Omitted variable bias in machine learned causal models

Victor Chernozhukov
Carlos Cinelli
Whitney K. Newey
Amit Sharma
Vasilis Syrgkanis

OMITTED VARIABLE BIAS IN MACHINE LEARNED CAUSAL MODELS

V. CHERNOZHUKOV, C. CINELLI, W. NEWEY, A. SHARMA, V. SYRGKANIS

ABSTRACT. We derive general, yet simple, sharp bounds on the size of the omitted variable bias for a broad class of causal parameters that can be identified as linear functionals of the conditional expectation function of the outcome. Such functionals encompass many of the traditional targets of investigation in causal inference studies, such as, for example, (weighted) average of potential outcomes, average treatment effects (including subgroup effects, such as the effect on the treated), (weighted) average derivatives, and policy effects from shifts in covariate distribution—all for general, nonparametric causal models. Our construction relies on the Riesz-Frechet representation of the target functional. Specifically, we show how the bound on the bias depends only on the additional variation that the latent variables create both in the outcome and in the Riesz representer for the parameter of interest. Moreover, in many important cases (e.g, average treatment effects in partially linear models, or in nonseparable models with a binary treatment) the bound is shown to depend on two easily interpretable quantities: the nonparametric partial R^2 (Pearson's "correlation ratio") of the unobserved variables with the treatment and with the outcome. Therefore, simple plausibility judgments on the maximum explanatory power of omitted variables (in explaining treatment and outcome variation) are sufficient to place overall bounds on the size of the bias. Finally, leveraging debiased machine learning, we provide flexible and efficient statistical inference methods to estimate the components of the bounds that are identifiable from the observed distribution.

KEYWORDS: sensitivity analysis, omitted variable bias, omitted confounders, causal models, machine learning, confidence bounds.

1. Introduction

Causal inference with observational data usually relies on the assumption that the treatment assignment mechanism is "exogenous" or "ignorable" (i.e, independent of potential outcomes) conditional on a set of observed variables; or, equivalently, that the set of observed covariates satisfy the "backdoor" (or, more generally, adjustment) criterion [Rosenbaum and Rubin, 1983a, Pearl, 1995, 2009a, Angrist and Pischke, 2009, Shpitser et al., 2012, Imbens and Rubin, 2015]. Investigators who rely on the conditional ignorability assumption for drawing causal inferences from non-experimental studies must, therefore, also be able to cogently argue that there are *no unobserved confounders* of the treatment-outcome relationship. Yet, claiming the absence of unmeasured confounders is not only fundamentally unverifiable from the data, but often an assumption that is very hard to defend in practice. What if it is wrong?

When the assumption of no unobserved confounders is called into question, researchers are advised to perform sensitivity analyses, consisting of a formal and systematic assessment of the robustness of their findings against plausible violations of unconfoundedness. The problem of sensitivity analysis has been studied across several disciplines, dating back to, at least, the classical work of Cornfield et al. [1959], and with more recent works from Rosenbaum and Rubin [1983b], Robins [1999], Frank [2000], Rosenbaum [2002], Imbens [2003], Brumback et al. [2004], Altonji et al. [2005], Hosman et al. [2010], Imai et al. [2010], Vanderweele and Arah [2011], Blackwell [2013], Frank et al. [2013], Carnegie et al. [2016], Dorie et al. [2016], Middleton et al. [2016], Oster [2017], VanderWeele and Ding [2017], Kallus and Zhou [2018], Kallus et al. [2019], Cinelli et al. [2019], Franks et al. [2020], Cinelli and Hazlett [2020a,b], Bonvini and Kennedy [2021], Scharfstein et al. [2021], Jesson et al. [2021], among others. Most of this work, however, either focus on a specific target estimand of interest (e.g, a causal risk-ratio, or a causal risk difference), or impose parametric assumptions on the observed data, or on the nature of unobserved confounding (or both).

In this paper, we generalize the traditional "omitted variable bias" framework for a broad class of causal parameters that can be identified as linear functionals of the conditional expectation function of the outcome. Such functionals encompass many of the traditional targets of investigation in causal inference studies, such as, for example, (weighted) average of potential outcomes, average treatment effects (including subgroup effects, such as the effect on the treated), (weighted) average derivatives, policy effects from shifts in covariate distribution, and others—all for general, nonparametric causal models. Our construction relies on the Riesz-Frechet representation of the target functional. Specifically, we show how the bound on the bias has a simple characterization, depending only on the additional variation that the latent variables create both in the outcome and in the Riesz representer (RR) for the parameter of interest. We can thus perform sensitivity analysis with respect to violations of conditional ignorability in a broad class of causal models and target estimands.

Moreover, in many important cases (e.g, average treatment effects in partially linear models, or in nonseparable models with a binary treatment), we further show how the bias can be

reparameterized in terms of two easily interpretable quantities: the nonparametric partial R^2 (Pearson's "correlation ratio") of the unobserved variables with the treatment and with the outcome. Therefore, simple plausibility judgments on the maximum explanatory power of omitted variables (in explaining treatment and outcome variation) are sufficient to place overall bounds on the size of the bias. These results recover and generalize recent works on sensitivity analysis such as Cinelli and Hazlett [2020a] and Detommaso et al. [2021].

Finally, we provide flexible and efficient statistical inference for these bounds using debiased machine learning (DML) and auto-DML [Chernozhukov et al., 2017, 2016, 2018a, 2020, 2018b] as well as targeted MLE [Van der Laan and Rose, 2011]. DML methods can be seen as implementing the "one-step" semi-parametric correction [Pfanzagl and Wefelmeyer, 1978, Bickel et al., 1993] combined with cross-fitting, an efficient form of data-splitting, which makes it possible to use modern machine learning methods for estimating the identifiable components of the bounds, including regression functions, Riesz representers, the norm of regression residuals, and the norm of RRs. Auto-DML further automates the process and estimates RRs using their variational or adversarial characterization, without needing to know their analytical form. Auto-TML [Chernozhukov et al., 2018b] provides further refinements.

In what follows, Section 2 presents our method in the simpler context of partially linear models. The results in that section serve not only as an introduction to the main ideas of the more general, abstract framework, but are also important in their own right, since partially linear models are widely used in applied work. Section 3 then develops a general theory of omitted variable bias for continuous linear functionals of the conditional expectation of the outcome, based on their Riesz-Frechet representations. In Section 4 we construct high-quality inference methods for the bounds on the target parameters by leveraging recent advances in debiased machine learning with Riesz representers. Section 5 expands on popular target functionals of interest more formally. We conclude with Section 6, by offering some final remarks, and suggesting possible extensions.

Notation. All random vectors are defined on the probability space with law P. We consider a random vector $Z = (Y, W)$ with distribution P taking values z in its support $\mathcal{Z}$. We use P_V to denote the probability law of any subvector V and $\mathcal{V}$ denote its support. Denote the $L^q(P)$ norm of a measurable function $f : \mathcal{Z} \to \mathbb{R}$ and also the $L^q(P)$ norm of random variable $f(Z)$ by $\|f\|_{P,q} = \|f(Z)\|_{P,q}$. For a differentiable map $x \mapsto g(x)$, from $\mathbb{R}^d$ to $\mathbb{R}^k$, we use $\partial_{x'} g$ to abbreviate the partial derivatives $(\partial/\partial x')g(x)$, and we use $\partial_{x'} g(x_0)$ to mean $\partial_{x'} g(x) \mid_{x=x_0}$, etc. We use x' to denote the transpose of a column vector x. We use $R^2_{U \sim V}$ to denote the R^2 from the orthogonal linear projection of a scalar random variable U on a random vector V. We use the conventional notation dL/dP to denote the Radon-Nykodym derivative of measure L with respect to P.

2. Omitted Variable Bias in Partially Linear Models

To fix ideas, we begin our discussion in the context of partially linear models (PLM), i.e, the case in which the conditional expectation functions (CEF) of the outcome are linearly separable in the treatment. These results not only provide the key intuitions and the building blocks for the general case of nonseparable, nonparametric models of Section 3, but they are also important in their own right, as these models are widely used in applied work.

2.1. **Problem Set-Up.** Consider the partially linear regression model of the form

$$Y = \theta D + f(X, A) + \epsilon. \tag{1}$$

Here Y denotes a real-valued outcome, D a real-valued treatment, X an observed vector of covariates, and A an *unobserved* vector of covariates. We refer to $W := (D, X, A)$ as the "long" list of regressors, and to equation (1) as the "long" regression. For now, we assume the error term ϵ obeys $\mathrm{E}[\epsilon|D, X, A] = 0$ and thus $\mathrm{E}[Y|D, X, A] = \theta D + f(X, A)$.[1]

Under the traditional assumption of conditional exogeneity (or ignorability), we have that

$$\mathrm{E}[Y(d+1) - Y(d)] = \mathrm{E}[\mathrm{E}[Y|D = d+1, X, A] - \mathrm{E}[Y|D = d, X, A]] = \theta,$$

where $Y(d)$ denotes the *potential outcome* of Y when the treatment D is experimentally set to d. In other words, the assumptions of ignorability and a linearly separable CEF endow the regression coefficient θ with a causal meaning: the average treatment effect of a unit increase of D on the outcome Y. The problem, however, is that A is not observed, and thus both the long regression, and the regression coefficient θ cannot be identified.

Since the latent variables A are not measured, an alternative route to obtain an approximate estimate of θ is to consider the regression of Y on the "short" list of *observed* regressors $W^s := (D, X) \subset W$, as in,

$$Y = \theta_s D + f_s(X) + \epsilon_s. \tag{2}$$

Following convention, we call equation (2) the "short" regression. Here, again, we assume the error term ϵ_s obeys $\mathrm{E}[\epsilon_s \mid D, X] = 0$ and we thus have $\mathrm{E}[Y|D, X] = \theta_s D + f_s(X)$.[2] We can then use the "short" regression parameter θ_s as a proxy for θ. Evidently, in general they are not equal, $\theta_s \neq \theta$, and this naturally leads to the question of how far our "proxy" θ_s can deviate from the true inferential target θ.

Our goal is, thus, to analyze the difference between the short and long parameters—the omitted variable bias (OVB):

$$\theta_s - \theta,$$

---

[1] We can also consider, more generally, the case where the error term ϵ is centered and simply obeys $\mathrm{E}[\epsilon(D - \mathrm{E}[D \mid X, A])] = 0$. In this case, we lose the interpretation of $\theta D + f(X, A)$ as the CEF of the outcome, and it can be interpreted as the projection of the CEF on the space of functions that are partially linear in D.

[2] As before, one can also consider the case where ϵ_s is centered and simply obeys the orthogonality condition $\mathrm{E}[\epsilon_s(D - \mathrm{E}[D \mid X])] = 0$.

and perform inference on this bias under various hypotheses on the strength of the latent confounders A.

2.2. OVB as the Covariance of Approximation Errors.

Recall that, using a Frisch-Waugh-Lovell partialling out argument, one can express the long and short regression parameters, θ and θ_s, as the linear projection coefficients of Y on the residuals $D - \mathrm{E}[D \mid X, A]$ and $D - \mathrm{E}[D \mid X]$, respectively. That is,

$$\theta = \mathrm{E}Y\alpha(W), \qquad \theta^s = \mathrm{E}Y\alpha_s(W^s); \tag{3}$$

where here we define

$$\alpha(W) := \frac{D - \mathrm{E}[D \mid X, A]}{\mathrm{E}(D - \mathrm{E}[D \mid X, A])^2}, \quad \alpha_s(W^s) := \frac{D - \mathrm{E}[D \mid X]}{\mathrm{E}(D - \mathrm{E}[D \mid X])^2}.$$

For reasons that will become clear in the next section, we can refer to $\alpha(W)$ and $\alpha_s(W^s)$ as the "long" and "short" Riesz representers (RR).

Now let $g(W) := \mathrm{E}[Y \mid D, X, A]$ and $g_s(W^s) := \mathrm{E}[Y \mid D, X]$ denote the long and short regression functions, respectively. Using the orthogonality conditions in (1) and (2), we can further express θ and θ_s as

$$\mathrm{E}Y\alpha(W) = \mathrm{E}g(W)\alpha(W), \qquad \mathrm{E}Y\alpha_s(W^s) = \mathrm{E}g_s(W^s)\alpha_s(W^s). \tag{4}$$

Our first characterization of the OVB is thus as follows, where we use the shorthand notation: $g = g(W)$, $g_s = g_s(W^s)$, $\alpha = \alpha(W)$, and $\alpha_s = \alpha_s(W^s)$.

Theorem 1 (OVB in PLM). *Assume that Y and D are square integrable with:*

$$\mathrm{E}(D - \mathrm{E}[D \mid X, A])^2 > 0$$

Then the OVB for the partially linear model of equations (1) - (2) is given by

$$\theta_s - \theta = \mathrm{E}(g_s - g)(\alpha_s - \alpha),$$

that is, it is the covariance between the regression error and the RR error. Furthermore, the squared bias can be bounded as

$$|\theta_s - \theta|^2 =: \rho^2 B^2 \leq B^2,$$

where

$$B^2 := \mathrm{E}(g - g_s)^2 \mathrm{E}(\alpha - \alpha_s)^2, \quad \rho^2 := \mathrm{Cor}^2(g - g_s, \alpha - \alpha_s).$$

The bound B^2 is the product of additional variations that omitted confounders generate in the regression function and in the RR. This bound is sharp for the adversarial confounding that maximizes ρ^2 to 1 over choices of α and g, holding $\mathrm{E}(\alpha - \alpha_s)^2$ and $\mathrm{E}(g - g_s)^2 \leq \mathrm{E}(Y - g_s)^2$ fixed, provided that the observed distribution of (Y, D, X) places no further constraints on the problem.

This result for partially linear regression models is new, and generalizes results for classical linear regression models. Moreover, this result naturally generalizes for completely nonseparable regression models, as we show in Section 3.

Sensitivity analysis requires making plausibility judgments on the values of the sensitivity parameters. Therefore, it is important that such parameters be well-understood, and easily interpretable in applied settings. Here we show how the bias of Theorem 1 can be further interpreted in terms of conventional R^2s. This interpretation is inspired by Imbens [2003] and, specifically, by the partial R^2 characterizations of the OVB in linear models by Cinelli and Hazlett [2020a]. Let us use $R^2_{V \sim U} = \mathrm{Cor}^2(U, V)$ to denote the R^2 from the orthogonal linear projection of random variable U on random variable V.

Corollary 1 (Interpreting OVB Bounds in Terms of R^2). *Under the conditions of Theorem 1, we can express the bound B^2 as*

$$B^2 = S^2 C_g^2 C_\alpha^2, \quad S^2 := \frac{\mathrm{E}\tilde{Y}_s^2}{\mathrm{E}\tilde{D}_s^2}, \quad C_g^2 := R^2_{\tilde{Y}_s \sim A_1}, \quad C_\alpha^2 := \frac{R^2_{\tilde{D}_s \sim A_2}}{1 - R^2_{\tilde{D}_s \sim A_2}}, \tag{5}$$

where $\tilde{Y}_s := Y - \mathrm{E}[Y \mid D, X]$ is the residualized outcome, and $\tilde{D}_s := D - \mathrm{E}[D \mid X]$ is the residualized treatment, using only the observed covariates, $A_1 := \mathrm{E}[Y \mid D, X, A] - \mathrm{E}[Y \mid D, X]$ is the effective confounder of the outcome, and $A_2 := \mathrm{E}[D \mid X, A] - \mathrm{E}[D \mid X]$ is the effective confounder of the treatment.

The bound is the product of the term S^2, which is directly identifiable from the observed distribution of (Y, D, X), and the term $C_g^2 C_\alpha^2$, which is not identifiable, and needs to be restricted through hypotheses that limit strength of confounding.

The factors C_g^2 and C_α^2 measure the strength of confounding that the omitted variables generate in the outcome and treatment regressions. They are stated in terms of simple R^2's. Specifically, $R^2_{\tilde{Y}_s \sim A_1}$ in the first factor stands for the proportion of variance of the residualized outcome explained by latent confounders. Further, $R^2_{\tilde{D}_s \sim A_2}$ in the second factor stands for the proportion of variance of the residualized treatment explained by latent confounders. In either case the effect of latent variables operate through the "effective" confounders, A_1 or A_2. While these quantities are given in terms of linear projection R^2's, it turns out they correspond to nonparametric partial R^2's, as given by Pearson's "correlation ratio" [Pearson, 1905], as we further explain below.

Returning to Theorem 1, the bound B^2 is the total potential amount of squared bias generated by confounding, and the actual amount of confounding is amortized by the correlation ρ. Adversarial confounding would select this correlation to maximize the bias, by setting $\rho^2 = 1$, while amicable confounding would minimize the bias, and set $\rho^2 = 0$. The latter corresponds to the case in which the "effective" confounders A_1 and A_2 are uncorrelated. In principle ρ^2 could be set to various values less than 1 (say, $\rho^2 = .5$) when confounding

is assumed to be "natural" rather than adversarial.[3] Here we focus on the case in which the researcher has no knowledge of the functional form of the CEFs in order to limit ρ, and accordingly, interpret the bias bounds B^2 as resulting from the presence of adversarial confounding.

Finally, the above results hold for population data. In practice, both θ_s and S^2 need to be estimated from finite samples. This can be readily done using debiased machine learning, as we discuss in Section 4. This enables efficient statistical inference on the bounds for θ under any hypothetical strength of the sensitivity parameters C_α and C_g. These results allow researchers to perform sharp sensitivity analyses in a flexible class of machine-learned causal models using very simple, and interpretable, tools.

2.3. **Characterization of the OVB Bounds in Terms of Nonparametric R^2's.** We now show that the previous residual R^2 quantities, $R^2_{\tilde{Y}_s \sim A_1}$ and $R^2_{\tilde{D}_s \sim A_2}$, can be interpreted as the non-parametric partial R^2s of A with Y given (D, X), and of A with D given X, correspondingly.

When the CEF is not linear, a natural measure of the strength of relationship between covariates W and Y is the *nonparametric R^2*, $\eta^2_{Y \sim W} := \frac{\mathrm{Var}(\mathrm{E}[Y|W])}{\mathrm{Var}(Y)}$. The nonparametric R^2 has been extensively studied in the context of nonparametric regression (see e.g Doksum and Samarov [1995]), and it was first introduced by Pearson [1905] as a generalization of the linear R^2 (and thus also known as the *Pearson's correlation ratio*). We define the nonparametric *partial R^2* of A with Y given (D, X), $\eta^2_{Y \sim A|D,X}$, as

$$\eta^2_{Y \sim A|D,X} := \frac{\mathrm{Var}(\mathrm{E}[Y|A, D, X]) - \mathrm{Var}(\mathrm{E}[Y|D, X])}{\mathrm{Var}(Y) - \mathrm{Var}(\mathrm{E}[Y|D, X])} = \frac{\eta^2_{Y \sim A,D,X} - \eta^2_{Y \sim D,X}}{1 - \eta^2_{Y \sim D,X}},$$

which measures the maximum proportion of the residual variation of the outcome that the latent confounders A explain, after taking into account the variation already explained by observed covariates. Simple algebra shows that the nonparametric partial R^2 can also be written as the linear R^2 of the residuals $\tilde{Y}_s := Y - \mathrm{E}[Y \mid D, X]$ with the "effective" confounder $A_1 := \mathrm{E}[Y \mid A, D, X] - \mathrm{E}[Y \mid D, X]$,

$$\eta^2_{Y \sim A|D,X} = \mathrm{Cor}^2(\tilde{Y}_s, A_1) = R^2_{\tilde{Y}_s \sim A_1}.$$

The definition of the non-parametric partial R^2 of A with D given X, $\eta^2_{D \sim A|X}$, follows the same logic, and stands for the maximum proportion of the residual variation of the treatment that the latent confounders A explain, after taking into account the variation already explained by observed covariates X. Using similar reasoning, we conclude that $\eta^2_{D \sim A|X}$ is identical to the linear R^2 of the residual $\tilde{D}_s := D - \mathrm{E}[D \mid X]$ with its corresponding "effective" confounder $A_2 := \mathrm{E}[D \mid X, A] - \mathrm{E}[D \mid X]$,

$$\eta^2_{D \sim A|X} = \mathrm{Cor}^2(\tilde{D}_s, A_2) = R^2_{\tilde{D}_s \sim A_2}.$$

We are now ready to re-express the bias in terms of Pearson's partial η^2.

---

[3]For instance, suppose that nature chooses $\rho^2 \sim U(0, 1)$. This yields an expected value for ρ^2 of .5.

Corollary 2 (**OVB in terms of Pearson's partial** η^2). *Under the conditions of Theorem 1, we can express the bounds components as*

$$C_g^2 := \eta_{Y \sim A|D,X}^2, \quad C_\alpha^2 := \frac{\eta_{D \sim A|X}^2}{1 - \eta_{D \sim A|X}^2}.$$

We now see that our bounds generalize the bounds of Cinelli and Hazlett [2020a] for partially linear models, by simply replacing linear R^2s with nonparametric R^2s. That is, consider linear CEFs for D and Y. We then have that $\eta_{Y \sim A|D,X}^2 = R_{Y \sim A|D,X}^2$, and $\eta_{D \sim A|D,X}^2 = R_{D \sim A|D,X}^2$. Thus:

$$B^2 = \left(\frac{\text{Var}(\tilde{Y}_s)}{\text{Var}(\tilde{D}_s)} \right) \left(\frac{R_{Y \sim A|D,X}^2 R_{D \sim A|X}^2}{1 - R_{D \sim A|X}^2} \right)$$

where $\tilde{Y}_s$ and $\tilde{D}_s$ denote the residuals of the *linear* projections of Y on (D, X), and of D on X, respectively. This recovers equation (8) of Cinelli and Hazlett [2020a, p.48].

2.4. **Sensitivity Analysis: A Conceptual Example.** The importance of the previous corollaries stems from the fact that it greatly reduces the complexity of plausibility judgments— no matter how complicated the nonlinear term $f(X, A)$ of $\text{E}[Y|D, X, A]$ actually is, or no matter how complicated $\text{E}[D|X, A]$ actually is, to place bounds on the size of the bias, researchers need only to reason about the *maximum explanatory power* that unobserved confounders A have in explaining treatment and outcome variation.

As an example of how these bounds can be used in practice, suppose that theory or prior studies suggest that unobserved confounders A can explain at most 10 percent of the variation of the treatment and of the outcome, above and beyond what observed covariates already explain. This implies $\eta_{Y \sim A|D,X}^2 = .1$ and $\eta_{D \sim A|X}^2 = .1$, which then translates into a bound on the squared bias of:

$$B^2 = S^2 \left(\frac{\eta_{Y \sim A|D,X}^2 \eta_{D \sim A|X}^2}{1 - \eta_{D \sim A|X}^2} \right) = S^2 \frac{(.1)(.1)}{(.9)} \approx S^2 \times .011$$

This further leads to bounds on the target parameter θ,

$$\theta_\pm := \theta_s \pm \sqrt{B^2} \approx \theta_s \pm S \times .105$$

We shall refer to this type of scenario analysis as benchmarking [Imbens, 2003, Altonji et al., 2005, Oster, 2017, Cinelli and Hazlett, 2020a]. The role of benchmarking is to examine the sensitivity of causal inferences to plausible strengths of the omitted confounders.

Notice there is a trade-off between the parameters that bound the bias: in order to maintain the same bound, a higher degree of confounding in the treatment can be offset by a lower degree of confounding in the outcome. Therefore, a useful tool for visualizing the whole sensitivity range of the target parameter, under different assumptions regarding the strength of confounding, is a bivariate contour plot [Imbens, 2003, Cinelli and Hazlett, 2020a] showing the collection of curves in the space of nonparametric partial R^2 values $(\eta_{Y \sim A|D,X}^2, \eta_{D \sim A|X}^2)$
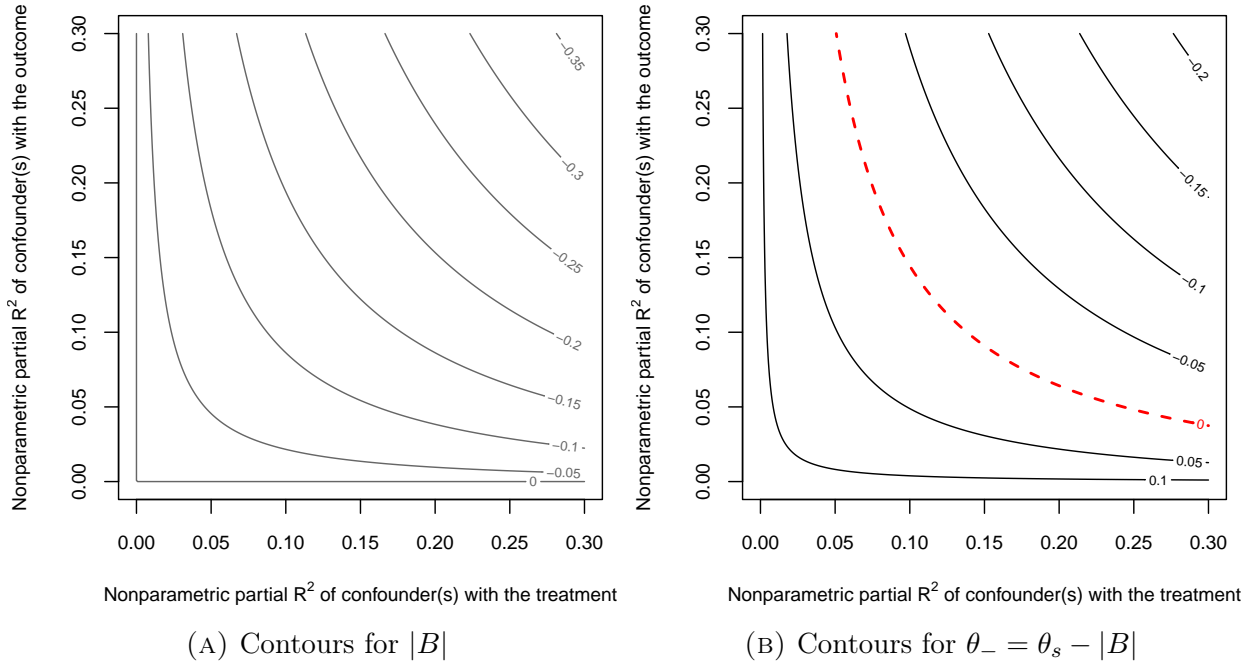
(A) Contours for $|B|$

(B) Contours for $\theta_- = \theta_s - |B|$

FIGURE 1. Sensitivity contour plots of a hypothetical example.

Note: The vertical axis shows the nonparametric partial R^2 of latent variables with the outcome, i.e, the maximum proportion of the residual variance of the outcome that could be explained by latent confounders. The horizontal axis shows the nonparametric partial R^2 of latent variables with the treatment, i.e, the maximum proportion of the residual variance of the treatment that could be explained by latent confounders. Fig 1a shows the contours for the absolute value of the bound on the bias, $|B|$. Fig 1b shows the contours for the lower bound of the target parameter itself, i.e, $\theta_- = \theta_s - |B|$, which could be brought to the critical value of zero (dashed red contour), or beyond zero. In both cases, the further the curve is from the origin, the higher the bias.

along which the bounds are constant. Figure 1 illustrates such curves for a hypothetical example, both for the bound on the absolute value of the bias $|B|$ (Fig 1a), and for the lower bound of the target parameter itself, i.e, $\theta_- = \theta_s - |B|$ (Fig 1b), assuming the original θ_s was positive. In this particular hypothetical example, for instance, confounders that explain 10 percent of the residual variation of the treatment and of the outcome would *not* be sufficiently strong to bring down the lower bound for θ_s to the critical threshold of zero.

3. OMITTED VARIABLE BIAS IN NONPARAMETRIC CAUSAL MODELS

In this section we derive the main partial identification theorems of the paper, and construct sharp bounds on the size of the omitted variable bias for a broad class of causal parameters that can be identified as linear functionals of the conditional expectation function of the outcome, all for general nonparametric causal models. Although more abstract, the presentation of this section largely parallels the special case of partially linear models given in Section 2.

3.1. **Problem Set-Up.** Consider the following modern acyclical structural equations model (SEM) as an example:

$$
\begin{aligned}
Y &:= g_Y(D, X, A, \epsilon_Y), \\
D &:= g_D(X, A, \epsilon_D), \\
A &:= g_A(X, \epsilon_A), \\
X &:= \epsilon_X,
\end{aligned}
$$

where Y is an outcome variable, D is a treatment variable, X is a vector-valued confounder variable, A is a vector-valued latent confounder variable, and $\epsilon_Y, \epsilon_D, \epsilon_A$ are vector-valued structural disturbances that are mutually independent. This model has an associated Directed Acyclic Graph (DAG) [Pearl, 1995, 2009a] as shown in Figure 2.
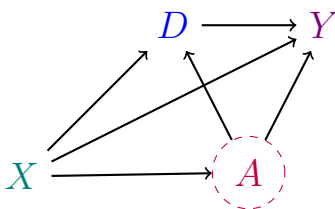

FIGURE 2. DAG associated with the SEM.

Note: D denotes the treatment, Y the outcome, X observed confounders, and A *unobserved* confounders. The node for the unobserved confounders is circled.

The SEM above induces the potential outcome $Y(d)$ under the intervention that replaces the structural equation of D by the fixed value d. That is,

$$Y(d) := g_Y(d, X, A, \epsilon_Y).$$

Additionally, the independence of the structural disturbances implies the following conditional exogeneity (or, ignorability) condition:

$$Y(d) \perp\!\!\!\perp D \mid \{X, A\} \tag{6}$$

which states that the realized treatment D is independent of the potential outcomes, conditional on X and A.

More generally, we can work with any structural equation model that implies the existence of the potential outcomes $Y(d)$, and such that the conditional exogeneity (6) holds. In fact there are many structural causal models that satisfy such assumptions; see e.g. Pearl [2009b] and Figure 3 for concrete examples. The causal interpretation of our results rely only on conditional exogeneity. Under this set-up, we then have the following (well-known) identification result

$$\mathrm{E}[Y(d) \mid D = d, X, A] = \mathrm{E}[Y \mid D = d, X, A] =: g(d, X, A),$$

that is, the conditional average potential outcome coincides with the "long" regression function of Y on D, X, and A. Therefore, we can identify various causal parameters—functionals of
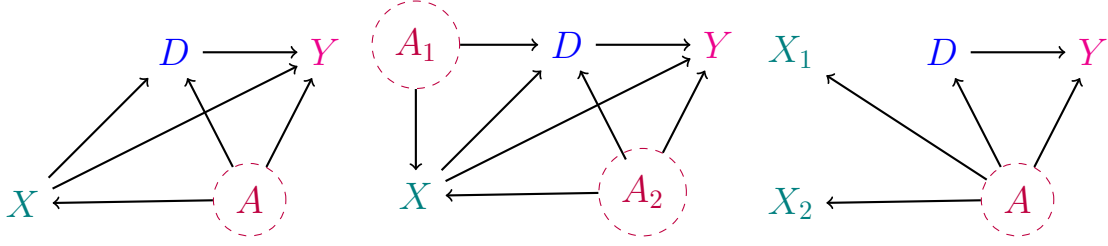
FIGURE 3. Examples of different DAGs that imply $Y(d) \perp\!\!\!\perp D \mid \{X, A\}$.

Note: Examples of DAGs (nonparametric SEMs) that imply the conditional exogeneity condition (6). Latent nodes are circled. In the left DAG, the arrow from $A \to X$ is in reverse order relative to the DAG of Figure 2. In this DAG we still need to condition on X and A to identify the causal effect of D on Y. In the center DAG, we need to condition on $A = (A_1, A_2)$ and X to identify causal effect of D on Y. The center DAG can be viewed as a special case of the left DAG by setting $A = (A_1, A_2)$. In the right DAG, it suffices to control for A to identify the average causal effects of D on Y, but we only observe X_1 and X_2, the so called "negative" controls, which are measurements, or proxies, of A. The conditional exogeneity condition (6) still holds in this case.

the average potential outcome—from the regression function. Important examples include: (i) the average causal effect (ACE)

$$\theta = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[g(1, X, A) - g(0, X, A)],$$

for the case of a binary treatment D; and, (ii) the average causal derivative (ACD)

$$\theta = \mathrm{E}[\partial_d Y(D)] = \mathrm{E}[\partial_d g(D, X, A)],$$

for the case of a continuous treatment D.

In fact, our framework is considerably more general, in that it covers any target parameter of the following general form.

Assumption 1 (Target "Long" Parameter). *The target parameter θ is a continuous linear functional of the long regression:*

$$\theta := \mathrm{E}m(W, g); \tag{7}$$

where the mapping $f \mapsto m(w; f)$ is linear in $f \in L^2(P)$, and the mapping $f \mapsto \mathrm{E}m(W, f)$ is continuous in g with respect to the $L^2(P)$ norm.

This formulation covers the two working examples above with $m(W, g) = g(1, X, A) - g(0, X, A)$ for the ACE and $m(W, g) = \partial_d g(D, X, A)$ for the ACD, and the continuity condition holds under the regularity condition provided in the remark below. In addition to these examples, we show that many other examples in Section 5 are of this form; and further examples of this form (e.g, consumer surplus, decomposition functionals) can be found in Chernozhukov et al. [2018b].

Remark 1 (Regularity Conditions for ACE and ACD). As regularity conditions for the ACE we assume $\mathrm{E}Y^2 < \infty$ and the weak overlap condition:

$$\mathrm{E}[P(D = 1 \mid X, A)^{-1} P(D = 0 \mid X, A)^{-1}] < \infty.$$

As regularity conditions for the ACD we assume $\mathrm{E}Y^2 < \infty$, that the conditional density $d \mapsto f(d|x, a)$ is continuously differentiable on its support $\mathcal{D}_{x,a}$, the regression function $d \mapsto g(d, x, a)$ is continuously differentiable on $\mathcal{D}_{x,a}$, and we have that $g(d, x, a)f(d|x, a) = 0$ on the boundary of $\mathcal{D}_{x,a}$. The above needs to hold for all values x and a in the support of (X, A). We also impose the bounded information assumption:

$$\mathrm{E}(\partial_d \log f(D \mid X, A))^2 < \infty.$$

These conditions imply that Assumption 1 holds, by Lemma 4 given in Section 5. $\qquad\square$

The *key problem* is that we do not observe A, and therefore we can only identify the "short" conditional expectation of Y given D and X, i.e.

$$g_s(D, X) := \mathrm{E}[Y \mid D, X] = \mathrm{E}[g(D, X, A)|D, X],$$

which is given by the projection of the long regression $g(D, X, A)$ on the subspace generated by the observed D and X. Given the short regression, we can compute proxies (or approximations) θ_s for θ. In particular, for the ACE, the short parameter consists of

$$\theta_s = \mathrm{E}[g_s(1, X) - g_s(0, X)],$$

and for the ACD,

$$\theta_s = \mathrm{E}[\partial_d g_s(D, X)].$$

In this general framework, the proxy parameters can also be expressed as the same linear functionals applied to the short regression, $g_s(W^s)$.

Assumption 2 (Proxy "Short" Parameter). *The proxy parameter θ_s is defined by replacing the long regression g with the short regression g_s in the definition of the target parameter:*

$$\theta_s := \mathrm{E}m(W, g_s).$$

We require $m(W, g_s) = m(W^s, g_s)$, i.e., the score depends only on W^s when evaluated at g_s.

Indeed, in the two working examples this assumption is satisfied, since $m(W, g_s) = m(W^s, g_s) = g_s(1, X) - g_s(0, X)$ for the ACE and $m(W, g_s) = m(W^s, g_s) = \partial_d g_s(D, X)$ for the ACD. Section 5 verifies this assumption for other examples.

Our goal is to provide bounds on the omitted variable bias (OVB), ie., the difference between the "long" and "short" functionals,

$$\theta_s - \theta,$$

under assumptions that limit the strength of confounding, and perform statistical inference on its size.

3.2. **Omitted Variable Bias for Linear Functionals of the CEF.** The key to bounding the bias is the following lemma that characterizes the target parameters and their proxies as inner products of regressions with terms called Riesz representers (RR).

Lemma 1 (**Riesz Representation**). *There exist unique square integrable random variables* $\alpha(W)$ *and* $\alpha_s(W^s)$, *the long and short Riesz representers, such that*

$$\theta = \mathrm{E}m(W, g) = \mathrm{E}g(W)\alpha(W), \quad \theta_s = \mathrm{E}m(W^s, g_s) = \mathrm{E}g_s(W^s)\alpha_s(W^s),$$

for all square-integrable g's and g_s. Furthermore, $\alpha_s(W^s)$ is the projection of α_s in the sense that

$$\alpha_s(W^s) = \mathrm{E}[\alpha(W) \mid W^s].$$

In the case of the ACE with a binary treatment, we have that

$$\alpha(W) = \frac{1(D = 1)}{P(D = 1 \mid X, A)} - \frac{1(D = 0)}{P(D = 0 \mid X, A)}, \quad \alpha_s(W) = \frac{1(D = 1)}{P(D = 1 \mid X)} - \frac{1(D = 0)}{P(D = 0 \mid X)},$$

and in the case of the ACD with a continuous treatment, we have that

$$\alpha(W) = -\partial_d \log f(D \mid X, A), \quad \alpha_s(W^s) = -\partial_d \log f(D \mid X).$$

Sometimes it is useful to impose restrictions on the regression functions, such as partial linearity or additivity. The next lemma describes the RR property for the long and short target parameters in this case.

Lemma 2 (**Riesz Representation for Restricted Regression Classes**). *Furthermore, if g is known to belong to a closed linear subspace Γ of $L^2(P_W)$, and g_s is known to belong to a closed linear subspace $\Gamma_s = \Gamma \cap L^2(P_{W^s})$, then there exist unique long RR $\bar{\alpha}$ in Γ and unique short RR $\bar{\alpha}_s$ in Γ_s that continue to have the representation property*

$$\theta = \mathrm{E}m(W, g) = \mathrm{E}g(W)\bar{\alpha}(W), \quad \theta_s = \mathrm{E}m(W^s, g_s) = \mathrm{E}g_s(W^s)\bar{\alpha}_s(W^s),$$

for all $g \in \Gamma$ and $g_s \in \Gamma_s$. Moreover, they are given by the orthogonal projections of α and α_s on Γ and Γ_s, respectively. Since projections reduce the norm, we have $\mathrm{E}\bar{\alpha}^2 \leq \mathrm{E}\alpha^2$ and $\mathrm{E}\bar{\alpha}_s^2 \leq \mathrm{E}\alpha_s^2$. Furthermore, the best linear projection of $\bar{\alpha}$ on $\bar{\alpha}_s$ is given by $\bar{\alpha}_s$, namely,

$$\min_{b \in \mathbb{R}} \mathrm{E}(\bar{\alpha} - b\bar{\alpha}_s)^2 = \mathrm{E}(\bar{\alpha} - \bar{\alpha}_s)^2 = \mathrm{E}\bar{\alpha}^2 - \mathrm{E}\bar{\alpha}_s^2.$$

To illustrate, suppose that the regression functions are partially linear, as in Section 2

$$g(W) = \beta D + f(X, A), \quad g_s(W^s) = \beta_s D + f_s(X),$$

then for either the ACE or the ACD we have that the RR are given by

$$\alpha(W) = \frac{D - \mathrm{E}[D \mid X, A]}{\mathrm{E}(D - \mathrm{E}[D \mid X, A])^2}, \quad \alpha_s(W^s) = \frac{D - \mathrm{E}[D \mid X]}{\mathrm{E}(D - \mathrm{E}[D \mid X])^2}.$$

In what follows we use the notation α and α_s without bars, with the understanding that if such restrictions have been made, then we work with $\bar{\alpha}$ and $\bar{\alpha}_s$.

Using these lemmas, we immediately obtain the following characterization of the OVB.

Theorem 2 (**OVB and Sharp Bounds**). *Consider the long and short parameters θ and θ_s as given by Assumptions 1 and 2. We then have that the OVB is*

$$\theta_s - \theta = \mathrm{E}(g_s - g)(\alpha_s - \alpha),$$

that is, it is the covariance between the regression error and the RR error. Therefore, the squared bias can be bounded as

$$|\theta_s - \theta|^2 = \rho^2 B^2 \leq B^2,$$

where

$$B^2 := \mathrm{E}(g - g_s)^2 \mathrm{E}(\alpha - \alpha_s)^2, \quad \rho^2 := \mathrm{Cor}^2(g - g_s, \alpha - \alpha_s).$$

The bound B^2 is the product of additional variations that omitted confounders generate in the regression function and in the RR. This bound is sharp for the adversarial confounding that maximizes ρ^2 to 1 over choices of α and g, holding $\mathrm{E}(\alpha - \alpha_s)^2$ and $\mathrm{E}(g - g_s)^2 \leq \mathrm{E}(Y - g_s)^2$ fixed, provided that the observed distribution of (Y, D, X) places no further constraints on the problem.

This is a general OVB formula that covers a wide variety of causal estimands of interest, as long as they can be written as linear functionals of the long regression. In particular, it recovers classical OVB formulas for linear regression; for the case of average causal derivatives, it recovers the OVB formula in Detommaso et al. [2021] (which was derived via a different method using a flow representation of a DAG). It applies to rich classes of examples analyzed in Section 5, and many other examples (e.g., consumer surplus, decomposition of total effect into direct and indirect and others) discussed in Chernozhukov et al. [2018b].

Finally, we note the following interesting fact.

Remark 2 (**Tighter Bounds under Restrictions**). When we work with restricted parameter spaces, the restricted RRs obey

$$\mathrm{E}(\bar{\alpha} - \bar{\alpha}_s)^2 \leq \mathrm{E}(\alpha - \alpha_s)^2,$$

since the orthogonal projection on a closed subspace reduces the $L^2(P)$ norm. This means that the bounds become tighter in this case. Therefore, by default, when restrictions have been made, we work with restricted RRs. □

3.3. **Characterization of the OVB Bounds.** In the same spirit of Section 2, we can further derive useful characterizations of the bounds.

Corollary 3 (**Interpreting Bounds**). *The bound of Theorem 2 can be re-expressed as*

$$B^2 = S^2 C_g^2 C_\alpha^2, \tag{8}$$

where $S^2 := \mathrm{E}(Y - g_s)^2 \mathrm{E}\alpha_s^2$ and

$$C_g^2 := \frac{\mathrm{E}(g - g_s)^2}{\mathrm{E}(Y - g_s)^2} = R^2_{Y - g_s \sim g - g_s}, \quad C_\alpha^2 := \frac{\mathrm{E}\alpha^2 - \mathrm{E}\alpha_s^2}{\mathrm{E}\alpha_s^2} = \frac{1 - R^2_{\alpha \sim \alpha_s}}{R^2_{\alpha \sim \alpha_s}}.$$

This generalizes the result of Corollary 1 to fully nonlinear models, and general target parameters defined as linear functionals of the long regression. As before, the bound is the product of the term S^2, which is directly identifiable from the observed distribution of (Y, D, X), and the term $C_g^2 C_\alpha^2$, which is not identifiable, and needs to be restricted through hypotheses that limit strength of confounding.

Moreover, we again have the following useful interpretation of C_g^2.

Remark 3 (Interpretation of C_g^2). The term C_g can be also written as

$$C_g^2 = R_{\tilde{Y}_s \sim A_1}^2 = \eta_{Y \sim A|D,X}^2,$$

where $Y_s = Y - g_s(X)$ is the short residual and $A_1 = g(X, A) - g_s(X)$ is the effective confounder for the outcome. As discussed in Section 2.3, this is exactly equal to the nonparametric partial R^2 of A with Y, given D and X, namely, $\eta_{Y \sim A|D,X}^2$. □

Thus, as before, the terms C_g^2 and C_α^2 generally measure the strength of confounding that the omitted variables generate in the outcome regression and in the treatment:

- $R_{\tilde{Y}_s \sim A_1}^2$ is the proportion of residual variance in the outcome explained by confounders;
- $1 - R_{\alpha \sim \alpha_s}^2$ is the proportion of variance of the long RR *not* explained by the short RR.

The case of zero adversarial confounding arises whenever one of these two parameters is zero. Figure 4 shows hypothetical contours in the space of R^2-squares $(R_{\tilde{Y}_s \sim A_1}^2, 1 - R_{\alpha \sim \alpha_s}^2)$ along which the bias bound is constant. Here, again, there is a trade-off: greater confounding with the outcome can be compensated by smaller confounding with the treatment, and vice-versa.

The interpretation of C_α^2 can be further refined for special cases. For instance, in the case of the partially linear regression model, the term C_α^2 reduces exactly to the terms given in Theorem 1, as well as Corollaries 1 and 2. Moreover, equivalent results can be obtained for the ACE with a binary treatment.

Remark 4 (Interpretation of C_α^2 for ACE with a Binary Treatment). For the ACE example, we have that

$$C_\alpha^2 = \frac{\mathrm{E}[\pi(X)(1 - \pi(X))]}{\mathrm{E}[\pi(X, A)(1 - \pi(X, A))]} - 1, \tag{9}$$

where $\pi(X) = \mathrm{P}(D = 1 \mid X)$ and $\pi(X, A) = \mathrm{P}(D = 1 \mid X, A)$, which is the ratio of the average short conditional variance of the treatment to the average long conditional variance of the treatment minus 1. This further leads to

$$C_\alpha^2 = \frac{R_{\tilde{D}_s \sim A_2}^2}{1 - R_{\tilde{D}_s \sim A_2}^2} = \frac{\eta_{D \sim A|X}^2}{1 - \eta_{D \sim A|X}^2}, \tag{10}$$
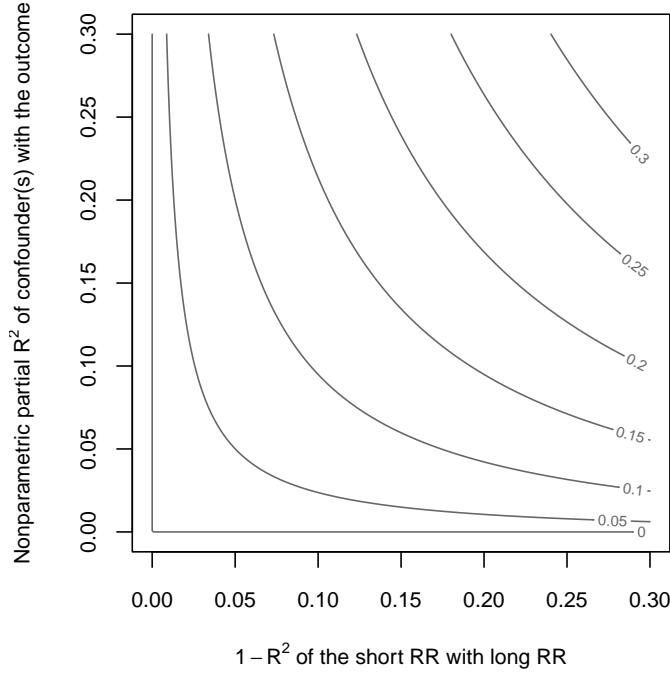
FIGURE 4. Hypothetical contours for the absolute value of the bias bound $|B|$.

Note: The horizontal axis shows the proportion of unexplained variation in the long RR as given by the short RR, i.e, $1 - R^2_{\alpha \sim \alpha_s}$. The vertical axis shows the amount of variation in the residualized outcome Y that can be explained by the "effective confounder" $A_1 := g - g_s$, or, equivalently, the nonparametric partial R^2 of A with Y, i.e, $R^2_{\tilde{Y}_s \sim A_1} = \eta^2_{Y \sim A|D,X}$. Contours show the bias bound $|B| = SC_g C_\alpha$. The further the curve is from the axes, the higher the bias.

where again $\tilde{D}_s = D - \pi(X)$ and $A_2 = \pi(X, A) - \pi(X)$ is the effective confounder for the treatment. Hence the interpretation of C^2_α for the ACE with a binary treatment is identical to the interpretation for the case of the partially linear model. □

And a similar interpretation applies for average causal derivatives.

Remark 5 (Interpretation of C_α for Average Causal Derivatives). For the ACE example,

$$C^2_\alpha = \frac{\mathrm{E}[(\partial_d \log f(D \mid X, A))^2]}{\mathrm{E}[(\partial_d \log f(D \mid X))^2]} - 1, \tag{11}$$

which can be interpreted as the relative increase in the information that the confounder A provides about the location of D. If D is homoscedastic Gaussian, conditional on both X and (X, A), we have

$$\partial_d \log f(D \mid X, A) = -\frac{D - \mathrm{E}[D \mid X, A]}{\mathrm{E}(D - \mathrm{E}[D \mid X, A])^2}, \quad \partial_d \log f(D \mid X, A) = -\frac{D - \mathrm{E}[D \mid X]}{\mathrm{E}(D - \mathrm{E}[D \mid X))^2},$$

so that C^2_α simplifies to the term C^2_α found for the partially linear model. □

4. STATISTICAL INFERENCE ON THE BOUNDS

The bounds for the target parameter θ take the form

$$\theta_{\pm} = \theta_s \pm |\rho| S C_g C_\alpha, \quad S^2 = \mathrm{E}(Y - g_s)^2 \mathrm{E}\alpha_s^2.$$

The components C_g, C_α are restricted through benchmarking hypotheses. The correlation $|\rho|$ can be set to 1 under adversarial confounding.[4] The unknown components of the bounds are S and θ_s. We can estimate these components via debiased machine learning (DML), which is a form of the classical "one-step" semi-parametric correction [Pfanzagl and Wefelmeyer, 1978, Bickel et al., 1993] combined with cross-fitting, an efficient form of data-splitting.

For debiased machine learning of θ_s, we exploit the representation

$$\theta_s = \mathrm{E}[m(W^s, g_s) + (Y - g_s)\alpha_s],$$

as in Chernozhukov et al. [2018b, 2021a]. This representation is Neyman orthogonal with respect to perturbations of (g_s, α_s), which is a key property required for DML. Another component to be estimated is

$$\mathrm{E}(Y - g_s)^2 =: \sigma_s^2,$$

which is also Neyman-orthogonal with respect to g_s. The final component to be estimated is $\mathrm{E}\alpha_s^2$. For this we explore the following formulation:

$$\mathrm{E}\alpha_s^2 = 2\mathrm{E}m(W^s, \alpha_s) - \mathrm{E}\alpha_s^2 =: \nu_s^2,$$

where the latter parameterization is Neyman-orthogonal. Application of DML theory in Chernozhukov et al. [2017] and the delta-method imply the statistical properties of the estimated bounds under the condition that machine learning of g_s and α_s is of sufficiently high quality, with rate faster than $n^{-1/4}$.

Specifically Neyman orthogonality refers to the property:

$$\partial_{g,\alpha}\mathrm{E}[m(W^s, g) + (Y - g)\alpha]\Big|_{\alpha=\alpha_s, g=g_s} = 0;$$

$$\partial_g \mathrm{E}(Y - g)^2\Big|_{g=g_s} = 0;$$

$$\partial_\alpha \mathrm{E}[2m(W^s, \alpha) - \alpha^2]\Big|_{\alpha=\alpha_s} = 0;$$

where ∂ is the Gateaux (pathwise derivative) operator over directions $h \in L^2(P_{W^s})$.

The estimation relies on the following generic algorithm.

Definition 1 (DML(ψ)). *Input the Neyman-orthogonal score $\psi(Z; \beta, \eta)$, where $\eta = (g, \alpha)$. Then (1), given a sample $(Z_i := (Y_i, D_i, X_i))_{i=1}^n$, randomly partition the sample into folds $(I_\ell)_{\ell=1}^L$ of approximately equal size. Denote by I_ℓ^c the complement of I_ℓ. (2) For each ℓ, estimate $\widehat{\eta}_\ell = (\widehat{g}_\ell, \widehat{\alpha}_\ell)$ from observations in I_ℓ^c. (3) Estimate β as a root of: $0 =$*

---

[4]Or other values less than 1, if one is willing to entertain non-adversarial confounding.

$n^{-1} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(\beta, Z_i; \widehat{\eta}_\ell)$. *Output $\widehat{\beta}$ and the estimated scores $\widehat{\psi}^o(Z_i) = \psi(\widehat{\beta}, Z_i; \widehat{\eta}_\ell)$ for each $i \in I_\ell$ and each ℓ.*

Therefore the estimators are defined as

$$\widehat{\theta}_s := \mathrm{DML}(\psi_\theta); \quad \widehat{\sigma}_s^2 := \mathrm{DML}(\psi_{\sigma^2}); \quad \widehat{\nu}_s^2 := \mathrm{DML}(\psi_{\nu^2});$$

for the scores

$$\psi_\theta(Z; \theta, g, \alpha) := m(W^s, g) + (Y - g(W^s))\alpha(W^s) - \theta;$$
$$\psi_{\sigma^2}(Z; \sigma^2, g) := (Y - g(W^s))^2 - \sigma^2;$$
$$\psi_{\nu^2}(Z; \nu^2, \alpha) := (2m(W^s, \alpha) - \alpha^2) - \nu^2.$$

We say that an estimator $\widehat{\beta}$ of β is asymptotically linear and Gaussian with the centered influence function $\psi^o(Z)$ if

$$\sqrt{n}(\widehat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi^o(Z_i) + o_{\mathrm{P}}(1) \rightsquigarrow N(0, \mathrm{E}\psi_0^2(Z)).$$

The application of the results in Chernozhukov et al. [2017] for linear score functions yields the following result.

Lemma 3 (DML for Bound Components). *Suppose that each of ψ's listed above and the machine learners $\widehat{\eta}_\ell = (\alpha_\ell, g_\ell)$ of $\eta_0 = (g_s, \alpha_s)$ in $L^2(P_{W^s})$ obey Assumptions 3.1 and 3.2 in Chernozhukov et al. [2017], in particular the rate of learning η_0 in the $L^2(P_{W^s})$ norm needs to be $o_P(n^{-1/4})$. Then the estimators are asymptotically linear and Gaussian with influence functions:*

$$\psi_\theta^o(Z) := \psi_\theta(Z; \theta_s, g_s, \alpha_s); \quad \psi_{\sigma^2}^o(Z) := \psi_{\sigma^2}(Z; \sigma_s^2, g_s); \quad \psi_{\nu^2}^o(Z) := \psi_{\nu^2}(Z; \nu_s^2, \alpha_s).$$

The covariance of the scores can be estimated by the empirical analogues using the covariance of the estimated scores.

The resulting plug-in estimator for the bounds is then:

$$\widehat{\theta}_\pm = \widehat{\theta}_s \pm \widehat{S} |\rho| C_g C_\alpha, \quad \widehat{S}^2 = \widehat{\sigma}_s^2 \widehat{\nu}_s^2.$$

Theorem 3 (DML Confidence Bounds for Bounds). *The bounds estimator $\widehat{\theta}_\pm$ is also asymptotically linear and Gaussian with the influence function:*

$$\varphi_\pm^o(Z) = \psi_\theta^o(Z) \pm \frac{|\rho|}{2} \frac{C_g C_\alpha}{S} (\sigma_s^2 \psi_{\nu^2}^o(Z) + \nu_s^2 \psi_{\sigma^2}^o(Z)).$$

Therefore, the confidence bound

$$[\ell, u] = \left[\widehat{\theta}_- - \Phi^{-1}(1 - a_-/2)\sqrt{\frac{\mathrm{E}\varphi_-^{o2}}{n}}, \ \widehat{\theta}_+ + \Phi^{-1}(1 - a_+/2)\sqrt{\frac{\mathrm{E}\varphi_+^{o2}}{n}} \right)\right]$$

covers $[\theta_-, \theta_+]$ *with probability* $1 - a_- - a_+ - o(1)$. *The same results continue to hold if* $\mathrm{E}\varphi_{\pm}^{o2}(Z)^2$ *are replaced by the empirical analogue*

$$\frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\varphi}_{\pm}^{o2}(Z_i).$$

Remark 6 (**Confidence Bounds**). If we are interested in one-sided confidence bound, we can set either $a_- = 0$ or $a_+ = 0$. The above confidence bound covers $[\theta_-, \theta_+]$ with a prescribed probability $1 - a = 1 - a_+ - a_-$ asymptotically. If the goal is to cover a fixed value in $\theta \in [\theta_-, \theta_+]$ with a prescribed probability, then it is possible to use the approach of Stoye [2009], that refines Imbens and Manski [2004]. This approach allows us to use the less conservative critical value $\Phi^{-1}(1 - a)$ instead of the more conservative $\Phi^{-1}(1 - a/2)$ when the bias bound B is bounded away from zero. In practice, this results in shorter two-sided confidence intervals than the approach above. $\qquad \square$

The following remark discusses learning the regression function g_s and the Riesz representer α_s.

Remark 7 (**Machine Learning of** α_s **and** g_s). Estimation of the short RR g_s is standard and a variety of modern methods can be used (neural networks, random forests, penalized regressions). Estimation of the short RR α_s can proceed in one of the following ways. First, we can use analytical formulas for α_s, see e.g., Chernozhukov et al. [2017], Semenova and Chernozhukov [2021] for practical details and references therein. Second, we can use variational characterization of α_s:

$$\alpha_s = \arg \min_{\alpha \in \mathcal{G}} \mathrm{E}[\alpha^2(W^s) - 2m(W^s, \alpha)],$$

proposed in Chernozhukov et al. [2021a, 2018b]. This avoids inverting propensity scores or conditional densities in the analytical approach. This approach is motivated by the first-order-conditions of the above variational characterization:

$$\mathrm{E}\alpha_s g = \mathrm{E}m(W^s, g) \quad \text{for all } g \text{ in } \mathcal{G},$$

which is the definition of the RR. Neural network (RieszNet) and random forest (ForestRiesz) implementations of this approach are given in Chernozhukov et al. [2021b], and the Lasso implementation in Chernozhukov et al. [2018b]. Third, we may use a minimax (adversarial) characterization of α_s, as in Chernozhukov et al. [2020, 2018a]:

$$\alpha_s = \arg \min_{\alpha \in \mathcal{G}} \max_{g \in \mathcal{G}} |\mathrm{E}m(Z, g) - \mathrm{E}\alpha g|.$$

This also avoids inverting propensity scores. The neural network implementation of this approach is given in Chernozhukov et al. [2020]. The Dantzig selector implementation of this approach is given in Chernozhukov et al. [2018a]. $\qquad \square$

5. Details of Leading Examples

We were using the ACE and ACD as working examples. Here we provide more general example classes, covering a wide variety of interesting and important causal estimands. The presentation of examples draws on Chernozhukov et al. [2018a].

5.1. **Examples.** We first present some examples for the binary treatment case, with the understanding that finitely discrete treatments can be analyzed similarly. Recall that we use $W = (D, X, A)$ to denote the "long" set of regressors and $W^s = (D, X)$ to denote the "short" list of regressors.

Example 1 (**Weighted Average Potential Outcome**). *Let $D \in \{0, 1\}$ be the indicator of the receipt of the treatment. Define the long parameter as*

$$\theta = \mathrm{E}[g(\bar{d}, X, A)\ell(W^s)],$$

where $w^s \mapsto \ell(w^s)$ is a bounded nonnegative weighting function and $\bar{d}$ is a fixed value in $\{0, 1\}$. We define the short parameter as

$$\theta_s = \mathrm{E}[g_s(\bar{d}, X)\ell(W^s)].$$

We assume $\mathrm{E}Y^2 < \infty$ and the weak overlap condition

$$\mathrm{E}[\ell^2(W^s)/P(D = \bar{d} \mid X, A)] < \infty.$$

The long parameter is a weighted average potential outcome (PO) when we set the treatment to $\bar{d}$, under the standard conditional exogeneity assumption (6). The short parameter is a statistical approximation based on the short regression.

In this example, setting

- $\ell(w^s) = 1$ gives the average PO in the entire population;
- $\ell(w^s) = 1(x \in \mathcal{N})/P(X \in \mathcal{N})$ the average PO for group $\mathcal{N}$;
- $\ell(w^s) = 1(d = 1)/P(D = 1)$ the average PO for the treated.

Above we can consider $\mathcal{N}$ as small regions shrinking in volume with the sample size, to make the averages local, as in Chernozhukov et al. [2018a], but for simplicity we take them as fixed in this paper.

Example 2 (**Weighted Average Treatment Effects**). *In the setting of the previous example, define the long parameter*

$$\theta = \mathrm{E}[(g(1, W) - g(0, W))\ell(W^s)],$$

and the short parameter as

$$\theta_s = \mathrm{E}[g_s(1, W) - g_s(0, W))\ell(W^s)].$$

We further assume $\mathrm{E}Y^2 < \infty$ *and the weak overlap condition*

$$\mathrm{E}[\ell^2(W^s)/\{P(D = 0 \mid X, A)P(D = 1 \mid X, A)\}] < \infty.$$

The long parameter is a weighted average treatment effect under the standard conditional exogeneity assumption.

In this example, setting

- $\ell(w^s) = 1$ gives ACE in the entire population;
- $\ell(w^s) = 1(x \in \mathcal{N})/P(X \in \mathcal{N})$ the ACE for group $\mathcal{N}$;
- $\ell(w^s) = 1(d = 1)/P(D = 1)$ the ACE for the treated;
- $\ell(x) = \pi(x)$ the average value of policy (APV) π,

where the policy π assigns a fraction $0 \le \pi(x) \le 1$ of the subpopulation with observed covariate value x to receive the treatment.

In what follows D does not need to be binary. We next consider a weighted average effect of changing observed covariates W^s according to a transport map $w \mapsto T(w^s)$, where T is deterministic measurable map from $\mathcal{W}^s$ to $\mathcal{W}^s$. For example, the policy

$$T(W^s) = (D + 1, X, A)$$

adds a unit to the treatment D. This has a causal interpretation if the policy induces the equivariant change in the regression function, namely the counterfactual outcome $\tilde{Y}$ under the policy obeys $\mathrm{E}[\tilde{Y}|X, A] = g(T(W^s), A)$, and the counterfactual covariates are given by $\tilde{W} = (T(W^s), A)$.

Example 3 (**Average Policy Effect from Transporting** W^s). *For a bounded weighting function* $w^s \mapsto \ell(w^s)$, *the long parameter is given by*

$$\theta = \mathrm{E}[\{g(T(W^s), A) - g(W^s, A)\}\ell(W^s)].$$

The short form of this parameter is

$$\theta_s = \mathrm{E}[\{g_s(T(W^s)) - g_s(W^s)\}\ell(W^s)].$$

As the regularity conditions we require that the support of $P_{\tilde{W}} = \mathrm{Law}(T(W^s), A)$ *is included in the support of* P_W, *and require the weak overlap condition*

$$\mathrm{E}[(\ell(dP_{\tilde{W}} - dP_W)/dP_W)^2] < \infty.$$

We now turn to examples with continuous treatments D taking values in $\mathbb{R}^k$. Consider the average causal effect of the policy that shifts the distribution of covariates via the map $W = (D, X, A) \mapsto T(W^s) = (D + rt(W^s), X, A)$ weighted by $\ell(W^s)$, keeping the long regression function invariant. The following long parameter θ is an approximation to $1/r$ times this average causal effect for small values of r. This example is a differential version of the previous example.

Example 4 (**Weighted Average Incremental Effects**). *Consider the long parameter taking the form of the average directional derivative:*

$$\theta = \mathrm{E}[\ell(W^s)t(W^s)'\partial_d g(D, X, A)],$$

where ℓ is a bounded weighting function and t is a bounded direction function. The short form of this parameter is

$$\theta_s = \mathrm{E}[\ell(W^s)t(W^s)'\partial_d g(D, X)].$$

As regularity conditions, we suppose that $\mathrm{E}Y^2 < \infty$. Further for each (x, a) in the support of (X, A), and each d in $\mathcal{D}_{x,a}$, the support of D given $(X, A) = (x, a)$, the derivative maps $d \mapsto \partial_d g(d, x, a)$ and $d \mapsto g(w)\omega(w)$, for $\omega(w) := \ell(d, x)t(d, x)f(d|x, a)$, are continuously differentiable; the set $\mathcal{D}_{x,a}$ is bounded, and its boundary is piecewise-smooth. Moreover we assume the weak overlap:

$$\mathrm{E}[(\mathrm{div}_d \omega(W)/f(D|X, A))^2] < \infty.$$

Another example is that of a policy that shifts the entire distribution of observed covariates, independently of A. The following long parameter corresponds to the average causal contrast of two policies that set the distribution of observed covariates W^s to F_0 and F_1, independently of A. Note that this example is different from the transport example, since here the dependence between A and W^s is eliminated under the interventions.

Example 5 (**Policy Effect from Changing Distribution of W^s**). *Define the long parameter as*

$$\theta = \int \left[\int g(w^s, a)dP_A(a) \right] \ell(w^s)d\mu(w^s); \quad \mu(w^s) = F_1(w^s) - F_0(w^s),$$

where ℓ is a bounded weight function, and the short parameter as

$$\theta_s = \int g_s(w^s)\ell(w^s)d\mu(w^s); \quad \mu(w^s) = F_1(w^s) - F_0(w^s).$$

As the regularity conditions we require that the supports of F_0 and F_1 are contained in the support of W^s, and that the measure $dP_A \times dF_k$ is absolutely continuous with respect to the measure dP_W on $\mathcal{A} \times \mathrm{support}(\ell)$. We further assume that $\mathrm{E}Y^2 < \infty$ and the weak overlap:

$$\mathrm{E}[(\ell[dP_A \times d(F_1 - F_0)]/dP)^2] < \infty.$$

The following main result for this section establishes that the OVB formulas and bounds are valid.

Lemma 4 (**OVB Validity in Examples 1-5**). *Under the conditions stated in Examples 1-5, Assumptions 1 and 2 are satisfied. Therefore, the general OVB formulas and bounds are valid, with the m-scores and RR described below.*

5.2. **Score and RRs for the Examples.** The m-scores in Examples 1-4 are given by:

(1) $m(w, g) = (g(\bar{d}, x, a))\ell(w^s)$;
(2) $m(w, g) = (g(1, x, a) - g(0, x, a))\ell(w^s)$;
(3) $m(w, g) = (g(T(w^s), a) - g(w^s, a))\ell(w^s)$;
(4) $m(w, g) = \ell(w^s)t(w^s)'\partial_d g(w)$;
(5) $m(w, g) = m(g) = \int[\int g(w^s, a)dP_A(a)]\ell(w^s)d\mu(w^s)$;

and the short m-scores are given by:

(1) $m(w^s, g_s) = (g_s(\bar{d}, x))\ell(w^s)$;
(2) $m(w^s, g_s) = (g_s(1, x) - g_s(0, x))\ell(w^s)$;
(3) $m(w_s, g) = (g_s(T(w^s)) - g_s(w^s))\ell(w^s)$;
(4) $m(w^s, g_s) = \ell(w^s)t(w^s)'\partial_d g_s(w^s)$;
(5) $m(w^s, g_s) = m(g_s) = \int g_s(w^s)\ell(w^s)d\mu(w^s)$.

The long RR are given by:

(1) $\alpha(w) = [(1(d = \bar{d}))/p(\bar{d} \mid x, a)]\bar{\ell}(x, a)$;
(2) $\alpha(w) = [(1(d = 1) - 1(d = 0))/p(d \mid x, a)]\bar{\ell}(x, a)$;
(3) $\alpha(w) = [(dP_{\tilde{W}}(w) - dP_W(w))/dP(w)]\ell(w^s)$;
(4) $\alpha(w) = -(\text{div}_d(\ell(w^s)t(w^s)f(d|x, a)))/f(d|x, a)$;
(5) $\alpha(w) = [dP_A(a) \times d(F_1(w^s) - F_0(w^s))/dP(w)]\ell(w^s)$;

and the short RR are given by:

(1) $\alpha_s(w^s) = [(1(d = \bar{d}))/p(\bar{d} \mid x)]\bar{\ell}(x)$;
(2) $\alpha_s(w^s) = [(1(d = 1) - 1(d = 0))/p(\bar{d} \mid x)]\bar{\ell}(x)$;
(3) $\alpha_s(w^s) = [(dP_{\tilde{W}_s}(w^s) - dP_{W^s}(w^s))/dP_{W^s}(w^s)]\ell(w^s)$
(4) $\alpha_s(w^s) = -(\text{div}_d(\ell(w^s)t(w^s)f(d|x)))/f(d|x)$;
(5) $\alpha_s(w^s) = [d(F_1(w^s) - F_0(w^s))/dP_{W^s}(w^s)]\ell(w^s)$;

where above we used the notations: $\bar{\ell}(X, A) := \mathrm{E}[\ell(W^s)|X, A], \bar{\ell}(X) := \mathrm{E}[\ell(W^s)|X], p(d \mid x, a) := \mathrm{P}(D = d|X = x, A = a), p(d \mid x) := \mathrm{P}(D = d|X = x)$. In Examples 1-2, when the weight function only depends on X, namely $\ell(W^s) = \ell(X)$, we have the simplifications $\bar{\ell}(X, A) = \bar{\ell}(X) = \ell(X)$.

6. EXTENSIONS AND CONCLUSIONS

We have derived simple, practical, yet sharp bounds for a rich class of estimands in causal models with unobserved confounders, and operationalized inference using debiased machine learning methods.

The causal estimands we address in this paper are given by linear functionals of the long regression that one would have run had they observed the latent confounders. These results can potentially be extended to nonlinear functionals. For example, consider a variant of the IV problem [Imbens and Angrist, 1994], where the instrumental variable Z is confounded by observed covariates X, and latent variables A. In this case, the IV estimand is given by the ratio of two average causal effects,

$$\text{IV} = \frac{ACE(Z \to Y)}{ACE(Z \to D)}$$

The numerator and denominator can be bounded using the methods for bounding the ACE proposed in this paper.

Another potentially interesting direction of investigation is to consider causal estimands that are functionals of the long quantile regression, or causal estimands that are values of a policy in dynamic stochastic programming. When the degree of confounding is small, it seems possible to use the results in Chernozhukov et al. [2016] to derive approximate bounds on the bias that can be estimated using debiased ML approaches. Another interesting direction for further explorations is the use of shape restrictions on the long regression g that can potentially sharpen the bounds.

Appendix A. Preliminaries

A.1. **Nonparametric R^2 and Linear R^2.** By the ANOVA theorem we can decompose the variance of Y as the variance explained by covariates W and the unexplained variance

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathrm{E}[Y|W]) + \mathrm{E}[\mathrm{Var}(Y|W)].$$

We can thus define the *nonparametric R^2*, $\eta^2_{Y\sim W}$, by

$$\eta^2_{Y\sim W} := \frac{\mathrm{Var}(\mathrm{E}[Y|W])}{\mathrm{Var}(Y)}$$

Note that $\eta^2_{Y\sim W}$ can also be written as the linear R^2 between Y and the CEF

$$\eta^2_{Y\sim W} = \mathrm{Cor}^2(Y, \mathrm{E}[Y|W]) = R^2_{Y\sim\mathrm{E}[Y|W]}.$$

For covariates $W = (A, D, X)$ and $W^s = (D, X) \subset W$ we define the nonparametric partial R^2 of A with Y given $W^s = (D, X)$, $\eta^2_{Y\sim A|W^s}$, as

$$\eta^2_{Y\sim A|W^s} = \frac{\mathrm{Var}(\mathrm{E}[Y|W]) - \mathrm{Var}(\mathrm{E}[Y|W^s])}{\mathrm{Var}(Y) - \mathrm{Var}(\mathrm{E}[Y|W^s])} = \frac{\eta^2_{Y\sim W} - \eta^2_{Y\sim W^s}}{1 - \eta^2_{Y\sim W^s}},$$

The nonparameteric partial R^2 can also be written as the linear R^2

$$\eta^2_{Y\sim A|W^s} = \mathrm{Cor}^2(Y - \mathrm{E}[Y|W^s], \mathrm{E}[Y|W] - \mathrm{E}[Y|W^s]) = R^2_{Y-g_s\sim g-g_s}.$$

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

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

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

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

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

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

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

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

where S is the surface measure induced by V.

APPENDIX B. DEFERRED PROOFS

B.1. **Proof of Theorem 1, Corollary 1 and Corollary 2.** The result follows from

$$\mathrm{E}g\alpha - \mathrm{E}g_s\alpha_s = \mathrm{E}(g_s + g - g_s)(\alpha_s + \alpha - \alpha_s) - \mathrm{E}g_s\alpha_s$$

$$= \mathrm{E}g_s(\alpha - \alpha_s) + \mathrm{E}\alpha_s(g - g_s) + \mathrm{E}(g - g_s)(\alpha - \alpha_s)$$

$$= \mathrm{E}(g - g_s)(\alpha - \alpha_s),$$

using the fact that α_s is orthogonal to $g - g_s$ and g_s is orthogonal to $\alpha - \alpha_s$ by definition of α, α_s and g_s.

Corollary 1 follows from observing that

$$\frac{\mathrm{E}(g - g_s)^2}{\mathrm{E}(Y - g_s)^2} = R^2_{Y - g_s \sim g - g_s} = R^2_{\tilde{Y}_s \sim A_1}$$

and from

$$\frac{\mathrm{E}(\alpha - \alpha_s)^2}{\mathrm{E}\alpha_s^2} = \frac{\mathrm{E}\alpha^2 - \mathrm{E}\alpha_s^2}{\mathrm{E}\alpha_s^2} = \frac{1/\mathrm{E}\tilde{D}^2 - 1/\mathrm{E}\tilde{D}_s^2}{1/\mathrm{E}\tilde{D}_s^2} = \frac{\mathrm{E}\tilde{D}_s^2 - \mathrm{E}\tilde{D}^2}{\mathrm{E}\tilde{D}^2} = \frac{R^2_{\tilde{D}_s \sim A_2}}{1 - R^2_{\tilde{D}_s \sim A_2}},$$

where $\tilde{D} := D - \mathrm{E}[D \mid X, A]$. Here we used the observation that

$$\mathrm{E}(\alpha - \alpha_s)^2 = \mathrm{E}\alpha^2 + \mathrm{E}\alpha_s^2 - 2\mathrm{E}\alpha\alpha_s = \mathrm{E}\alpha^2 - \mathrm{E}\alpha_s^2,$$

holding because

$$\mathrm{E}\alpha\alpha_s = \frac{\mathrm{E}\tilde{D}\tilde{D}_s}{\mathrm{E}\tilde{D}^2\mathrm{E}\tilde{D}_s^2} = \frac{\mathrm{E}\tilde{D}^2}{\mathrm{E}\tilde{D}^2\mathrm{E}\tilde{D}_s^2} = \frac{1}{\mathrm{E}D_s^2} = \mathrm{E}\alpha_s^2.$$

Corollary 2 follows immediately from the definitions of η^2, since $R^2_{\tilde{Y}_s \sim A_1} = \eta^2_{Y \sim A|D,X}$ and $R^2_{\tilde{D}_s \sim A_2} = \eta^2_{D \sim A|X}$.

To show the bound is sharp we need to show that

$$1 = \max\{\rho^2 \mid (\alpha, g) : \mathrm{E}(\alpha - \alpha_s)^2 = B_\alpha^2, \quad \mathrm{E}(g - g_s)^2 = B_g^2\},$$

where B_α and B_g are nonnegative constants such that $B_g^2 \leq \mathrm{E}(Y - g_s)^2$. To do so, choose any g of the partially linear form such that $\mathrm{E}(g - g_s)^2 = B_g^2$, then set

$$\alpha - \alpha_s = B_\alpha(g - g_s)/B_g.$$

This yields an admissible RR, and sets $\rho^2 = 1$. □

B.2. **Proof of Lemma 1.** The existence of the unique long RR $\alpha \in L^2(P_W)$ follows from the Riesz-Frechet representation theory. To show that we can take $\alpha_s(W^s) := \mathrm{E}[\alpha(W) \mid W^s]$ to be the short RR, we first observe that the long RR obeys

$$\mathrm{E}m(W, g_s) = \mathrm{E}g_s(W^s)\alpha(W)$$

for all $g_s \in L^2(P_{W^s})$. That is, the long RR α can represent the linear functionals over the smaller space $L^2(P_{W^s}) \subset L^2(P_W)$, but α itself is not in $L^2(P_{W^s})$. Then, we decompose the long RR into the orthogonal projection α_s and the residual e:

$$\alpha(W) = \alpha_s(W^s) + e(W); \quad \mathrm{E}e(W)g_s(W) = 0, \text{ for all } g_s \text{ in } L^2(P_{W^s}).$$

Then

$$\mathrm{E}g_s(W)\alpha(W) = \mathrm{E}\big[g_s(W^s)\big(\alpha_s(W^s) + e(W^s)\big)\big],$$
$$= \mathrm{E}\big[g_s(W^s)\alpha_s(W^s)\big].$$

Therefore $\mathrm{E}[\alpha(W) \mid W^s]$ is a short RR, and it is unique in $L^2(P_{W^s})$ by the RF theory. We also have that $\mathrm{E}\alpha^2 = \mathrm{E}\alpha_s^2 + \mathrm{E}e^2$, establishing that $\mathrm{E}\alpha^2 \geq \mathrm{E}\alpha_s^2$. □

B.3. **Proof of Lemma 2.** We have from the Riesz-Frechet theory that

$$\mathrm{E}m(W, g_r) = \mathrm{E}g_r(W)\alpha(W),$$

for all $g_r \in \Gamma$, that is the RR α continues to represent the functional over the restricted linear subspace $\Gamma \subset L^2(P_W)$. Decompose α in the orthogonal projection $\bar{\alpha}$ and the residual e:

$$\alpha(W) = \bar{\alpha}(W) + e(W), \quad \mathrm{E}e(W)g_r(W) = 0, \text{ for all } g_r \text{ in } \Gamma.$$

Then we have that

$$\mathrm{E}g_r(W)\alpha(W) = \mathrm{E}g_r(W)\bar{\alpha}(W) + \mathrm{E}g_r(W)e(W) = \mathrm{E}g_r(W)\alpha_r(W).$$

That is, $\bar{\alpha}$ is a RR, and it is unique in Γ by the RF theory. We also have that $\mathrm{E}\alpha^2 = \mathrm{E}\bar{\alpha}^2 + \mathrm{E}e^2$, establishing that $\mathrm{E}\alpha^2 \geq \mathrm{E}\bar{\alpha}^2$.

Analogous argument yields the result for the closed linear subsets Γ_s of $L^2(P_{W^s})$.

Here we show that $\bar{\alpha}_s$ is given by a projection of $\bar{\alpha}$ onto Γ_s. Indeed, $\bar{\alpha}$ represents the functionals over Γ_s but it is not itself in Γ_s. However, its projection onto Γ_s therefore can also represent the functionals, using the same arguments as above. By uniqueness of the RR over Γ_s, we must have that the projected $\bar{\alpha}$ coincides with $\bar{\alpha}_s$. Further,

$$\mathrm{E}(\bar{\alpha} - \bar{\alpha}_s)^2 \geq \min_{b \in \mathbb{R}} \mathrm{E}(\bar{\alpha} - b\bar{\alpha}_s)^2 \geq \min_{a \in \Gamma_s} \mathrm{E}(\bar{\alpha} - a)^2 = \mathrm{E}(\bar{\alpha} - \bar{\alpha}_s)^2.$$

This shows that the linear orthogonal projection of $\bar{\alpha}$ on $\bar{\alpha}_s$ is given by $\bar{\alpha}_s$. The latter means that we can decompose:

$$\mathrm{E}(\bar{\alpha} - \bar{\alpha}_s)^2 = \mathrm{E}\alpha^2 - \mathrm{E}\alpha_s^2. \quad \square$$

B.4. **Proof of Theorem 2 and Corollary 3.** We decompose $L^2(P_W)$ into $L^2(P_{W^s})$ and its orthocomplement $L^2(P_{W^s})^\perp$,

$$L^2(P_W) = L^2(P_{W^s}) + L^2(P_{W^s})^\perp.$$

So that any element $m_s \in L^2(P_{W^s})$ is orthogonal to any $e \in L^2(P_{W^s})^\perp$ in the sense that

$$\mathrm{E}m_s(W^s)e(W) = 0.$$

The claim of the theorem follows from

$$\mathrm{E}g\alpha - \mathrm{E}g_s\alpha_s = \mathrm{E}(g_s + g - g_s)(\alpha_s + \alpha - \alpha_s) - \mathrm{E}g_s\alpha_s$$
$$= \mathrm{E}g_s(\alpha - \alpha_s) + \mathrm{E}\alpha_s(g - g_s) + \mathrm{E}(g - g_s)(\alpha - \alpha_s)$$
$$= \mathrm{E}(g - g_s)(\alpha - \alpha_s),$$

using the fact that $\alpha_s \in L^2(P_{W^s})$ is orthogonal to $g - g_s \in L^2(P_{W^s})^\perp$ and $g_s \in L^2(P_{W^s})$ is orthogonal to $\alpha - \alpha_s \in L^2(P_{W^s})^\perp$.

Corollary 3 follows from observing that

$$\frac{\mathrm{E}(g - g_s)^2}{\mathrm{E}(Y - g_s)^2} = R^2_{Y - g_s \sim g - g_s},$$

as before, and from

$$\frac{\mathrm{E}(\alpha - \alpha_s)^2}{\mathrm{E}\alpha_s^2} = \frac{\mathrm{E}\alpha^2 - \mathrm{E}\alpha_s^2}{\mathrm{E}\alpha_s^2} = \frac{\mathrm{E}\alpha^2 - \mathrm{E}\alpha_s^2}{\mathrm{E}\alpha^2} \frac{\mathrm{E}\alpha^2}{\mathrm{E}\alpha_s^2} = \frac{1 - R^2_{\alpha \sim \alpha_s}}{R^2_{\alpha \sim \alpha_s}}.$$

The proof for the case where g's and α's are restricted follows similarly, replacing $L^2(P_W)$ with $\Gamma \subset L^2(P_W)$ and $L^2(P_{W^s})$ with $\Gamma_s = \Gamma \cap L^2(P_{W_s})$, and decomposing $\Gamma = \Gamma_s + \Gamma_s^\perp$, where $\Gamma_s^\perp$ is the orthogonal complement of Γ_s relative to Γ. The remaining arguments are the same, utilizing Lemma 2.

To show the bound is sharp we need to show that

$$1 = \max\{\rho^2 \mid (\alpha, g) : \mathrm{E}(\alpha - \alpha_s)^2 = B_\alpha^2, \quad \mathrm{E}(g - g_s)^2 = B_g^2\},$$

where B_α and B_g are nonnegative constants such that $B_g^2 \leq \mathrm{E}(Y - g_s)^2$. To do so, choose any α of an admissible form such such that $\mathrm{E}(\alpha - \alpha_s)^2 = B_\alpha^2$, then set

$$g - g_s = B_g(\alpha - \alpha_s)/B_\alpha.$$

This yields an admissible long regression function, and sets $\rho^2 = 1$. $\quad \square$

Remark 8. We note here that distribution of observed data P can place other restrictions on the problem, restricting admissible values of B_α^2 or B_g^2 or $\rho^2 < 1$. For example, we have $0 \leq g, g_s \leq 1$ when $0 \leq Y \leq 1$. This implies $\|g - g_s\|_\infty \leq 1$, which can potentially result in the adversarial $\rho^2 < 1$. $\quad \square$

B.5. **Proof of Lemma 3 and Theorem 3.** The Lemma follows from the application of Theorem 3.1 and Theorem 3.2 in Chernozhukov et al. [2017]. Valid estimation of covariance follows similarly to the proof of Theorem 3.2 in Chernozhukov et al. [2017]. The first result of Theorem 3 follows from the delta method in van der Vaart and Wellner [1996]. The validity of the confidence intervals follows from using the standard arguments for confidence intervals based on asymptotic normality. $\qquad\square$

B.6. **Proof of Lemma 4.** Here the argument is similar to Chernozhukov et al. [2018a], but we provide details for completeness.

The assumptions directly imply that the candidate long RR obey $\alpha \in L^2(P)$ with $\|\alpha\|_{P,2} \leq C$ in each of the examples, for some constant C that depends on P. By $EY^2 < \infty$, we have $g \in L^2(P)$. Therefore, $|E\alpha g| < \|\alpha\|_{P,2}\|g\|_{P,2} < \infty$ in any of the calculations below.

We first verify that long RR α's can indeed represent the functionals $g \mapsto \theta(g) := Em(W, g)$ in Examples 1,2,3,5 over $g \in L^2(P)$. In Example 4, the long RR represents the Hanh-Banach extension of the mapping $g \mapsto \theta(g)$ to $L^2(P)$ over $L^2(P)$.

In Example 1, recall that $\bar{\ell}(X, A) := E[\ell(W^s)|X, A]$. Then since $dP(d, x, a) = \sum_{j=0}^1 1(j = d)P[D = j|X = x, A = a]dP(x, a)$ by the Bayes rule, we have

$$Eg(W)\alpha(W) = \int g(d, x, a) \frac{1(d = \bar{d})\bar{\ell}(x, a)}{P[D = \bar{d}|X = x, A = a]} dP(d, x, a)$$

$$= \int g(\bar{d}, x, a)\bar{\ell}(x, a)dP(x, a) = Eg(\bar{d}, X, A)\bar{\ell}(X, A) = Eg(\bar{d}, X, A)\ell(W^s) = \theta(g),$$

where the penultimate equality follows by the law of iterated expectations. The claim for Example 2 follows from the claim for Example 1.

Example 3 follows by the change of measure of dP_W to dP_{W^s}, given the assumed absolutely continuity of the former with respect to the latter. Then we have

$$Eg(W)\alpha(W) = \int g\ell \left(\frac{dP_{\tilde{W}} - dP_W}{dP_W}\right) dP_W$$

$$= \int g\ell(dP_{\tilde{W}} - dP_W) = \int \ell(w^s)(g(T(w^s), a) - g(w^s, a))dP_W(w) = \theta(g).$$

In Example 4, we can write for any $g's$ that have the properties stated in this example:

$$Eg(W)\alpha(W) = -\int \int g(w) \frac{\text{div}_d(\ell(w^s)t(w^s)f(d|x, a))}{f(d|x, a)} f(d|x, a)\text{d}d\text{d}P(x, a)$$

$$= -\int \int g(w)\text{div}_d(\ell(w^s)t(w^s)f(d|x, a))\text{dd}d P(x, a)$$

$$= \int \int \partial_d g(w)'t(w^s)\ell(w^s)f(d|x, a)\text{dd}dP(x, a) = \theta(g),$$

where we used the integration by parts and that $g(w)\ell(w^s)t(w^s)f(d|x,a)$ vanishes for any d in the boundary of $\mathcal{D}_{x,a}$.

Example 5 follows by the change of measure $dP_A \times dF_k$ to dP_W, given the assumed absolutely continuity of the former with respect to the latter on $\mathcal{A} \times \text{support}(\ell)$. Then we have

$$\mathrm{E}g(W)\alpha(W) = \int g\ell \left(\frac{[dP_A \times d(F_1 - F_0)]}{dP_W} \right) dP_W$$

$$= \int g(w^s, a)\ell(w^s)dP_A(a)d(F_1 - F_0)(w^s) = \theta(g).$$

In all examples, the continuity of $g \mapsto \theta(g)$ required in Assumption 1 now follows from the representation property and from $|\mathrm{E}\alpha g| \leq \|\alpha\|_{P,2}\|g\|_{P,2} \leq C\|g\|_{P,2}$.

Verification of Assumption 2 follows directly from the inspection of m-scores given in Section 5.

Note that we don't need the analytical form of the short RRs to verify Assumptions 1 or 2. However, their analytical form can be found by exactly the same steps as above, or by taking the conditional expectation. $\qquad\square$

References

Joseph G Altonji, Todd E Elder, and Christopher R Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184, 2005.

Joshua D. Angrist and Jorn-Steffan Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, 2009.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press, 1993.

Matthew Blackwell. A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, 22(2):169–182, 2013.

Matteo Bonvini and Edward H Kennedy. Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–11, 2021.

Babette A Brumback, Miguel A Hernán, Sebastien JPA Haneuse, and James M Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine*, 23(5):749–767, 2004.

Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

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

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

Victor Chernozhukov, Whitney Newey, and Rahul Singh. De-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018a.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *arXiv preprint arXiv:1809.05224*, 2018b.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Victor Chernozhukov, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737*, 2021a.

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests, 2021b.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 39–67, 2020a.

Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Work. Pap*, 2020b.

Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. *International Conference on Machine Learning*, 2019.

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.

Gianluca Detommaso, Michael Brückner, Philip Schulz, and Victor Chernozhukov. Causal bias quantification for continuous treatment, 2021.

Kjell Doksum and Alexander Samarov. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, pages 1443–1473, 1995.

Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470, 2016.

Kenneth A Frank. Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2):147–194, 2000.

Kenneth A Frank, Spiro J Maroulis, Minh Q Duong, and Benjamin M Kelcey. What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, 35(4):437–460, 2013.

AlexanderM Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical*

Association, 115(532):1730–1746, 2020.

Carrie A Hosman, Ben B Hansen, and Paul W Holland. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, pages 849–870, 2010.

Kosuke Imai, Luke Keele, Teppei Yamamoto, et al. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 25(1):51–71, 2010.

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62:467–475, 1994.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *arXiv preprint arXiv:2103.04850*, 2021.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *arXiv preprint arXiv:1805.08593*, 2018.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019.

Joel A Middleton, Marc A Scott, Ronli Diakow, and Jennifer L Hill. Bias amplification and bias unmasking. *Political Analysis*, 24(3):307–323, 2016.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, pages 1–18, 2017.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. Causal inference in statistics: an overview. *Stat. Surv.*, 3:96–146, 2009a. ISSN 1935-7516. doi: 10.1214/09-SS057. URL http://dx.doi.org/10.1214/09-SS057.

Judea Pearl. *Causality.* Cambridge university press, 2009b.

Karl Pearson. On the general theory of skew correlation and non-linear regression. *Dulau and Company*, 1905.

J. Pfanzagl and W. Wefelmeyer. A third-order optimum property of the maximum likelihood estimator. *Journal of Multivariate Analysis*, 8:1–29, 1978.

James M Robins. Association, causation, and marginal structural models. *Synthese*, 121(1):151–179, 1999.

Paul R Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983a.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical*

Society. Series B (Methodological), pages 212–218, 1983b.

Daniel O Scharfstein, Razieh Nabi, Edward H Kennedy, Ming-Yueh Huang, Matteo Bonvini, and Marcela Smid. Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*, 2021.

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

Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*, 2012.

Jörg Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

Tyler J. Vanderweele and Onyebuchi A. Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, 22(1):42–52, January 2011.

Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.