0.00	0.92	0.12	0.04	0.90	0.16	−0.00	0.94
setup 2	600	6	1	0.15	0.12	0.52	0.11	0.09	0.74	**0.08**	0.02	0.94	0.09	0.00	0.92
	600	6	2	0.23	0.22	0.08	0.21	0.20	0.04	**0.09**	0.07	0.85	0.10	0.00	0.94
	600	12	1	0.16	0.14	0.44	0.12	0.11	0.62	**0.08**	0.03	0.93	0.08	0.00	0.98
	600	12	2	0.27	0.26	0.02	0.25	0.24	0.00	**0.11**	0.09	0.76	0.10	0.01	0.95
	1200	6	1	0.12	0.11	0.30	0.09	0.08	0.52	**0.05**	0.01	0.95	0.06	−0.00	0.96
	1200	6	2	0.20	0.20	0.00	0.20	0.19	0.00	**0.06**	0.04	0.90	0.06	−0.00	0.96
	1200	12	1	0.12	0.11	0.31	0.10	0.09	0.48	**0.05**	0.01	0.96	0.06	−0.00	0.98
	1200	12	2	0.22	0.22	0.00	0.21	0.20	0.00	**0.07**	0.04	0.86	0.07	0.00	0.94
setup 3	600	6	3	0.23	0.23	0.04	0.14	0.13	0.44	**0.11**	0.09	0.72	0.08	−0.00	0.96
	600	6	4	0.20	0.20	0.12	0.13	0.11	0.54	**0.10**	0.09	0.72	0.07	−0.00	0.96
	600	12	3	0.25	0.24	0.03	0.21	0.20	0.10	**0.12**	0.10	0.70	0.08	−0.01	0.95
	600	12	4	0.21	0.20	0.09	0.18	0.17	0.16	**0.11**	0.10	0.72	0.08	−0.01	0.94
	1200	6	3	0.20	0.19	0.01	0.10	0.09	0.55	**0.07**	0.05	0.78	0.05	−0.01	0.97
	1200	6	4	0.18	0.18	0.01	0.08	0.07	0.68	**0.06**	0.05	0.85	0.05	−0.01	0.96
	1200	12	3	0.23	0.22	0.00	0.16	0.15	0.02	**0.08**	0.07	0.76	0.05	−0.00	0.96
	1200	12	4	0.19	0.19	0.00	0.14	0.14	0.13	**0.08**	0.07	0.70	0.05	0.00	0.94
setup 4	600	6	4	0.22	0.16	0.84	0.16	−0.03	0.94	**0.11**	−0.02	1.00	0.16	0.03	0.94
	600	6	5	0.20	0.14	0.88	0.15	−0.05	0.93	**0.11**	−0.02	1.00	0.15	0.00	0.93
	600	12	4	0.23	0.15	0.86	0.18	−0.09	0.88	**0.14**	−0.04	0.96	0.17	−0.01	0.91
	600	12	5	0.24	0.17	0.82	0.19	−0.09	0.89	**0.13**	−0.05	0.97	0.17	−0.01	0.94
	1200	6	4	0.13	0.09	0.90	0.10	−0.03	0.94	**0.07**	−0.01	1.00	0.10	0.00	0.96
	1200	6	5	0.14	0.08	0.91	0.11	−0.05	0.94	**0.08**	−0.01	1.00	0.11	0.00	0.94
	1200	12	4	0.14	0.08	0.88	0.13	−0.07	0.88	**0.08**	−0.02	0.98	0.11	−0.00	0.94
	1200	12	5	0.14	0.09	0.87	0.13	−0.07	0.90	**0.08**	−0.02	1.00	0.11	−0.00	0.96

error, our simple AML estimator already performs well relative to the main baseline (i.e., plug-in doubly robust estimation), and the AML+ estimator does better yet. Perhaps more surprisingly, our methods sometimes also beat the doubly robust oracle, achieving comparable control of bias with a substantial decrease in variance. This reduction in variance arises from shrinkage due to the penalty term in (8). It costs us little bias then because, although the oracle weights must be large to control bias for all square integrable regression errors $\hat{m} - m$ (i.e., to solve (4)), large weights are not necessary to control bias for $\hat{m} - m$ in $\mathcal{F}$ (i.e., to solve (5)).

In terms of coverage, some of our simulation designs are extremely difficult and all nonoracle estimators have substantial relative bias. However, in settings 1 and 4, the asymptotics appear to kick in and our estimators get close to nominal coverage.

4. The effect of lottery winnings on earnings. To test the behavior of our method in practice, we revisit a study of Imbens, Rubin and Sacerdote (2001) on the effect of lottery

winnings on long-term earnings. It is of considerable policy interest to understand how people react to reliable sources of unearned income; such questions come up, for example, in discussing how universal basic income would affect employment. In an attempt to get some insight about this effect, Imbens, Rubin and Sacerdote (2001) study a sample of people who won a major lottery whose prize is paid out in installments over 20 years. The authors then ask how $1 in yearly lottery income affects the earnings of the winner.

To do so, the authors consider $n = 194$ people who all won the lottery, but got prizes of different sizes ($1000–$100,000 per year).[11] They effectively use a causal model $\mathbb{E}[Y_i(w) \mid X_i = x] = m(x) + \tau w$ for observations $Y_i = Y_i(W_i)$ of the average yearly earnings in the six years following winning W_i in yearly lottery payoff, where X_i denotes a set of $p = 12$ pre-win covariates (year won, number of tickets bought, age at win, gender, education, whether employed at time of win, earnings in six years prior to win). Here, $Y_i(w)$ represents the average yearly earnings that would have occurred had, possibly contrary to fact, unit i won a prize paying w dollars annually (e.g., Imbens and Rubin (2015)). The authors also consider several other model specifications.

As discussed at length by Imbens, Rubin and Sacerdote (2001), although the lottery winnings were presumably randomly assigned, we cannot assume exogeneity of the form $W_i \perp\!\!\!\perp \{Y_i(w) : w \in \mathbb{R}\}$ because of survey nonresponse. The data was collected by mailing out surveys to lottery winners asking about their earnings, etc., so there may have been selection effects in who responded to the survey. A response rate of 42% was observed, and older people with big winnings appear to have been relatively more likely to respond than young people with big winnings. For this reason, the authors only assume exogeneity conditionally on the covariates, that is, $W_i \perp\!\!\!\perp \{Y_i(w) : w \in \mathbb{R}\} | X_i$, which suffices to establish that the aforementioned causal model is identified as a regression model $m(x) + \tau w = \mathbb{E}[Y_i \mid X_i = x, W_i = w]$.

Here, we examine the robustness of the conclusions of Imbens, Rubin and Sacerdote (2001) to potential effect heterogeneity. Instead of assuming that the slope τ in this model is a constant, we let it vary with x and seek to estimate $\psi = \mathbb{E}[\tau(X)]$; this corresponds exactly to an average partial effect in the conditionally linear model, which we studied in Section 3. In our comparison, we consider 3 estimators that implicitly assume constant slope and estimate τ, and 6 that allow $\tau(x)$ to vary and estimate $\mathbb{E}[\tau(X)]$.

Among methods that assume constant slope, the first runs ordinary least squares for Y_i on W_i, ignoring potential confounding due to nonresponse. The second, which most closely resembles the method used by Imbens, Rubin and Sacerdote (2001), controls for the X_i using ordinary least squares, that is, it regresses Y_i on (X_i, W_i) and considers the coefficient on W_i. The third uses the method of Robinson (1988) with cross-fitting as in Chernozhukov et al. (2018): it first estimates the marginal effect of X_i on W_i and Y_i via a nonparametric adjustment and then regresses residuals $Y_i - \widehat{\mathbb{E}}[Y_i|X_i]$ on $W_i - \widehat{\mathbb{E}}[W_i|X_i]$. In each case, we report robust standard errors obtained via the `R`-package `sandwich` (Zeileis (2004)).

The six methods that allow for treatment effect heterogeneity correspond to the five methods discussed in Section 3, along with a pure weighting estimator using the estimated Riesz representer, $\hat{\psi} = n^{-1} \sum_{i=1}^{n} \hat{\gamma}_\psi(X_i) Y_i$, with the same choice of $\hat{\gamma}_\psi(\cdot)$ as used in (24). For all nonparametric regression adjustments, we run penalized regression as in Section 3, on a basis obtained by taking order-3 Hermite interactions of the 10 continuous features, and then creating full interactions with the two binary variables (gender and employment), resulting in a total of 1140 basis elements. For AML+, we include propensity strata of widths 0.05, 0.1 and 0.2 in the class $\mathcal{H}_+$.

[11] The paper also considers some people who won very large prizes (more than $100k per year) and some who won smaller prizes (not paid in installments); however, we restrict our analysis to the smaller sample of people who won prizes paid out in installments worth $1k–$100k per year.

TABLE 2

Various estimates, estimators, and estimands for the effect of unearned income on earnings, using the data set of Imbens, Rubin and Sacerdote (2001). The first three methods are justified under the assumption of no heterogeneity in $\tau(x)$ (i.e., $\tau(x) = \tau$), and estimate τ, while the latter six allow for heterogeneity and estimate $\mathbb{E}[\tau(X)]$

Estimand	Estimator	Estimate	std. err
partial effect	OLS without controls	−0.176	0.039
partial effect	OLS with controls	−0.106	0.032
partial effect	residual-on-residual OLS	−0.110	0.032
avg. partial effect	plug-in Riesz weighting	−0.175	—
avg. partial effect	doubly robust plugin	−0.108	0.042
avg. partial effect	minimax linear weighting	−0.074	—
avg. partial effect	augm. minimax linear	−0.091	0.044
avg. partial effect	minimax linear+ weighting	−0.083	—
avg. partial effect	augm. minimax linear+	−0.097	0.045

Table 2 reports results using the nine estimators described above, along with standard error estimates. We do not report standard errors for the three pure weighting methods, as these may not be asymptotically unbiased and so confidence intervals should also account for bias. The reported estimates are unitless; in other words, the majority of the estimators suggest that survey respondents on average respond to a $1 increase in unearned yearly income by reducing their yearly earnings by roughly $0.10.

Substantively, it appears reassuring that most point estimates are consistent with each other, whether or not they allow for heterogeneity in $\tau(x)$. The only two divergent estimators are the one that does not control for confounding at all, and the one that uses pure plug-in weighting (which may simply be unstable here). From a methodological perspective, it is encouraging that our method (and here, also the plug-in doubly robust method) can rigorously account for potential heterogeneity in $\tau(x)$ without excessively inflating uncertainty.

Acknowledgments. We are grateful for stimulating discussions with Timothy Armstrong, Vitor Hadad, Guido Imbens, Whitney Newey, Jamie Robins, Florian Stebegg and José Zubizarreta, as well as for comments from seminar participants at several venues. We also thank Guido Imbens for sharing the lottery data with us. We initiated this research while D.H. was a Ph.D. candidate at Columbia University and S.W. was visiting Columbia as a postdoctoral research scientist.

SUPPLEMENTARY MATERIAL

Appendices (DOI: 10.1214/21-AOS2080SUPP; .pdf). We provide complete proofs for the results in the main text, details about our simulation study, and a discussion of computational issues.

REFERENCES

ARMSTRONG, T. B. and KOLESÁR, M. (2018). Optimal inference in a class of regression models. *Econometrica* **86** 655–683. MR3783342 https://doi.org/10.3982/ECTA14434

ARMSTRONG, T. B. and KOLESÁR, M. (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica* **89** 1141–1177. MR4325179 https://doi.org/10.3982/ECTA16907

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CAI, T. T. and LOW, M. G. (2003). A note on nonparametric estimation of linear functionals. *Ann. Statist.* 1140–1153.

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63** 615–620. MR0445666 https://doi.org/10.1093/biomet/63.3.615

CHAN, K. C. G., YAM, S. C. P. and ZHANG, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 673–700. MR3506798 https://doi.org/10.1111/rssb.12129

CHEN, X., HONG, H. and TAROZZI, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.* **36** 808–843. MR2396816 https://doi.org/10.1214/009053607000000947

CHERNOZHUKOV, V., NEWEY, W. and ROBINS, J. (2018). Double/de-biased machine learning using regularized Riesz representers. Preprint. Available at arXiv:1802.08667.

CHERNOZHUKOV, V., ESCANCIANO, J. C., ICHIMURA, H. and NEWEY, W. K. (2016). Locally robust semiparametric estimation. Preprint. Available at arXiv:1608.00033.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144 https://doi.org/10.1093/biomet/asn055

D'AMOUR, A., DING, P., FELLER, A., LEI, L. and SEKHON, J. (2021). Overlap in observational studies with high-dimensional covariates. *J. Econometrics* **221** 644–654. MR4215042 https://doi.org/10.1016/j.jeconom.2019.10.014

DOMAHIDI, A., CHU, E. and BOYD, S. (2013). ECOS: An SOCP solver for embedded systems. In *European Control Conference* (*ECC*) 3071–3076.

DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082 https://doi.org/10.1214/aos/1176325367

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. III. *Ann. Statist.* **19** 668–701. MR1105839 https://doi.org/10.1214/aos/1176348114

FAN, J., IMAI, K., LIU, H., NING, Y. and YANG, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical Report, Princeton Univ.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

FU, A., NARASIMHAN, B., DIAMOND, S. and MILLER, J. (2017). CVXR: Disciplined Convex Optimization. R package version 0.94-4.

GRAHAM, B. S., DE XAVIER PINTO, C. C. and EGEL, D. (2012). Inverse probability tilting for moment condition model with missing data. *Rev. Econ. Stud.* **79** 1053–1079. MR2986390 https://doi.org/10.1093/restud/rdr047

GRAHAM, B. S. and PINTO, C. C. D. X. (2018). Semiparametrically efficient estimation of the average linear regression function. Technical Report, National Bureau of Economic Research.

GRAHAM, B. S., PINTO, C. C. D. X. and EGEL, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *J. Bus. Econom. Statist.* **34** 288–301. MR3475879 https://doi.org/10.1080/07350015.2015.1038544

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

HIRSHBERG, D. A., MALEKI, A. and ZUBIZARRETA, J. (2019). Minimax linear estimation of the retargeted mean. Preprint. Available at arXiv:1901.10296.

HIRSHBERG, D. A. and WAGER, S. (2021). Supplement to "Augmented minimax linear estimation." https://doi.org/10.1214/21-AOS2080SUPP

IBRAGIMOV, I. A. and KHAS'MINSKIĬ, R. Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18–32.

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941 https://doi.org/10.1111/rssb.12027

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821 https://doi.org/10.1093/biomet/87.3.706

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

IMBENS, G. W., RUBIN, D. B. and SACERDOTE, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *Am. Econ. Rev.* **91** 778–794.

IMBENS, G. and WAGER, S. (2019). Optimized regression discontinuity designs. *Rev. Econ. Stat.* **101** 264–278.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

JOHNSTONE, I. M. (2015). Gaussian estimation: Sequence and wavelet models. Manuscript.

JUDITSKY, A. B. and NEMIROVSKI, A. S. (2009). Nonparametric estimation by convex programming. *Ann. Statist.* **37** 2278–2300. MR2543692 https://doi.org/10.1214/08-AOS654

KALLUS, N. (2018). Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems* 8909–8920.

KALLUS, N. (2020). Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.* **21** Paper No. 62, 54. MR4095341

KENNEDY, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. Preprint. Available at arXiv:2004.14497.

LANG, S. (1993). *Real and Functional Analysis*, 3rd ed. *Graduate Texts in Mathematics* **142**. Springer, New York. MR1216137 https://doi.org/10.1007/978-1-4612-0897-6

LECUÉ, G. and MENDELSON, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.* **46** 611–641. MR3782379 https://doi.org/10.1214/17-AOS1562

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LUGOSI, G. and ZEGER, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inf. Theory* **41** 677–687. MR1331260 https://doi.org/10.1109/18.382014

NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237 https://doi.org/10.2307/2951752

NEWEY, W. K. and ROBINS, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Preprint. Available at arXiv:1801.09138.

NIE, X. and WAGER, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. Preprint. Available at arXiv:1712.04912.

NING, Y., PENG, S. and IMAI, K. (2017). High dimensional propensity score estimation via covariate balancing.

PEYPOUQUET, J. (2015). *Convex Optimization in Normed Spaces: Theory, Methods and Examples. SpringerBriefs in Optimization*. Springer, Cham. MR3310025 https://doi.org/10.1007/978-3-319-13710-0

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117 https://doi.org/10.2307/1913713

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J., SUED, M., LEI-GOMEZ, Q. and ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable [MR2420458]. *Statist. Sci.* **22** 544–559. MR2420460 https://doi.org/10.1214/07-STS227D

ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762 https://doi.org/10.2307/1912705

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811 https://doi.org/10.1214/aos/1176350055

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. With discussion and a reply by the author. MR0443204

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DER LAAN, M. J., BENKESER, D. and CAI, W. (2019). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. Preprint. Available at arXiv:1908.05607.

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11, 40. MR2306500 https://doi.org/10.2202/1557-4679.1043

VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204. MR1091845 https://doi.org/10.1214/aos/1176347976

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

WANG, Y. and ZUBIZARRETA, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142 https://doi.org/10.1093/biomet/asz050

WONG, R. K. W. and CHAN, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105** 199–213. MR3768874 https://doi.org/10.1093/biomet/asx069

ZEILEIS, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Softw.* **11** 1–17.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHAO, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47** 965–993. MR3909957 https://doi.org/10.1214/18-AOS1698

ZHAO, Q., SMALL, D. S. and ERTEFAIE, A. (2017). Selective inference for effect modification via the lasso. Preprint. Available at arXiv:1705.08020.

ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning. Springer Ser. Statist.* 459–474. Springer, New York. MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672 https://doi.org/10.1080/01621459.2015.1023805