LONG STORY SHORT:
OMITTED VARIABLE BIAS IN CAUSAL MACHINE LEARNING

VICTOR CHERNOZHUKOV[†], CARLOS CINELLI[*], WHITNEY K. NEWEY[‡], AMIT SHARMA[‖],

AND VASILIS SYRGKANIS[§]

ABSTRACT. We develop a general theory of omitted variable bias for a wide range of common causal parameters, including (but not limited to) averages of potential outcomes, average treatment effects, average causal derivatives, and policy effects from covariate shifts. Our theory applies to nonparametric models, while naturally allowing for (semi-)parametric restrictions (such as partial linearity) when such assumptions are made. We show how simple plausibility judgments on the maximum explanatory power of omitted variables are sufficient to bound the magnitude of the bias, thus facilitating sensitivity analysis in otherwise complex, nonlinear models. Finally, we provide flexible and efficient statistical inference methods, which can leverage modern machine learning algorithms for estimation. These results allow empirical researchers to perform sensitivity analyses in a flexible class of machine-learned causal models using very simple, and interpretable, tools. Empirical examples demonstrate the utility of our approach.

1. INTRODUCTION

Unmeasured confounding is a pervasive issue in studies that aim to draw causal inferences from observational data. Such studies typically rely on a conditional ignorability (also known as unconfoundedness) assumption, which states that the treatment assignment is independent of potential outcomes given a set of observed covariates (Rosenbaum and Rubin, 1983a; Pearl, 2009; Angrist and Pischke, 2009; Imbens and Rubin, 2015). This assumption, however, requires that

[†] Dept. of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA. Email: vchern@mit.edu.

[*] Dept. of Statistics, University of Washington, Seattle, WA, USA. Email: cinelli@uw.edu.

[‡] Dept. of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA. Email: wnewey@mit.edu.

[‖] Microsoft Research India, Bangalore, India. Email: amshar@microsoft.com.

[§] Dept. of Mgmt Science and Engineering, Stanford University, Stanford, CA, USA. Email: vsyrgk@stanford.edu.

there are no unobserved confounders influencing both the treatment and the outcome. When such variables are omitted from the analysis, empirical estimates may differ from the true causal effect of interest, giving rise to what is now commonly known as "omitted variable bias."

The omitted variable bias (OVB) problem is one of the most significant threats to the identification of causal effects. In the context of linear models, this bias amounts to the difference between the coefficients of the treatment variable from two distinct outcome regressions: one that controls only for observed covariates (the "short" regression) and another that would additionally control for unobserved variables (the "long" regression). Formulas characterizing this difference play a foundational role in statistics, econometrics, and related fields (see, e.g., discussions in classical and modern textbooks, such as Goldberger, 1991; Angrist and Pischke, 2009 and Wooldridge, 2010).

But while linear models are widely used in applied work, they are often overly restrictive. For example, in the binary treatment case, using linear models when treatment effects are heterogeneous may yield unintuitive or even misleading estimates of the causal effects of interest (Aronow and Samii, 2016; Słoczyński, 2022). To address these limitations, many empirical analysts have turned their attention to more flexible nonlinear or nonparametric models, often leveraging modern machine learning techniques for estimation and inference (van der Laan and Rose, 2011; Belloni et al., 2013; Chernozhukov et al., 2018a; Athey et al., 2019). These tools offer the flexibility to capture complex relationships between variables, avoiding stringent functional form assumptions in causal effect estimation. Yet, though sensitivity analyses for nonparametric models do exist (see related literature below), we currently lack general results characterizing the form of omitted variable bias in nonlinear models, as we have for the linear case. Our work aims to address this gap.

In this paper we develop a general theory of omitted variable bias for a wide range of common causal parameters that can be identified as linear functionals of the conditional expectation function (CEF) of the outcome. Such functionals encompass many (if not most) of the traditional targets of investigation in causal inference studies, such as averages of potential outcomes, average treatment effects, average causal derivatives, and policy effects from covariate shifts. We allow for arbitrary treatment (e.g., continuous or binary) and outcome variables. Our theory applies to general nonparametric models, while naturally allowing for (semi-)parametric restrictions (such as partial linearity) when such assumptions are made. Our formulation recovers well-known and familiar OVB results for linear models as a special case, and it can be seen as its natural generalization to nonlinear models. Importantly, we show that the general nonparametric formula of the bias still has a simple and interpretable form.

More specifically, we first formalize the OVB problem in the nonparametric setting. Paralleling the linear case, we define the OVB as the difference between the "short" and "long" functionals of the outcome regression, where the former omits and the latter includes the latent variables. To derive the OVB, our construction then leverages the Riesz-Frechet representation of the target functionals, which allows us to rewrite the parameters of interest as weighted averages of the outcome regression, with weights given by the Riesz representers (RRs).[1] We show that the OVB arises as a by-product of confounders introducing systematic errors in both the outcome regression and in the RRs for the parameter of interest. Furthermore, plausibility judgments on the maximum explanatory power of latent variables suffice to place overall bounds on the bias, simplifying the task of sensitivity analysis even when using nonparametric or otherwise complex models.

Although these results may initially seem abstract to those not familiar with Riesz representation theory, in many leading examples the RRs in fact correspond to quantities that are well-known to empirical researchers. For instance, when estimating the average treatment effect in a partially linear model, the RR is the (variance scaled) residualized treatment, after "partialling out" the control covariates. Or, when estimating an average treatment effect in a general nonparametric model with a binary treatment, the RR is now given by another familiar quantity—the inverse probability of treatment weights (IPTW). In such cases, we show that the bias can be reparameterized in terms of simple percentage gains in variance explained (or precision) in the treatment and the outcome regression due to unmeasured confounders, again facilitating the interpretation and use of the OVB

[1] The RRs are also important in asymptotic variance formulae for estimators of parameters of interest (Newey, 1994).

formulas in practice. We further help analysts make plausibility judgments on the magnitude of sensitivity parameters by means of comparison of the relative strength of unobserved confounders against the strength of observed covariates.

Finally, we provide statistical inference using debiased machine learning (DML) and auto-DML (Chernozhukov et al., 2018a,b, 2020, 2022c). Our construction makes it possible to use modern machine learning methods for estimating the identifiable components of the omitted variable bias formula, including regression functions, Riesz representers, the norm of regression residuals, and the norm of Riesz representers. These results enable flexible and efficient statistical inference on the bounds of the target parameter, allowing researchers to perform sensitivity analyses against unmeasured confounding in a flexible class of machine-learned causal models using simple and interpretable tools.

Related Literature. Our work is most closely related to the literature that derives OVB formulas for linear models, such as those found in traditional textbooks and recent extensions (Goldberger, 1991; Angrist and Pischke, 2009; Wooldridge, 2010; Frank, 2000; Oster, 2019; Cinelli and Hazlett, 2020, 2025). We advance this literature by providing analogous, easily explainable OVB formulas for a broad and rich class of causal parameters, all for general nonlinear models, *with or without further parametric restrictions*. Importantly, we provide a single unifying framework that covers all these cases, and that can be easily specialized depending on the target parameter and on whether additional parametric assumptions (if any) are made. We further advance the OVB literature by providing flexible and efficient statistical inference methods, leveraging modern machine learning algorithms with debiased machine learning.

More broadly, our work is related to the extensive literature on sensitivity analysis against unmeasured confounders. Here we highlight the key differences between our approach and existing methods, while relegating a more detailed review to the Appendix, Section D. First, many prior works on sensitivity analysis either focus exclusively on binary treatments (e.g., Rosenbaum, 1987; Masten and Poirier, 2018; Zhao et al., 2019; Bonvini and Kennedy, 2021), target a single estimand of interest, such as a causal risk ratio (Ding and VanderWeele, 2016; VanderWeele and Ding, 2017), or impose parametric assumptions on the observed data or on the nature of unobserved confounding (Rosenbaum and Rubin, 1983b; Imbens, 2003; Dorie et al., 2016; Cinelli et al., 2019; Zhang et al., 2021). Our approach differs from these in that (i) it is not limited to binary treatments, (ii) it covers a broader range of target parameters, such as average causal derivatives and average policy effects from covariate shifts, and (iii) it does not require parametric assumptions on the observed data nor on the nature of confounding.

Even if we focus solely on the important special case of estimating an average treatment effect (ATE) or average treatment effect on the treated (ATT) with a binary treatment, our OVB results usefully complement other seminal approaches to this problem such as those of Rosenbaum (1987) or the marginal sensitivity models of Tan (2006). Whereas such approaches limit the strength of confounding through its impact on the *worst-case* change in the odds ratio of treatment assignment, our OVB approach shows that this assumption is much stronger than what is needed to bound the bias—all that is required by the OVB formula are claims about changes in *averages* (see Section 3.4). Moreover, even in stylized models of treatment assignment (e.g., a logistic model with a Gaussian latent confounder), worst-case approaches such as the ones in Rosenbaum (1987) and Tan (2006) have a naturally unbounded sensitivity parameter, no matter how small the actual degree of confounding is (see Appendix C.9 for an example).

Importantly, our OVB-based approach generally differs from traditional sensitivity analyses in that it derives the exact OVB formula for the target parameters we cover. For example, our results show that the bias of the ATE in the binary treatment case is *not* determined by deviations on the odds of treatment; rather, it is determined by three quantities: (i) the maximum explanatory power of confounders in the treatment regression, as given by gains in precision, (ii) the maximum explanatory power of confounders in the outcome regression, as given by gains in variance explained, and (iii) by the correlation of errors in the regression function and the IPTW. Therefore, beyond being a tool for sensitivity analysis, OVB results such as ours provide a precise characterization of how different estimands respond differently to unmeasured confounding.

Overview of the paper. Section 2 presents our method in the simpler context of partially linear models. Section 3 derives the main result of the paper, i.e., formalize the omitted variable bias problem and derive its formula for the general class of continuous linear functionals of (projections of) the conditional expectation function of the outcome—all for general, nonparametric causal models. In Section 4 we construct high-quality methods for statistical inference under omitted variable bias, by leveraging recent advances in debiased machine learning with Riesz representers. Section 5 demonstrates the use of our tools to assess the robustness of causal claims in a detailed empirical example that estimates the average treatment effect of 401(k) eligibility on net financial assets. Section 6 concludes with suggestions for possible extensions. The appendix contains all proofs, a more extensive literature review, as well as an additional empirical example that illustrates sensitivity analyses for average causal derivatives with continuous treatments.

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

We use $R^2_{V\sim U}$ to denote the R^2 from the orthogonal linear projection of a scalar random variable V on a random vector U. We follow Doksum and Samarov (1995) and define the *nonparametric* R^2:

$$\eta^2_{V\sim U} := \frac{\mathrm{Var}(\mathrm{E}[V \mid U])}{\mathrm{Var}(V)} = 1 - \frac{\mathrm{E}[\mathrm{Var}(V \mid U)]}{\mathrm{Var}(V)} = R^2_{V\sim \mathrm{E}[V|U]},$$

which also equals the linear R^2 of the projection of V on $\mathrm{E}[V \mid U]$. The nonparametric *partial* R^2 of U with another variable V *given* X measures the additional gain in the explanatory power that U provides, beyond what is already explained by X. This equals the relative decrease in the average residual variance, or the linear R^2 of the projection of the residual variation of V onto the residual variation of $\mathrm{E}[V \mid X, U]$:

$$\eta^2_{V\sim U|X} := \frac{\eta^2_{V\sim UX} - \eta^2_{V\sim X}}{1-\eta^2_{V\sim X}} = 1 - \frac{\mathrm{E}[\mathrm{Var}(V \mid X,U)]}{\mathrm{E}[\mathrm{Var}(V \mid X)]} = R^2_{V-\mathrm{E}[V|X]\sim \mathrm{E}[V|X,U]-\mathrm{E}[V|X]}. \tag{1}$$

For a binary random variable V, we use $\mathrm{Odds}(V \mid U) := P(V=1 \mid U)/(1-P(V=1 \mid U))$ to denote the conditional odds of V given random vector U. These metrics will be useful for characterizing the strength of unmeasured confounders in the next sections.

2. Warm-Up: Omitted Variable Bias in Partially Linear Models

To fix ideas, we begin our discussion in the context of partially linear models (PLM). 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.

Suppose an empirical researcher is interested in the regression coefficient θ from a partially linear regression model of the following form

$$Y = \theta D + f(X,U) + \varepsilon, \tag{2}$$

where here Y denotes a real-valued outcome, D a real-valued treatment, X an observed vector of covariates, and U an *unobserved* vector of confounders. We refer to $W := (D,X,U)$ as the "long" list of regressors, and to equation (2) as the "long" regression. This regression model need not be correctly specified, and the error term simply obeys the orthogonality condition $\mathrm{E}[\varepsilon(D-\mathrm{E}[D \mid X,U])] = 0$.[2]

[2] That is $g(W) := \theta D + f(X,U)$ is the projection of the CEF on the space of functions that are partially linear in D.

However, since U is unobserved, the coefficient θ cannot be identified from the observed data. The researcher then naturally approximates θ by estimating the analogous regression coefficient θ_s from a partially linear model that *omits* U:

$$Y = \theta_s D + f_s(X) + \varepsilon_s. \tag{3}$$

As before, this model need not be correctly specified, and we simply have $\mathrm{E}[\varepsilon_s(D - \mathrm{E}[D \mid X])] = 0$. We call equation (3) the "short regression," and $W_s := (D, X)$ the "short" list of observed regressors.

Evidently, in general θ_s is not equal to θ, and this naturally leads to the question of how far this "proxy" parameter deviates from the true inferential target. Our goal is to analyze the difference between the long and short parameters—the omitted variable bias (OVB):

$$\theta - \theta_s,$$

and perform inference on this bias under various hypotheses on the strength of the unobserved confounders U.

2.2. **Omitted variable bias for partially linear models.** Using a Frisch-Waugh-Lovell (FWL) partialling out argument (Frisch and Waugh, 1933; Lovell, 1963), one can express the long and short parameters, θ and θ_s, as the linear regression coefficients of Y on the residuals $D - \mathrm{E}[D \mid X, U]$ and $D - \mathrm{E}[D \mid X]$, respectively.

That is,

$$\theta = \mathrm{E} Y \alpha(W), \qquad \theta_s = \mathrm{E} Y \alpha_s(W_s); \tag{4}$$

where we define

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

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 $g(W) := \theta D + f(X,U)$ and $g_s(W_s) := \theta_s D + f_s(X)$ denote the long and short partially linear projections, respectively. Using the orthogonality conditions in (2) and (3), we can further express θ and θ_s as

$$\mathrm{E}Y\alpha(W) = \mathrm{E}g(W)\alpha(W), \qquad \mathrm{E}Y\alpha_s(W_s) = \mathrm{E}g_s(W_s)\alpha_s(W_s). \tag{5}$$

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—PLM). *Assume that Y and D are square-integrable with* $\mathrm{E}(D - \mathrm{E}[D \mid X,U])^2 > 0$. *Then the OVB for the partially linear regression coefficients of equation (4) is given by the covariance between the regression error and the RR error,*

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

This bias can be further reparameterized in terms of R^2 measures,

$$\theta - \theta_s = \underbrace{\mathrm{Cor}(g - g_s, \alpha - \alpha_s)}_{\rho} \times \underbrace{\sqrt{R^2_{Y-g_s\sim g-g_s}}}_{C_Y} \times \underbrace{\sqrt{\frac{(1-R^2_{\alpha\sim\alpha_s})}{R^2_{\alpha\sim\alpha_s}}}}_{C_D} \times \underbrace{\sqrt{\mathrm{E}(Y-g_s)^2\mathrm{E}\alpha_s^2}}_{S},$$

where, $1 - R^2_{\alpha\sim\alpha_s} = \eta^2_{D\sim U|X}$, *and* $E\alpha_s^2 = \frac{1}{\mathrm{E}(D-\mathrm{E}[D|X])^2} = \frac{1}{\mathrm{E}[\mathrm{Var}(D|X)]}$. *Moreover, if both* $\mathrm{E}[Y \mid D,X,U]$ *and* $\mathrm{E}[Y \mid D,X]$ *are partially linear in D—that is, if g and g_s are correctly specified—we also have* $\mathrm{E}(Y-g_s)^2 = \mathrm{E}[\mathrm{Var}(Y \mid D,X)]$ *and* $R^2_{Y-g_s\sim g-g_s} = \eta^2_{Y\sim U|DX}$.

This result for partially linear models is new and it naturally generalizes traditional OVB formulas for linear models (see Section C.2 in the Appendix). Note the theorem first states the bias formula in terms of the Riesz representers α and α_s, and then specializes its interpretations. This is intentional, and it aims to facilitate the transition to the general case of nonparametric models in the next section.

2.3. **Making sense of the OVB formula.** Theorem 1 shows that the bias decomposes into the product of an identifiable scaling factor, S, and a non-identifiable bias factor, $\rho C_Y C_D$. The scaling factor $S = \sqrt{\frac{\mathrm{E}(Y-g_s)^2}{\mathrm{E}(D-\mathrm{E}[D|X])^2}}$ is given by the ratio of the observed residual variances of the treatment and the outcome. Intuitively, it shows that the more variation the observed covariates explain of the outcome, the less room there is for omitted variable bias to arise. On the other hand, the more variation the observed covariates explain of the treatment, the less residual variation of the treatment is being used to identify the target parameter—this, in its turn, *inflates* the potential biasing effect of omitted variables.

Moving to the bias factor $\rho C_Y C_D$, this term decomposes the bias in terms of the capacity of confounders to induce errors in the outcome and in the treatment equation, as well as on how systematically related these errors are. The sensitivity parameter $R^2_{Y-g_s \sim g-g_s}$ in C_Y captures the potential of the confounder to induce errors in the outcome regression, as measured by how much residual variation the confounder explains of the outcome. The sensitivity parameter $1-R^2_{\alpha \sim \alpha_s} = \eta^2_{D \sim U|X}$ in C_D captures the potential of the confounder to induce errors in the treatment regression, as measured by the proportion of residual variation of the treatment explained by U. Collectively, the pair $(1-R^2_{\alpha\sim\alpha_s}, R^2_{Y-g_s\sim g-g_s})$ measures the *maximum explanatory power* of the confounders, which sets an upper bound on how much bias U *could* produce. However, for U to actually create bias, it is not sufficient for it to create errors in the treatment and the outcome equations—it must do so *systematically*. This is captured by the correlation of errors ρ.[3]

2.4. **Using the OVB formula to bound the bias.** Given plausibility judgments on the maximum value that the components of the OVB formula can take, $|\rho| \leq \rho^{\max}$, $R^2_{Y-g_s\sim g-g_s} \leq R^{2\max}_Y$ and

[3] For an example, consider the model $D = U^2$, $Y = \theta D + U$, with $U \sim N(0,1)$. Here $\eta^2_{D\sim U|X} = \eta^2_{Y\sim U|D,X} = 1$, that is, the latent variable U nonparametrically explains 100% of the residual variation of the treatment and of the outcome. However, since U^2 is uncorrelated with U, we have $\rho = 0$, and omitting this confounder induces no bias in the estimation of θ.

$1 - R^2_{\alpha\sim\alpha_s} \leq R^{2\,\max}_D$, we immediately obtain the following bound on the bias,

$$|\theta - \theta_s| \leq \rho^{\max} \times \sqrt{R^{2\,\max}_Y \times \left(\frac{R^{2\,\max}_D}{1 - R^{2\,\max}_D}\right)} \times S. \tag{6}$$

Additionally, if investigators are unwilling to make any assumptions about $\rho^{\max}$ and $R^{2\,\max}_Y$ (i.e., setting them to the trivial bound of 1), then (6) reduces to a simple, single-parameter bound[4]

$$|\theta - \theta_s| \leq \sqrt{\frac{R^{2\,\max}_D}{1 - R^{2\,\max}_D}} \times S.$$

Note how this simplifies the complexity of plausibility judgments for sensitivity analysis. Researchers need only reason about the maximum explanatory power that unobserved confounders have in explaining treatment variation in order to place bounds on the bias. Should the researcher later choose to postulate the maximum explanatory power of confounders with the outcome via $R^{2\,\max}_Y$, or the degree to which the confounder is adversarial, via $\rho^{\max}$, the bounds become tighter.

Finally, one might wonder whether the bound (6) is the tightest possible bound one could obtain given the model assumptions and the distribution of observed variables. Sometimes the answer is yes. For example, in Appendix C.3, we demonstrate that when the observed residuals are homoskedastic Gaussian, it is always possible to construct a latent confounder U, compatible with the observed distribution, that achieves any combination of values $(|\rho|,\ R^2_{Y-g_s\sim g-g_s},\ 1 - R^2_{\alpha\sim\alpha_s}) \in [0,1)^3$. This implies that, in such a scenario, the bound (6) is always sharp and cannot be improved without additional assumptions. Other times, the observed distribution may impose restrictions on the values the triple can take, and the bound (6) may be tightened by leveraging such restrictions. See Appendix C.12 for further discussion.

3. Main Results: Omitted Variable Bias in Nonparametric Causal Models

We now derive the main results of the paper, and characterize the size of the omitted variable bias for a broad class of causal parameters that can be identified as linear functionals of the conditional

[4] This is similar in spirit to Manski (1990), in which, absent further assumptions, we consider the worst case scenario.

expectation function of the outcome. 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.** As a motivating example, consider any causal or structural equation model that implies the existence of potential outcomes, $Y(d)$, under the intervention that sets the treatment *experimentally* to d. Assume that *consistency* holds, namely, observed outcomes equal to the potential outcome for the treatment actually received, $Y := Y(D)$. Further, let the treatment assignment D obey the ignorability (conditional exogeneity) condition:

$$Y(d) \perp\!\!\!\perp D \mid \{X, U\}, \tag{7}$$

which states that the realized treatment D is independent of the potential outcomes, conditionally on observed covariates X and *unobserved* confounders U. Under this set-up, and when d is in the support of D given X, U, we then have the following (well-known) identification result

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

that is, the conditional average potential outcome coincides with the "long" regression function of Y on D, X, and U. Therefore, we can identify various causal parameters—functionals of the average potential outcome—from the regression function. Important examples include: (i) the average treatment effect (ATE)

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

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

$$\theta = \mathrm{E}\left[\partial_d \mathrm{E}[Y(D) \mid X, U]\right] = \mathrm{E}[\partial_d g(D, X, U)],$$

for the case of a continuous treatment D.

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

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

$$\theta := \mathrm{E}m(W,g), \tag{8}$$

where g is the projection of the CEF $\mathrm{E}[Y \mid W]$ onto a closed linear subspace $\Gamma \subseteq L^2(P_W)$:

$$g := \arg\min_{\gamma\in\Gamma} \mathrm{E}(Y-\gamma(W))^2 \equiv \arg\min_{\gamma\in\Gamma} \mathrm{E}(\mathrm{E}[Y \mid W]-\gamma(W))^2 \tag{9}$$

and the mapping $\gamma \mapsto m(w;\gamma)$ is linear in $\gamma \in \Gamma$, and continuous in $\gamma \in \Gamma$ with respect to the $L^2(P_W)$ norm.

This formulation covers the two previous examples with scores $m(W,g) = g(1,X,U) - g(0,X,U)$ for the ATE and $m(W,g) = \partial_d g(D,X,U)$ for the ACD. Moreover, it covers settings where the true CEF does not lie in the subspace Γ, allowing for mis-specification. Hence, it enables omitted variable bias for coefficients in best (partially) linear models, without any assumption of (partial) linearity of the CEF, as in Section 2.

The continuity condition holds under the regularity conditions provided in the remark below.

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

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

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

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

These conditions imply that Assumption 1 holds, by Theorem 5 given in Appendix A. □

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

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

With the short regression in hand, we can compute proxies (or approximations) θ_s for θ. In particular, for the ATE and ACD the short parameter consists of

$$\theta_s = \mathrm{E}[g_s(1,X) - g_s(0,X)] \text{ and } \theta_s = \mathrm{E}[\partial_d g_s(D,X)].$$

In our general framework, the proxy parameter can also be expressed as the same linear functional applied to the short regression, $g_s(W_s)$ and the short regression is allowed to only be a mean-squared-projection of the CEF $\mathrm{E}[Y \mid D,X]$ onto a closed linear subspace Γ_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:*

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

where g_s is the projection of the CEF $\mathrm{E}[Y \mid W_s]$ onto a closed linear subspace $\Gamma_s \subseteq \Gamma \cap L^2(P_{W_s})$, i.e.,

$$g_s := \arg\min_{\gamma_s \in \Gamma_s} \mathrm{E}(Y - \gamma_s(W_s))^2 \equiv \arg\min_{\gamma_s \in \Gamma_s} \mathrm{E}(\mathrm{E}[Y \mid W_s] - \gamma_s(W_s))^2 \tag{10}$$

We require $m(W,g_s)$ to be a measurable function of W_s alone, i.e., the score depends only on W_s when evaluated at the short regression g_s. Given this assumption, we use the shorthand notation $m(W_s,g_s)$ in place of $m(W,g_s)$.

In the two working examples this assumption is satisfied, since $m(W_s,g_s) = g_s(1,X) - g_s(0,X)$ for the ATE and $m(W_s,g_s) = \partial_d g_s(D,X)$ for the ACD.

3.2. **Omitted variable bias for linear functionals.** We now characterize the omitted variable bias, i.e., the difference between the "long" and "short" functionals, $\theta - \theta_s$, for general linear functionals of (projections of) the CEF. A key step to our approach is the following lemma that characterizes the target parameters and their proxies as inner products of regression functions with terms called Riesz representers (RR).

Lemma 1 (**Riesz Representation**). *Consider the long and short parameters θ and θ_s as given by Assumptions 1 and 2. There exist unique functions $\alpha \in \Gamma$ and $\alpha_s \in \Gamma_s$, referred to as the long and short Riesz Representers (RRs), such that*

$$\mathrm{E}m(W,\gamma) = \mathrm{E}\gamma(W)\alpha(W), \qquad \mathrm{E}m(W_s,\gamma_s) = \mathrm{E}\gamma_s(W_s)\alpha_s(W_s),$$

for all $\gamma \in \Gamma$ and $\gamma_s \in \Gamma_s$. Furthermore, α_s is the projection of α onto Γ_s, namely,

$$\alpha_s \in \arg\min_{\gamma_s \in \Gamma_s} \mathrm{E}(\alpha(W) - \gamma_s(W_s))^2$$

and therefore we also have that:

$$\mathrm{E}(\alpha(W) - \alpha_s(W_s))^2 = \mathrm{E}\alpha^2 - \mathrm{E}\alpha_s^2.$$

An important aspect of Lemma 1 is the property that the short RR α_s is the orthogonal projection of the long RR α onto Γ_s. This property is crucial for the proof of our main omitted variable bias formula later in this section. Note that when Γ, Γ_s are un-restricted spaces, i.e., $\Gamma = L^2(P_W)$ and $\Gamma_s = L^2(P_{W_s})$, then the latter property between the long and short RRs translates to:

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

For restricted spaces, this is a more subtle and relatively surprising statement.

Intuitively, the Riesz representers can be seen as weights that map the outcome to the target parameter. In the case of the ATE with a binary treatment, and imposing no restrictions on the regression function, then the representers are just the classical inverse probability of treatment (Horvitz-Thompson) weights:

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

This follows from change of measure arguments. One can also verify that $\alpha_s = \mathrm{E}[\alpha \mid D,X]$, by applying Bayes' rule.

In the case of the ACD with a continuous treatment, using integration by parts we can verify that the representers are logarithmic derivatives of the conditional densities:

$$\alpha(W) = -\partial_d \log f(D \mid X,U), \quad \alpha_s(W_s) = -\partial_d \log f(D \mid X).$$

If we further restrict the regression functions to be partially linear as in Section 2

$$g(W) = \beta D + f(X,U), \quad g_s(W_s) = \beta_s D + f_s(X),$$

then the restricted Riesz representers of either the ATE or the ACD functionals take the form:

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

That is, the representer is given by the (scaled) residualized treatment, which we previously derived using the classical Frisch-Waugh-Lovell theorem, without invoking Riesz representation per se. We remark here that Lemma 1 proves that the short RR in this case is the orthogonal projection of the long RR onto the space of functions of D,X that are partially linear in D. This is a highly non-trivial statement that comes out of our general framework. For illustrative purposes, in Appendix C.5, we provide a proof of this fact from first principles for this special case.

Using these lemmas, we obtain the following characterization of the OVB for general linear functionals of (projections of) the CEF.

Theorem 2 (OVB—General). *Consider the long and short parameters θ and θ_s as given by Assumptions 1 and 2. Then the OVB is given by the covariance between the regression error and the*

RR error,

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

with α, α_s as defined in Lemma 1. This bias can be further reparameterized in terms of R^2 measures,

$$\theta - \theta_s = \underbrace{\mathrm{Cor}(g - g_s, \alpha - \alpha_s)}_{\rho} \times \underbrace{\sqrt{R^2_{Y - g_s \sim g - g_s}}}_{C_Y} \times \underbrace{\sqrt{\frac{(1 - R^2_{\alpha \sim \alpha_s})}{R^2_{\alpha \sim \alpha_s}}}}_{C_D} \times \underbrace{\sqrt{\mathrm{E}(Y - g_s)^2 \mathrm{E}\alpha_s^2}}_{S}.$$

This is the main conceptual result of the paper, and it is new. It covers a rich variety of causal estimands of interest, as long as they can be written as linear functionals of the long regression. Theorem 5 in Appendix A verifies the OVB validity for many interesting examples beyond the running examples of the ATE and ACD in the main text. These include: (i) weighted averages of potential outcomes (Example 1); (ii) weighted average treatment effects, such as the average treatment effect on the treated (ATT), or the average value of a policy (Example 2); (iii) average policy effects when transporting the short covariates W_s (Example 3); (iv) weighted average incremental effects (Example 4); and, (v) policy effects from changing the distribution of W_s (Example 5).

Remark 2 (Restricted vs Unrestricted Riesz Representers)**.** When the linear functional is assumed to be continuous for all $\gamma \in L^2(P_W)$ and not just for $\gamma \in \Gamma$, then the functional also admits short and long RRs $\bar{\alpha} \in L^2(P_W)$ and $\bar{\alpha}_s \in L^2(P_{W_s})$, that represent the functional for all square-integrable functions. The RR $\bar{\alpha}$ and $\bar{\alpha}_s$ do not necessarily coincide with their restricted counterparts α, α_s. Moreover, it is possible that α and α_s exist but $\bar{\alpha}$ and $\bar{\alpha}_s$ do not exist if it happens that the linear functional is continuous only on the subspace Γ. Whenever the un-restricted RRs exist then we show here that they satisfy the following relationship with the restricted RRs.

Lemma 2. *Suppose that the linear functional $f \to \mathrm{E}m(W, f)$ is continuous for all $f \in L^2(P_W)$. Then there exists unique $\bar{\alpha} \in L^2(P_W)$ and $\bar{\alpha}_s \in L^2(P_{W_s})$, such that for all $f \in L^2(P_W)$ and all $f_s \in L^2(P_{W_s})$:*

$$\mathrm{E}m(W, f) = \mathrm{E}f(W)\bar{\alpha}(W), \qquad \mathrm{E}m(W_s, f_s) = \mathrm{E}f_s(W_s)\bar{\alpha}_s(W_s),$$

Moreover, the restricted RRs $\alpha \in \Gamma$ *and* $\alpha_s \in \Gamma_s$ *are given by the orthogonal projections of* $\bar{\alpha}$ *and* $\bar{\alpha}_s$ *on* Γ *and* Γ_s*, respectively. Since projections reduce the norm,* $\mathrm{E}\alpha^2 \leq \mathrm{E}\bar{\alpha}^2$ *and* $\mathrm{E}\alpha_s^2 \leq \mathrm{E}\bar{\alpha}_s^2$. □

3.3. **Making sense of the OVB formula.** Theorem 2 generalizes the warm-up result to fully nonlinear models, and general target parameters defined as linear functionals of the long regression. Though more abstract, the same discussions of Sections 2.3 and 2.4 apply. As before, the bias decomposes into an identified scaling factor S, and an unidentified bias factor $\rho C_Y C_D$. The terms C_Y and C_D capture the components of the bias factor that measure the capacity of the omitted variables in generate errors in the outcome regression and in the Riesz representer. In particular, $R^2_{Y-g_s \sim g-g_s}$ in the first factor measures the proportion of residual variance in the outcome explained by confounders; whereas $1-R^2_{\alpha\sim\alpha_s}$ in C_D measures the proportion of residual variation of the long RR generated by latent confounders. These two components capture the maximum potential of U to create bias, which is modulated by how systematically these errors are correlated, as measured by ρ.

When we take $g = \mathrm{E}[Y \mid D, X, U]$ and $g_s = \mathrm{E}[Y \mid D, X]$ we have the same useful interpretation of $R^2_{Y-g_s\sim g-g_s} = \eta^2_{Y\sim U|D,X}$. The interpretation of $1-R^2_{\alpha\sim\alpha_s}$ and $C_D^2 = (1-R^2_{\alpha\sim\alpha_s})/R^2_{\alpha\sim\alpha_s}$ can be further specialized for different estimands, as follows.

Remark 3 (**Interpretation of OVB for the ATE with a Binary Treatment**)**.** For the ATE,

$$1-R^2_{\alpha_{\mathrm{ATE}}\sim\alpha_{\mathrm{ATE},s}} = 1-\frac{\mathrm{E}[1/\operatorname{Var}(D \mid X)]}{\mathrm{E}[1/\operatorname{Var}(D \mid X,U)]}, \qquad C_D^{2,\mathrm{ATE}} = \frac{\mathrm{E}[1/\operatorname{Var}(D \mid X,U)]}{\mathrm{E}[1/\operatorname{Var}(D \mid X)]}-1. \tag{11}$$

That is, the OVB of the ATE depends on the relative gain in the average precision of the treatment model due to U.[5] Thus, the interpretation of $1-R^2_{\alpha\sim\alpha_s}$ for the ATE with a binary treatment parallels that of the partially linear model—compare it to equation (1). The sole distinction being that, here, gains in predictive power are measured by the relative increase in precision rather than the relative decrease in variance. □

[5] Precision is the inverse of the variance.

Remark 4 (**Interpretation of OVB for the ATT with a Binary Treatment**)**.** For the ATT,

$$1 - R^2_{\alpha_{\text{ATT}} \sim \alpha_{\text{ATT},s}} = 1 - \frac{\mathrm{E}[\mathrm{Odds}(D \mid X)]}{\mathrm{E}[\mathrm{Odds}(D \mid X, U)]}, \qquad C_D^{2,\text{ATT}} = \frac{\mathrm{E}[\mathrm{Odds}(D \mid X, U)]}{\mathrm{E}[\mathrm{Odds}(D \mid X)]} - 1. \tag{12}$$

That is, the OVB of the ATT depends on the relative gain in the *average odds* of treatment due to U. In particular, note the similarities with the traditional sensitivity parameter of Rosenbaum's sensitivity model (Rosenbaum, 1987; Yadlowsky et al., 2022) or marginal sensitivity models (Tan, 2006; Dorn and Guo, 2023). We elaborate on this point further below in Section 3.4. □

Remark 5 (**Interpretation of OVB for Average Causal Derivatives**)**.** For the ACD example,

$$1 - R^2_{\alpha_{\text{ACD}} \sim \alpha_{\text{ACD},s}} = 1 - \frac{\mathrm{E}[(\partial_d \log f(D \mid X))^2]}{\mathrm{E}[(\partial_d \log f(D \mid X, U))^2]}, \qquad C_D^{2,\text{ ACD}} = \frac{\mathrm{E}[(\partial_d \log f(D \mid X, U))^2]}{\mathrm{E}[(\partial_d \log f(D \mid X))^2]} - 1, \tag{13}$$

which can be interpreted as the relative gain in information that the confounder U provides about the location of D. Furthermore, if D is homoscedastic Gaussian, conditional on both X and (X, U), we then have

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

so that $1 - R^2_{\alpha \sim \alpha_s}$ simplifies to the nonparametric R^2 of the latent variable with the treatment, similarly to the partially linear model, i.e, $1 - R^2_{\alpha \sim \alpha_s} = \eta^2_{D \sim U \mid X}$. □

3.4. **Relationship with marginal sensitivity models.** Given their popularity and importance, here we expand on the differences between our omitted variable bias approach for sensitivity analysis, and sensitivity models based on worst-case assumptions on the change in odds-ratios, such as the "marginal sensitivity model" (MSM) (Tan, 2006; Dorn and Guo, 2023; Dorn et al., 2025). As these approaches usually restrict D to be binary, we focus on this case, with the understanding that this is not necessary for our approach. We note that similar discussion applies to Rosenbaum's sensitivity model (Rosenbaum, 1987; Yadlowsky et al., 2022).

3.4.1. *Omitted variable bias, sensitivity analysis and worst-case assumptions.* Traditional approaches to sensitivity analysis, such as the marginal sensitivity model (MSM), often start with some assumption about the strength of U, and then determine how this assumption constrains the target parameter of interest, such as the average potential outcome, the ATE or the ATT. For example, the MSM posits that unmeasured confounders U can change the odds of treatment by *at most* $\Lambda \geq 1$, for any value of the observed covariates X, i.e.,

$$\Lambda^{-1} \leq \frac{\text{Odds}(D \mid X = x, U = u)}{\text{Odds}(D \mid X = x)} \leq \Lambda, \qquad \forall x \in \mathscr{X}, u \in \mathscr{U}.$$

Given this constraint, one derives bounds on various causal effects. See, for example, Tan (2006), and more recently Dorn and Guo (2023) and Dorn et al. (2025) for sharp bounds under the MSM.

The omitted variable bias approach we take here inverts this logic entirely. Instead of starting from a primitive assumption about confounder strength, and then deriving its logical implications, we ask a targeted question: what exactly must be known about the confounder to characterize the bias of a given estimand? By dissecting the anatomy of the bias, the OVB formula reveals how confounders must be restricted to bound that particular target estimand—and it also shows that different estimands have different sensitivities to unmeasured confounding.

Concretely, for example, if one wants to bound the bias of the ATE, the OVB formula shows that it suffices to restrict the gains in *average* precision of treatment assignment (Remark 3), whereas if one wants to bound the ATT, it suffices to restrict the ratio of *average* odds of treatment (Remark 4). If, instead, interest lies in a partially linear projection coefficient, it suffices to restrict the gains in variation explained of the treatment (Section 2). Note these sensitivity parameters are not arbitrary; rather, they emerge naturally from the bias structure. Moreover, any restrictions beyond those are stronger than what is needed to bound each of these estimands.

The OVB analysis thus reveals the marginal sensitivity model imposes stronger assumptions than what is needed to bound the bias of any of these estimands. The case of the ATT is particularly instructive. In that case, the sensitivity parameters of the MSM and the OVB formula are very similar in nature. The key distinction is simply that the MSM asks researchers to make a strong worst-case assumption (i.e., bounding the worst-case change in odds) while the OVB formula shows that it suffices to make a weaker, average-case assumption (i.e., bounding the change in average odds). Understanding this distinction is important because Λ can be unbounded even when $1-R^2_{\alpha\sim\alpha_s}$ is not only finite, but small (see Appendix C.9 for an example).

3.4.2. *Using Λ to calibrate judgments on OVB parameters.* A researcher accustomed to the marginal sensitivity model might find this estimand-specific approach unfamiliar—after all, the appeal of the MSM lies partly in having a single parameter Λ that can be used to bound various causal quantities. Theorem 3 bridges this gap, and shows how one can still leverage a single belief about the worst-case odds ratio Λ to bound the maximum strength of $1-R^2_{\alpha\sim\alpha_s}$ for various estimands.[6] We state the theorem in terms of C_D^2 with the understanding that $1-R^2_{\alpha\sim\alpha_s} = C_D^2/(1+C_D^2)$.

Theorem 3 (Sharp Bounds on C_D given Λ)**.** *Under the marginal sensitivity model we have,*

$$
\begin{aligned}
C_D^{2,ATT} &= \frac{\mathrm{E}[\mathrm{Odds}(D\mid X,U)]}{\mathrm{E}[\mathrm{Odds}(D\mid X)]}-1 \;\le \frac{(\Lambda-1)^2}{\Lambda}\times\left(1-\frac{P(D=1)}{\mathrm{E}[\mathrm{Odds}(D\mid X)]}\right) \;\le \frac{(\Lambda-1)^2}{\Lambda},\\
C_D^{2,ATE} &= \frac{\mathrm{E}[1/\mathrm{Var}(D\mid X,U)]}{\mathrm{E}[1/\mathrm{Var}(D\mid X)]}-1 \le \frac{(\Lambda-1)^2}{\Lambda}\times\left(1-\frac{3}{\mathrm{E}[1/\mathrm{Var}(D\mid X)]}\right) \;\le \frac{(\Lambda-1)^2}{\Lambda},\\
C_D^{2,PLM} &= \frac{\mathrm{E}[\mathrm{Var}(D\mid X)]}{\mathrm{E}[\mathrm{Var}(D\mid X,U)]}-1 \;\le \frac{(\Lambda-1)^2}{\Lambda}\times\left(\frac{\mathrm{E}\left[\omega(X)\times\mathrm{Var}(D\mid X)\right]}{\mathrm{E}[\omega(X)]}\right)\le \frac{(\Lambda-1)^2}{4\Lambda},
\end{aligned}
$$

where $\omega(X) := \mathrm{Var}(D\mid X)/(\Lambda+(\Lambda-1)^2\,\mathrm{Var}(D\mid X))$*. The first bound is sharp given Λ and the observed data. The second bound is sharp given Λ alone.*

How could one use these relationships? If a researcher truly believes the worst-case odds ratio to be at most Λ, then they should use the recent sharp bounds of Dorn and Guo (2023) and Dorn et al. (2025). If, however, researchers find the worst-case assumption too strong, they can use Λ as a

[6] The proof of Theorem 3 leverages the fact that the long propensity score is bounded under the MSM. Thus, these results can potentially be extended to other sensitivity models that imply bounded propensity scores, such as Rosenbaum's model (Rosenbaum, 1987; Yadlowsky et al., 2022) or the model of Masten and Poirier (2018).

useful *heuristic* to calibrate plausible values for the OVB parameter $1-R^2_{\alpha\sim\alpha_s}$ from first principles, across various estimands, while effectively relaxing the strict worst-case assumption of the MSM.

3.5. **Using observed covariates to calibrate the strength of U.** Beyond making direct plausibility judgments on the strength of confounding using the above quantities, analysts can also leverage judgments of relative importance of variables to bound the size of the bias (see, e.g. Imbens, 2003; Cinelli and Hazlett, 2020). For instance, if one has reasons to believe that U would not generate as much gains in explanatory power as certain key observed covariates X_j, this can be used to formally place bounds on the strength of confounding due to U. This allows one to assess, for instance, whether confounders as strong or stronger than observed covariates would be sufficient to overturn an empirical result. We elaborate the benchmarking procedure formally in Section E of the appendix and illustrate its use in the empirical example. These results extend previous benchmarking ideas for linear regression models to the general case.

4. Statistical Inference Under Omitted Variable Bias

Suppose the components ρ, C_Y, C_D of the OVB formula in Theorem 2 are set through hypotheses on the maximum explanatory power of omitted variables. We then have the following bounds $[\theta_-,\theta_+]$ for the target parameter θ:

$$\theta_\pm = \theta_s \pm |\rho| C_Y C_D S, \quad S^2 = \mathrm{E}(Y-g_s)^2 \mathrm{E}\alpha_s^2. \tag{14}$$

The estimable 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 (Levit, 1975; Hasminskii and Ibragimov, 1978; Pfanzagl and Wefelmeyer, 1985; Bickel et al., 1993; Newey, 1994; Chernozhukov et al., 2018a, 2022a) based on Neyman orthogonal scores we give for these components, 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. (2022c, 2021). This representation is Neyman orthogonal with respect to perturbations of (g_s, α_s), which is a key property required for DML.[7] 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. Specifically Neyman orthogonality refers to the property:

$$\partial_{g,\alpha}\mathrm{E}[m(W_s, g) + (Y - g)\alpha]\Big|_{\alpha_s, g_s} = 0; \ \partial_g \mathrm{E}(Y - g)^2\Big|_{g_s} = 0; \ \partial_\alpha \mathrm{E}[2m(W_s, \alpha) - \alpha^2]\Big|_{\alpha_s} = 0;$$

where ∂ is the Gateaux (pathwise derivative) operator over directions $h \in \Gamma_s$. These Neyman orthogonality properties follow from the definition of g_s as the minimizer over a closed linear space Γ_s, the representation property of α_s over Γ_s and the fact that $g_s, \alpha_s, h \in \Gamma_s$, which imply the conditions:

$$\forall \gamma_s \in \Gamma_s : \mathrm{E}(Y - g_s(W_s))\gamma_s(W_s) = 0, \qquad \mathrm{E}(m(W_s, \gamma_s) - \alpha_s(W_s))\gamma_s(W_s) = 0.$$

Application of DML theory in Chernozhukov et al. (2018a) and the delta-method gives 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 learning rate faster than $n^{-1/4}$. 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 $\widehat{\eta}_\ell = (\widehat{g}_\ell, \widehat{\alpha}_\ell)$*

[7] For the partial linear model of Section 2, one can alternatively use the partialling-out approach of Robinson (1988).

from observations in I_ℓ^c. (3) Estimate β as a root of: $0 = n^{-1}\sum_{\ell=1}^L \sum_{i \in I_\ell} \psi(\beta, Z_i; \widehat{\eta}_\ell)$. 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

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

for the scores

$$\psi_{\theta_s}(Z; \theta_s, g_s, \alpha_s) := m(W_s, g_s) + (Y - g_s(W_s))\alpha_s(W_s) - \theta_s;$$

$$\psi_{\sigma_s^2}(Z; \sigma_s^2, g_s) := (Y - g_s(W_s))^2 - \sigma_s^2;$$

$$\psi_{\nu_s^2}(Z; \nu_s^2, \alpha_s) := (2m(W_s, \alpha_s) - \alpha_s^2) - \nu_s^2.$$

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

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi^o(Z_i) + o_{\mathrm{P}}(1) \rightsquigarrow N(0, \mathrm{E}\psi^{o2}(Z)).$$

The application of the results in Chernozhukov et al. (2018a) 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 $\hat{\eta}_\ell = (\hat{\alpha}_\ell, \hat{g}_\ell)$ of $\eta_0 = (g_s, \alpha_s)$ in $L^2(P_{W_s})$ obey Assumptions 3.1 and 3.2 in Chernozhukov et al. (2018a), 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:*

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

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:

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

Confidence bounds for the bounds can be constructed using the following result.

Theorem 4 (DML Confidence Bounds for Bounds). *Under the conditions of Lemma 3, the plug-in estimator $\widehat{\theta}_{\pm}$ is also asymptotically linear and Gaussian with the influence function:*

$$\varphi^o_{\pm}(Z) = \psi^o_{\theta_s}(Z) \pm \frac{|\rho|}{2} \frac{C_Y C_D}{S} (\sigma_s^2 \psi^o_{\nu_s^2}(Z) + \nu_s^2 \psi^o_{\sigma_s^2}(Z)), \quad S > 0.$$

Therefore, the confidence bound

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

has the one-sided covering property, namely

$$\mathrm{P}(\theta_- \geq \ell) \to 1 - a \textit{ and } \mathrm{P}(\theta_+ \leq u) \to 1 - a, \textit{ as } n \to \infty.$$

The same results continue to hold if $\mathrm{E}\varphi^{o2}_{\pm}(Z)^2$ *is replaced by the empirical analogue* $\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i \in I_\ell} \hat{\varphi}^{o2}_{\pm}(Z_i)$.

We focus on the one-sided covering property stated in the theorem, since in applications the relevant hypotheses are typically one-sided. We can use further adjustments of Stoye (2009) to construct uniformly valid two-sided intervals (in broad terms, Stoye's approach smoothly switches to two-sided bands slightly before the bounds cannot be distinguished from a singleton by any statistical test).

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

Remark 6 (Machine Learning of α_s and g_s). Estimation of the short regression 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. (2018a); Semenova and Chernozhukov (2021), and references therein, for practical details). Second, we can use a variational characterization of α_s:

$$\alpha_s = \arg\min_{\alpha \in \mathscr{A}} \mathrm{E}[\alpha^2(W_s) - 2m(W_s, \alpha)],$$

where $\mathscr{A} \subseteq \Gamma_s$ is the parameter space for α_s, as proposed in Chernozhukov et al. (2021, 2022c). This avoids inverting propensity scores or conditional densities, as usually required when using analytical formulas. This approach is motivated by the first-order-conditions of the variational characterization:

$$\mathrm{E}\alpha_s g = \mathrm{E}m(W_s, g) \quad \text{for all } g \text{ in } \mathscr{G},$$

which is the definition of the RR. Neural network (RieszNet) and random forest (ForestRiesz) implementations of this approach are given in Chernozhukov et al. (2022b), and the Lasso implementation in Chernozhukov et al. (2022c).[8] □

5. Omitted Firm Characteristics in Evaluating the Effects of 401(k) Plan.

In this section we demonstrate the utility of our approach in an empirical example that estimates the average treatment effect of 401(k) eligibility on net financial assets (Poterba et al., 1994, 1995; Chernozhukov et al., 2018a). Our goal is to determine whether prior conclusions, reached under the assumption of conditional ignorability, are robust to plausible scenarios of unmeasured confounding. This example illustrates sensitivity analysis for the ATE in a partially linear model and in a nonparametric model with a binary treatment. In the Appendix we provide an additional example that estimates the price elasticity of gasoline demand (Blundell et al., 2012, 2017; Chetverikov and Wilhelm, 2017) and illustrates sensitivity analysis for the average causal derivative with a continuous treatment.

[8] A third option is to use a minimax (adversarial) characterization of α_s, as in Chernozhukov et al. (2018b, 2020): $\alpha_s = \arg\min_{\alpha \in \mathscr{A}} \max_{g \in \mathscr{G}} |\mathrm{E}m(W_s, g) - \mathrm{E}\alpha g|$, where $\mathscr{A}$ is the parameter space for α_s. The Dantzig selector implementation of this approach is given in Chernozhukov et al. (2018b). The neural network implementation of this approach is given in Chernozhukov et al. (2020).

5.1. **Estimates under conditional ignorability.** A 401(k) plan is an employed sponsored tax-deferred savings option that allows individuals to deduct contributions from their taxable income, and accrue tax-free interest on investments within the plan. Introduced in the early 1980s as an incentive to increase individual savings for retirement, an important question in the savings literature is precisely to quantify the *causal* impact of 401(k) eligibility on net financial assets. Indeed, a naive comparison of net financial assets between those individuals with and without 401(k) eligibility suggests a positive and large impact: using data from the 1991 *Survey of Income and Program Participation* (SIPP), this difference amounts to $19,559.

The problem of this naive comparison, however, is that 401(k) plans can be obtained only by those individuals that work for a firm that offers such savings option—and employment decisions are far from randomized. As an attempt to overcome this lack of random assignment, Poterba et al. (1994), Poterba et al. (1995), and more recently Chernozhukov et al. (2018a), leveraged the 1991 SIPP data to adjust for potential confounding factors between 401(k) eligibility and the financial assets of an individual. As explained in Poterba et al. (1994), at least around the time 401(k) plans initially became available, people were unlikely to make employment decisions based on whether an employer offered a 401(k) plan; instead, their main focus was on salary and other aspects of the job. Thus, as a first approximation, whether one is eligible for a 401(k) plan could be taken as ignorable once we condition on income and other covariates related to job choice.

It is useful to think about causal diagrams (Pearl, 2009) that represent this identification strategy. One possible model is shown Figure 1a. Here the outcome variable, Y, consists of net financial assets;[9] the treatment variable, D, is an indicator for being eligible to enroll in a 401(k) plan; finally, the vector of observed covariates, X, consists of: (i) age; (ii) income; (iii) family size; (iv) years of education; (iv) a binary variable indicating marital status; (v) a "two-earner" status indicator; (vi) an

[9] Defined as the sum of IRA balances, 401(k) balances, checking accounts, U.S. saving bonds, other interest-earning accounts in banks and other financial institutions, other interest-earning assets (such as bonds held personally), stocks, and mutual funds less non-mortgage debt.

IRA participation indicator; and, (vii) a home ownership indicator. We consider that the decision to work for a firm that offers a 401(k) plan depends both on the observed covariates X, but also on *unobserved* firm characteristics, denoted by U; moreover, X, U, and D are jointly affected by a set of latent factors L. Most importantly, note the assumption of *absence* of direct arrows, both from U and L, to Y. Under such assumption, conditional ignorability holds adjusting for X only. The story represented by the DAG of Figure 1a is one way of rationalizing the identification strategy used in earlier papers.

The first three columns of Table 1 shows the estimates for the average treatment effect (ATE) of 401(k) eligibility on net financial assets under this conditional ignorability assumption. For these estimates, we follow the same strategy used in Chernozhukov et al. (2018a), and we estimate the ATE using DML with Random Forests, considering both a partially linear model (PLM), and a nonparametric model (NPM).[10] As we can see, after flexibly taking into account observed confounding factors, although the estimates of the effect of 401(k) eligibility on net financial assets are substantially attenuated, they are still large, positive and statistically significant (approximately \$9,000 for the PLM and \$8,000 for the NPM). With the nonparametric model, we further explore heterogeneous treatment effects, by analyzing the ATE within income quartile groups. The results are shown in Figure 2a. We see that the ATE varies substantially across groups, with effects ranging from approximately \$4,000 (first quartile) to almost \$18,000 (last quartile).

5.2. **Sensitivity analysis.** It is now useful to consider scenarios in which conditional ignorability fails. Figure 1b presents one such scenario, where a violation of conditional ignorability is credible.[11] Employers often offer a benefit in which they "match" a proportion of an employee's contribution

[10] We use Random Forest both for the outcome and treatment regression and estimate the parameters using DML with 5-fold cross-fitting. In order to reduce the variance that stems from sample splitting, we repeat the procedure 5 times. Estimates are then combined using the median as the final estimate, incorporating variation across experiments into the standard error as described in Chernozhukov et al. (2018a).

[11] We note that Figure 1b is just one example, and our sensitivity analysis results hold for any model in which conditional ignorability holds given observed variables and unobserved confounders.

to their 401(k) up to 5% of the employee's salaries. The model in Figure 1b allows this "matched amount," denoted by M, to be determined by unobserved firm characteristics U, observed worker characteristics X, and by 401(k) eligibility D. In this model, adjustment for X alone is *not* sufficient for control of confounding. Instead, we now need to condition *both* on observed covariates X *and* unobserved confounders U for ignorability to hold. How strong would the omitted firm characteristics U have to be in order to overturn our previous conclusions? And how plausible are the strengths revealed to be problematic? In what follows, we use our sensitivity analysis results to address these questions.

5.2.1. *Minimal sensitivity reporting.* In reporting empirical results, the following definition will be useful. This definition extends the suggestion of Cinelli and Hazlett (2020, 2025) made for linear regression to the general case.

Definition 2 (Robustness Values). *The robustness value $RV_{\theta,a}$ stands for the minimum upper bound RV on sensitivity parameters, $R^2_{Y-g_s\sim g-g_s} \leq RV$ and $1 - R^2_{\alpha\sim\alpha_s} \leq RV$, $|\rho| \leq 1$, such that the confidence bound $[l,u]$ of Theorem 4 includes θ, at the significance level a.*

Whereas standard errors, t-values or p-values communicate how robust the short estimate is to *sampling errors*, the idea of robustness values is to quickly communicate how robust the short estimate is to *systematic errors* due to residual confounding. For example, $\text{RV}_{\theta=0,\ a=.05}$ measures the minimal strength on both confounding factors such that the estimated confidence bound for the ATE would include zero, at the 5% significance level.[12]

Table 1 illustrates our proposal for a minimal sensitivity reporting of causal effect estimates. Beyond the usual estimates under the assumption of conditional ignorability, it reports the robustness values of the short estimate. Starting with the PLM, the $\text{RV}_{\theta=0,a=0.05} = 5.4\%$ means that unobserved

[12] As put by Rosenbaum (2005), "a sensitivity analysis [...] asks what the unmeasured covariate would have to be like to alter the conclusions of the study." Measures such as the RV provide such a characterization, and have a long history in the sensitivity analysis literature, see, e.g. Rosenbaum (1987) and VanderWeele and Ding (2017).

confounders that explain less than 5.4% of the residual variation, *both* of the treatment, and of the outcome, are not sufficiently strong to bring the lower limit of the confidence bound to zero, at the 5% significance level. Moving to the nonparametric model, we obtain a similar, but somewhat lower value of $\mathrm{RV}_{\theta=0,a=0.05} = 4.5\%$. The RV thus provides a quick and meaningful reference point that summarizes the robustness of the short estimate against unobserved confounding—any postulated confounding scenario that does not meet this minimal criterion of strength cannot overturn the results of the original study.

5.2.2. *Main confounding scenario.* We now proceed to construct a particular confounding scenario, based on the contextual details of the problem. We start with the assumption that U explains as much variation in net financial assets as the total variation of the maximal matched amount of income (5%) over the period of three years (roughly the period over which the effect is measured).[13] In the worst case scenario, this would lead to an additional 3% of *total* variation explained, resulting in a *partial* R^2 of outcome with omitted firm characteristics U of $C_Y^2 = \eta^2_{Y\sim U|DX} = 4\%$.[14] This amounts to a relative increase of approximately 10% in the baseline R^2 of the outcome regression of 28%. Following similar reasoning, and more conservatively, we posit that omitted firm characteristics can explain an additional 2.5% of the variation in 401(k) eligibility, corresponding to a 22% relative increase in the baseline R^2 of the treatment regression of 11.4%. For the partially linear model, this results in $1 - R^2_{\alpha\sim\alpha_s} = \eta^2_{D\sim U|X} \approx 3\%$ (and also $C_D^2 \approx 3\%$).[15] We adopt the same scenario for the nonparametric model, with the understanding that now this would correspond to gains in precision rather than gains in variation explained (see Remark 3). Since both $\eta^2_{Y\sim U|DX} \approx 4\%$ and $1 - R^2_{\alpha\sim\alpha_s} \approx 3\%$ are below the robustness value of 5.4% (or 4.5%), we immediately conclude that such confounding scenario is *not* capable of bringing the lower limit of the confidence bound of the ATE to zero.

[13] This strategy is based on a suggestion by James Poterba.

[14] $\eta^2_{Y\sim U|DX} = \frac{\eta^2_{Y\sim UDX} - \eta^2_{Y\sim DX}}{1-\eta^2_{Y\sim DX}} = \frac{0.28+0.03-0.28}{1-0.28} \approx 4\%$,

[15] $1 - R^2_{\alpha\sim\alpha_s} = \eta^2_{D\sim U|X} = \frac{\eta^2_{D\sim UX} - \eta^2_{D\sim X}}{1-\eta^2_{D\sim X}} = \frac{0.114+.025-0.114}{1-0.114} \approx 3\%$.

The exact bias, bounds, and confidence bounds on the ATE implied by the posited scenario are shown in Table 2.[16] Starting with the partially linear model, the confounding scenario has an estimated absolute value of the bias of $\$4,196$. Accounting for statistical uncertainty, we obtain a lower limit for the confidence bound of $\$2,497$. The results for the nonparametric model are qualitatively similar, with a bias of similar magnitude, and point estimates, bounds, and confidence bounds for the ATE shifted down by roughly one thousand dollars. Confidence bounds for group-wise ATEs can also be computed, and are shown in Figure 2b. Note how the bounds are still largely positive, with only a small excursion into the negative side in the case of the second quartile group. These results suggest that the main qualitative findings reported in earlier studies are relatively robust to plausible violations of unconfoundedness, such as the one specified by our confounding scenario.

5.2.3. *Sensitivity contour plots and benchmarks.* 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 showing the collection of curves in the space of R^2 values along which the confidence bounds are constant (Imbens, 2003; Cinelli and Hazlett, 2020). These plots allow investigators to quickly and easily assess the robustness of their findings against *any* postulated confounding scenario. Here we focus on contour plots for the lower limit of the confidence bounds, as this is the direction of the bias that threatens the preferred hypothesis in this empirical example. Analogous contours can be constructed for the upper limit of the confidence bounds, and are omitted.

Starting with the partially linear model, the results are shown in Figure 3a. The horizontal axis describes the fraction of residual variation of the treatment explained by unobserved confounders, whereas the vertical axis describes the share of residual variation of the outcome explained by unobserved confounders. The contour lines show the lower limit of the confidence bounds $[l,u]$ for the ATE (see Theorem 4), given a pair of hypothesized values of partial R^2. Note $\text{RV}_{\theta=0,a=0.05}$ of Table 1 is simply the point where the 45-degree line crosses the contour line of zero (red dashed

[16] We use the same estimation procedure as described in footnote 10.

line), offering a convenient summary of the critical contour. We can further place reference points on the contour plots, indicating plausible bounds on the strength of confounding, under alternative assumptions about the maximum explanatory power of omitted variables. The red triangle point on the plot—*Max Match*—shows the bounds on the partial R^2 as previously discussed, resulting in a lower limit of the confidence bound for the ATE of \$2,497, in accordance with Table 2. Note here the correlation $|\rho|$ is set to its upper bound of 1.

Another approach to construct confounding scenarios is to use observed covariates to bound the plausible strength of unobserved covariates. For instance, in our empirical example, we know that employment decisions are largely driven by salary considerations. Similarly, salary is clearly an important determinant of net financial assets. One could therefore argue that it is implausible to imagine other latent firm characteristics that would be even a fraction as strong as the observed *income* of individuals, in terms of explanatory power in predicting 401(k) eligibility and net financial assets. Whenever such claims of relative importance can be made, they can be used to set plausible bounds on the strength of unmeasured confounding. Formal details of this benchmarking procedure are provided in Appendix E.[17]

The red diamonds of Figure 3a shows the bounds on the strength of the latent variable U if it were as strong as (i) income (*1 x Income*), (ii) whether a worker has an individual retirement account (*1 x Part. in IRA*), and (iii) whether the worker's family has a two-earner status (*1 x Two Earners*). Note that, apart from income, latent variables as strong as these covariates would result in a weaker confounding scenario than the one we have previously considered (*Max Match*). As for income, the worst-case bound indicates that omitted firm characteristics as important as income could be sufficiently strong to overturn the original results. However, one could argue such scenario to be implausible, as it is hard to imagine latent firm characteristics that would explain more variation in job choice than income itself. A more realistic, but still conservative, scenario is thus provided by

[17] Since the benchmarks are estimated from data, one could also propagate the uncertainty due to estimation errors of the benchmarks via the delta method. Details are provided in Appendix E.6.

the benchmark point *1/4 x Income*, which shows the bound on the strength of U if it were 25% as strong as income in predicting treatment and outcome variation. Note this scenario is comparable to the *Max Match* scenario, and not enough to bring the lower limit of the confidence bound to zero.

All results of Figure 3a were computed under the very conservative assumption that, given a pair of partial R^2 values for the latent variable U, the confounders enter both the outcome and treatment equations in a way that maximizes the bias, resulting in $|\rho| = 1$. This may be a conservative scenario, especially in nonlinear models. Thus, similar benchmarking procedures used for assessing the plausibility of the R^2 values can also be employed to calibrate judgments on the magnitude of ρ. Section E of the appendix shows that in fact none of the observed covariates result in $|\rho|$ values exceeding 1/2. With this in mind, Figure 3b presents the same contour plots as before, but now with $|\rho|$ set to a less conservative value of 1/2. Note how this substantially attenuates the bias, with the lower limits of the confidence bounds reaching approximately $\$4,600$ and $\$5,400$ for the *Max Match* and *1/4 x Income*, respectively.

Sensitivity contour plots for the nonparametric model are similar but slightly more conservative, and are provided in Figures 3c and 3d. The interpretation of the contours is the same as before, with the main difference being that the horizontal axis now describes gains in precision instead of gains in variance explained (see, e.g., Remark 3).

6. CONCLUSION

In this paper we provide a general theory of omitted variable bias for continuous linear functionals of the conditional expectation function of the outcome—all for general, nonparametric, causal models, while naturally allowing for (semi-)parametric restrictions (such as partial linearity), when such assumptions are made. We allow for arbitrary (e.g., binary or continuous) treatment and outcome variables, and we show that the bounds on the bias depend only on the maximum explanatory power of latent variables. We provide theoretical details of many leading causal estimands, and, in particular, we derive novel results for the important special cases of average treatment effects in partially linear models, in nonparametric models with a binary treatment, as well as for average causal derivatives. Finally, we leverage the Riesz representation of our bounds to offer flexible statistical inference through (debiased) machine learning, with rigorous coverage guarantees.

Many interesting extensions of our work are possible. For example, our results can potentially be extended to nonlinear functionals, such as those arising in instrumental variable (IV) methods. For instance, consider a variant of the IV problem (Imbens and Angrist, 1994), where the instrumental variable Z is valid only when conditioning both on observed covariates X, and latent variables U. In this case, the IV estimand is given by the ratio of two average treatment effects, and both the numerator and denominator can be bounded using the methods for the ATE proposed in this paper. Another interesting direction 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. (2022a) to derive approximate bounds on the bias that can be estimated using debiased ML approaches. Finally, our results can potentially be extended to the richer class of "mixed bias" functionals (Rotnitzky et al., 2021), which includes "doubly robust" functionals (Robins et al., 2008; Cui and Tchetgen Tchetgen, 2024) as a special case.

Acknowledgments, Data Availability, Conflict of Interests, and Funding

Acknowledgments. We thank the Editor and anonymous reviewers for very helpful comments. We are also grateful to Isaiah Andrews, Elias Bareinboim, Ben Deaner, Jacob Dorn, Xavier Gabaix, David Green, Guido Imbens, Steve Lehrer, Judith Lok, Esfandiar Maasoumi, Anna Mikusheva, Richard Nickl, Judea Pearl, Jack Porter, James Poterba, Alejandro Schuler, Eric Tchetgen Tchetgen, Ingrid Van Keilegom, and the participants of the Chamberlain seminar, Canadian Economic Association meeting, Institute for Nonparametric Statistics meeting, Bernoulli–IMS World Congress, EuroCIM, ACIC, PCIC, SBE, JSM, PolMeth, and UAI, as well as participants of seminars at

Harvard-MIT, Stanford, Wisconsin, Emory, Berkeley Methods Workshop, University of Washington, Cornell, Syracuse University, Penn State University, and BU Causal Seminar for their valuable feedback. We are grateful to Jack Porter for suggesting the “long story short” title. As usual, all remaining errors are our own.

Data availability. All data is publicly available in our GitHub repository: `https://github.com/carloscinelli/dml.sensemakr`. An R package for the methods developed in this paper is also available at the same repository. Alternative R and Python implementations are available via the DoubleML ecosystem at `https://docs.doubleml.org/stable/index.html`.

Conflict of interest. There are no relevant financial or nonfinancial competing interests to report.

Funding. This research was partially funded by the Royalty Research Fund at the University of Washington and by the National Science Foundation under Grants No. MMS-2417955 and 2242447.

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 M Aronow and Cyrus Samii. Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60(1):250–267, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program Evaluation with High-Dimensional Data. *ArXiv e-prints*, November 2013.

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.

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

Richard Blundell, Joel Horowitz, and Matthias Parey. Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *Review of Economics and Statistics*, 99 (2):291–304, 2017.

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.

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*, 2018a. ArXiv 2016; arXiv:1608.00060.

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*, 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*, 2021.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *Econometrica*, 2022a.

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. *International Conference on Machine Learning*, 2022b.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 2022c.

Denis Chetverikov and Daniel Wilhelm. Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, 85(4):1303–1320, 2017. doi: https://doi.org/10.3982/ECTA13639.

Denis Chetverikov, Dongwoo Kim, and Daniel Wilhelm. Nonparametric instrumental-variable estimation. *The Stata Journal*, 18(4):937–950, 2018.

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, 2020.

Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Biometrika*, page asaf004, 2025.

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

Yifan Cui and EJ Tchetgen Tchetgen. Selective machine learning of doubly robust functionals. *Biometrika*, 111(2):517–535, 2024.

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

Peng Ding and Tyler J VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368, 2016.

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.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 120(549):331–342, 2025.

H. P. Edmundson. Bounds on the expectation of a convex function of a random variable. RAND Paper P-982, The RAND Corporation, Santa Monica, CA, April 1957. April 9, 1957.

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

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.

Ragnar Frisch and Frederick V Waugh. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401, 1933.

Arthur Stanley Goldberger. *A course in econometrics*. Harvard University Press, 1991.

Rafail Z Hasminskii and Ildar A Ibragimov. On the nonparametric estimation of functionals. In *Proceedings of the 2nd Prague Symposium on Asymptotic Statistics*, pages 41–51, 1978.

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

Boris Ya Levit. On efficiency of a class of non-parametric estimates. *Teoriya Veroyatnostei i ee Primeneniya*, 20(4):738–754, 1975.

Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14(6): 570–580, 2013.

Michael C Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963.

Alexander R Luedtke, Ivan Diaz, and Mark J van der Laan. The statistics of sensitivity analyses. 2015.

Albert Madansky. Bounds on the expectation of a convex function of a multivariate random variable. *The Annals of Mathematical Statistics*, 30(3):743–746, 1959.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Matthew A Masten and Alexandre Poirier. Identification of treatment effects under conditional partial independence. *Econometrica*, 86(1):317–351, 2018.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

Oak Ridge National Laboratory ORNL. 2001 national household travel survey: User guide.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

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

J Pfanzagl and W Wefelmeyer. Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388, 1985.

James M. Poterba, Steven F. Venti, and David A. Wise. 401(k) plans and tax-deferred savings. In D. A. Wise, editor, *Studies in the Economics of Aging*. Chicago, IL: University of Chicago Press, 1994.

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

Amy Richardson, Michael G Hudgens, Peter B Gilbert, and Jason P Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical Science*, 29(4):596, 2014.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pages 335–422. Institute of Mathematical Statistics, 2008.

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

P. M. Robinson. Root-*N*-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

Paul R Rosenbaum. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science*, 4:1809–1814, 2005.

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.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

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.

Tymon Słoczyński. Interpreting ols estimands when treatment effects are heterogeneous: smaller groups get larger weights. *Review of Economics and Statistics*, 104(3):501–509, 2022.

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

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

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.

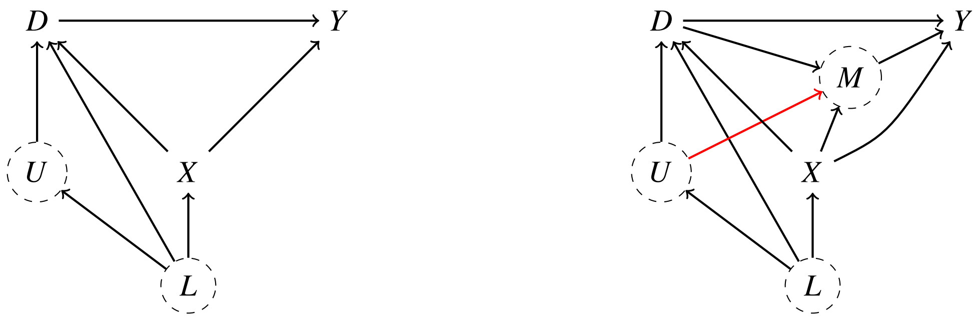
Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press, second edition, 2010.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5), 2022.

Chi Zhang, Carlos Cinelli, Bryant Chen, and Judea Pearl. Exploiting equality constraints in causal inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1630–1638. PMLR, 2021.

Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B*, 81(4):735–761, 2019.



(A) Ignorability holds conditional on X only.

(B) Ignorability holds conditional on X *and* U.

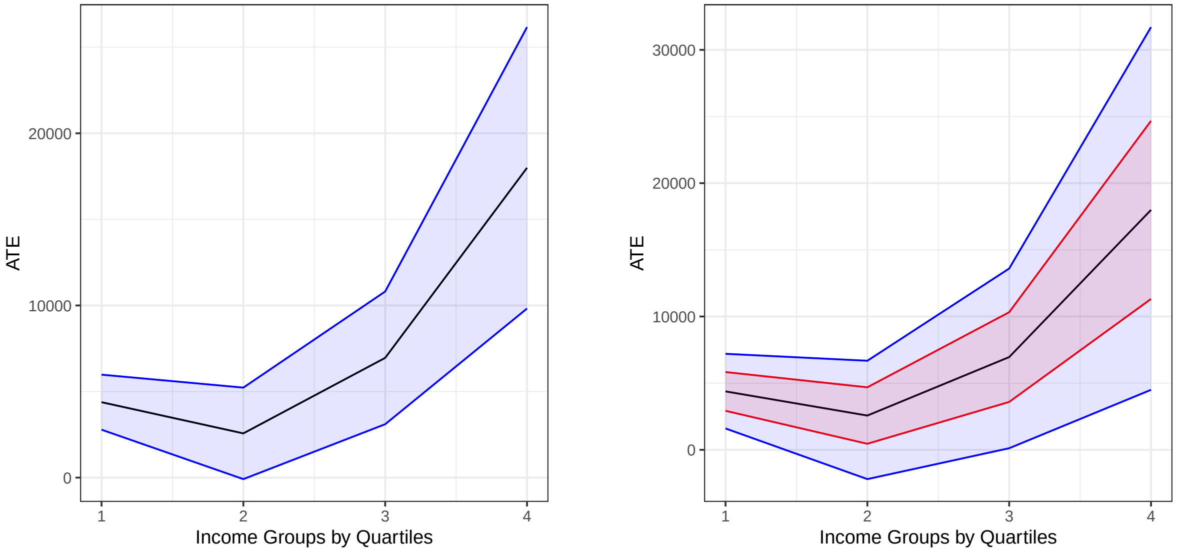
FIGURE 1. Two possible causal DAGs for the 401(k) example.

TABLE 1. Minimal sensitivity reporting. Significance level of 5%.

	Results Under Conditional Ignorability			Robustness Values
Model	Short Estimate	Std. Error	Confidence Bounds	$\mathrm{RV}_{\theta=0,\ a=0.05}$
Partially Linear	9,002	1,394	[6,271; 11,733]	5.4%
Nonparametric	7,949	1,245	[5,509; 10,388]	4.5%

TABLE 2. Estimate, bias, and bounds for the ATE. Significance level of 5%. Standard errors in parenthesis. Confounding scenario: $\rho^2 = 1$; $C_Y^2 \approx 0.04$; $C_D^2 \approx 0.03$.

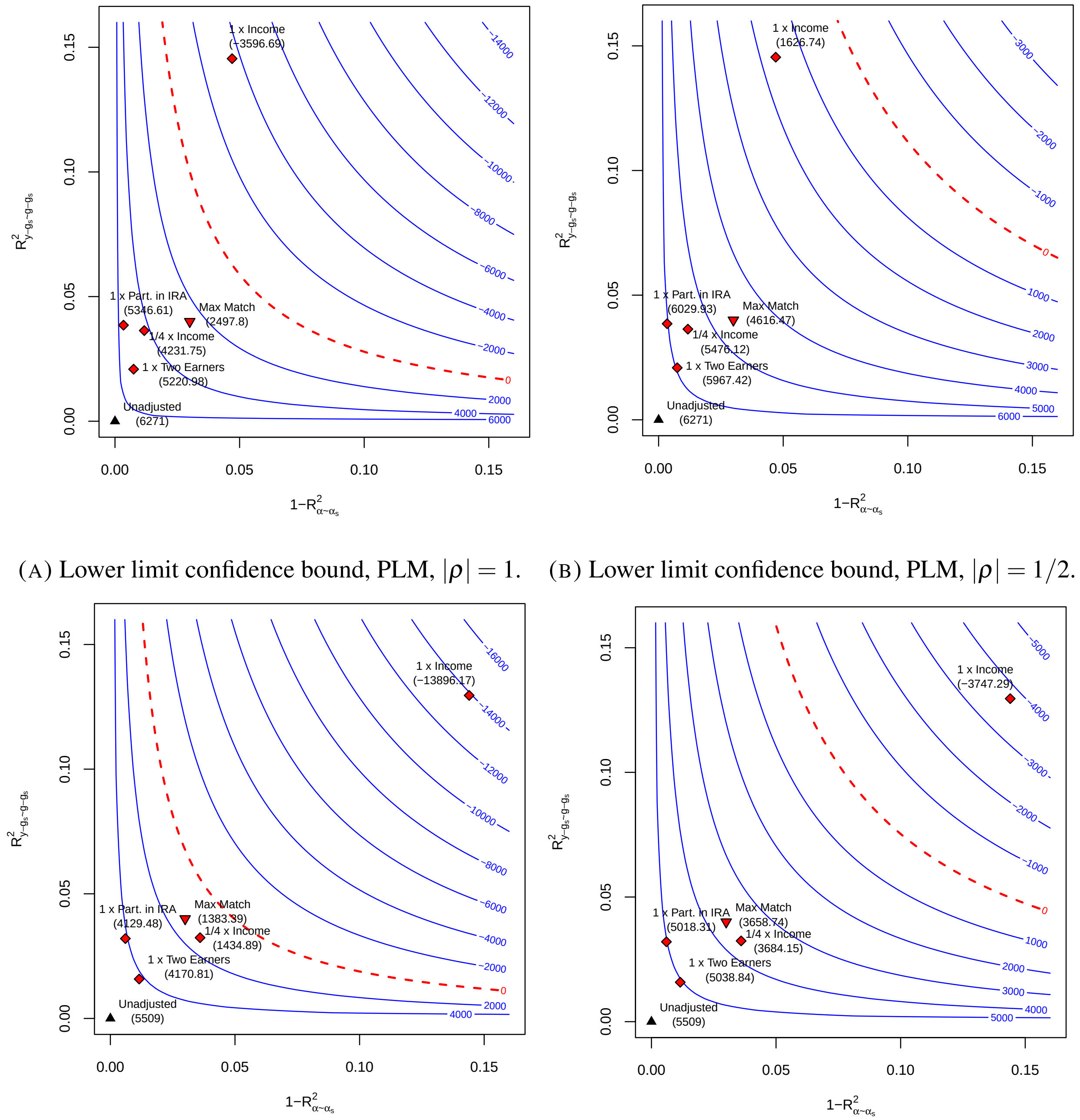
Model	Short Estimate	|Bias| Bound	ATE Bounds	Confidence Bounds
Partially Linear	9,002 (1,394)	4,196 (316)	[4,808; 13,196]	[2,497; 15,582]
Nonparametric	7,949 (1,245)	4,516 (336)	[3,452; 12,460]	[1,383; 14,630]



(A) Estimates under no confounding.

(B) Bounds under posited confounding.

FIGURE 2. Estimate (black), bounds (red), and confidence bounds (blue) for the ATE by income quartiles. Confounding scenario: $\rho^2 = 1$; $C_Y^2 \approx 0.04$; $C_D^2 \approx 0.03$. Significance level of 5%.



(A) Lower limit confidence bound, PLM, $|\rho| = 1$.

(B) Lower limit confidence bound, PLM, $|\rho| = 1/2$.

(C) Lower limit confidence bound, NPM, $|\rho| = 1$.

(D) Lower limit confidence bound, NPM, $|\rho| = 1/2$.

FIGURE 3. Sensitivity contour plots 401(k). Significance level $a = 0.05$.

Appendix A. Theoretical details for leading causal estimands

In this section we provide theoretical details for a wide variety of interesting and important causal estimands. Recall that we use $W = (D,X,U)$ to denote the "long" set of regressors and $W_s = (D,X)$ to denote the "short" list of regressors. For all estimands below, we consider the fully nonparametric regressions $g(D,X,U) := \mathrm{E}[Y \mid D,X,U]$ and $g_s(D,X) := \mathrm{E}[Y \mid D,X]$.

Let us start with examples for the binary treatment case, with the understanding that finitely discrete treatments can be analyzed similarly.

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 $\theta = \mathrm{E}[g(\bar{d},X,U)\ell(W_s)]$, where $W_s \mapsto \ell(W_s)$ is a bounded non-negative 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 $\mathrm{E}[\ell^2(W_s)/P(D = \bar{d} \mid X,U)] < \infty$.

The long parameter is a weighted average potential outcome (PO) when we set the treatment to $\bar{d}$, under the standard conditional ignorability assumption (7). 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 $\mathcal{N}$;
- $\ell(W_s) = D/P(D = 1)$ the average PO for the treated;
- $\ell(W_s) = (1-D)/P(D = 0)$ the average PO for the untreated.

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. (2018b), 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 $\theta = \mathrm{E}[(g(1,X,U) - g(0,X,U))\ell(W_s)]$, and the short parameter as $\theta_s =$

$\mathrm{E}[(g_s(1,X)-g_s(0,X))\ell(W_s)]$. We further assume $\mathrm{E}Y^2<\infty$ and the weak overlap condition $\mathrm{E}[\ell^2(W_s)/\{P(D=0\mid X,U)P(D=1\mid X,U)\}]<\infty$.

The long parameter is a weighted average treatment effect under the standard conditional ignorability assumption. In this example, setting

- $\ell(W_s)=1$ gives ATE in the entire population;
- $\ell(W_s)=1(X\in\mathcal{N})/P(X\in\mathcal{N})$ the ATE for group $\mathcal{N}$;
- $\ell(W_s)=D/P(D=1)$ the ATE for the treated (ATT);
- $\ell(W_s)=(1-D)/P(D=0)$ the ATE for the untreated (ATU);
- $\ell(x)=\pi(x)$ the average value of policy (APV) π,

where the policy π assigns a fraction $0\leq\pi(x)\leq 1$ 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 $W_s\mapsto T(W_s)$, where T is deterministic measurable map from $\mathcal{W}^s$ to $\mathcal{W}^s$. For example, the policy

$$(D,X,U)\mapsto(D+1,X,U)$$

adds a unit to the treatment D, that is $T(W_s)=(D+1,X)$. 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 $\mathrm{E}[\tilde{Y}\mid X,U]=g(T(W_s),U)$, and the counterfactual covariates are given by $\tilde{W}=(T(W_s),U)$.

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 $\theta=\mathrm{E}[\{g(T(W_s),U)-g(W_s,U)\}\ell(W_s)]$. 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 $P_{\tilde{W}}=\mathrm{Law}(T(W_s),U)$ 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 $W = (D,X,U) \mapsto (T(W_s),U) = (D + rt(W_s),X,U)$ 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: $\theta = \mathrm{E}[\ell(W_s)t(W_s)'\partial_d g(D,X,U)]$, where ℓ is a bounded weighting function and t is a bounded direction function. The short form of this parameter is $\theta_s = \mathrm{E}[\ell(W_s)t(W_s)'\partial_d g_s(D,X)]$. As regularity conditions, we suppose that $\mathrm{E}Y^2 < \infty$. Further for each (x,u) in the support of (X,U), and each d in $\mathscr{D}_{x,u}$, the support of D given $(X,U) = (x,u)$, the derivative maps $d \mapsto \partial_d g(d,x,u)$ and $d \mapsto g(w)\omega(w)$, for $\omega(w) := \ell(d,x)t(d,x)f(d \mid x,u)$, are continuously differentiable; the set $\mathscr{D}_{x,u}$ is bounded, and its boundary is piecewise-smooth; and $\omega(w)$ vanishes for each d in this boundary. Moreover, we assume the weak overlap: $\mathrm{E}[(\mathrm{div}_d\omega(W)/f(D \mid X,U))^2] < \infty$.

Another example is that of a policy that shifts the entire distribution of observed covariates, independently of U. 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 U. Note that this example is different from the transport example, since here the dependence between U and W_s is eliminated under the interventions.

Example 5 (Policy Effect from Changing Distribution of W_s). Define the long parameter as $\theta = \int [\int g(W_s,u)dP_U(u)]\,\ell(W_s)d\mu(W_s); \quad \mu(W_s) = F_1(W_s) - F_0(W_s)$, 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 $\mathscr{A} \times \text{support}(\ell)$. 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 result establishes the validity of the OVB formula for all examples.

Theorem 5 (OVB Validity in Examples 1-5). *Under the conditions stated in Examples 1,2,3,5, Assumptions 1 and 2 are satisfied. Under conditions stated in Example 4, Assumptions 1 and 2 are satisfied for the Hahn-Banach extension of the mapping $g \mapsto \mathrm{E}m(W,g)$ to the entire $L^2(P_W)$, given by $g \mapsto \mathrm{E}g(W)\alpha(W)$. The scores for Examples 1-5 are given by:*

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

The long RR and corresponding short RR are given by:

(1) $\alpha(w) = \frac{1(d=\bar{d})}{p(\bar{d}\mid x,u)}\bar{\ell}(x,u)$;	(1) $\alpha_s(W_s) = \frac{1(d=\bar{d})}{p(\bar{d}\mid x)}\bar{\ell}(x)$;
(2) $\alpha(w) = \frac{1(d=1)-1(d=0)}{p(d\mid x,u)}\bar{\ell}(x,u)$;	(2) $\alpha_s(W_s) = \frac{1(d=1)-1(d=0)}{p(\bar{d}\mid x)}\bar{\ell}(x)$;
(3) $\alpha(w) = \frac{dP_{\tilde{W}}(w)-dP_W(w)}{dP(w)}\ell(W_s)$;	(3) $\alpha_s(W_s) = \frac{dP_{\tilde{W}_s}(W_s)-dP_{W_s}(W_s)}{dP_{W_s}(W_s)}\ell(W_s)$
(4) $\alpha(w) = -\frac{\mathrm{div}_d(\ell(W_s)t(W_s)f(d\mid x,u))}{f(d\mid x,u)}$;	(4) $\alpha_s(W_s) = -\frac{\mathrm{div}_d(\ell(W_s)t(W_s)f(d\mid x))}{f(d\mid x)}$;
(5) $\alpha(w) = \frac{dP_U(u)\times d(F_1(W_s)-F_0(W_s))}{dP(w)}\ell(W_s)$;	(5) $\alpha_s(W_s) = \frac{d(F_1(W_s)-F_0(W_s))}{dP_{W_s}(W_s)}\ell(W_s)$;

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

As we have seen in Remarks 3, 4, and 5, it may be useful to further specialize the interpretation of the sensitivity parameters $1 - R^2_{\alpha\sim\alpha_s}$ for the many cases encompassed by the examples of Theorem 5. As this would be an extensive task, we leave such specializations to future work.

Statistical Inference. The examples above involve an extra nuisance function, such as the weighting function $\ell(W_s)$. If $\ell(W_s)$ is treated as a known function, then statistical inference can be conducted as before as in Theorem 4. If $\ell(W_s)$ is estimated from data, then its sampling variability needs to also be taken into account. For instance, in the case of the ATE, ATT and ATU of Example 2, the influence function for θ_s becomes,

$$\psi_{\theta_s}(Z;\theta_s,g_s,\alpha_s,\ell) := m(W_s,g_s)+(Y-g_s(W_s))\alpha_s(W_s)-\theta_s\times\ell(W_s). \tag{15}$$

More generally, if we have a moment of the form:

$$m(W_s,g_s) := m^*(W_s,g_s)\times\ell(W_s),$$

where the score $m^*(W_s,g_s)$ does not depend on any unknown nuisance function, and the weighting function $\ell(W_s)$ is known up to some unknown nuisance component $\eta_\ell(V)$, with V being a subset of the variables W_s, i.e. $\ell(W_s,\eta_\ell(V))$, then the influence function for θ_s also needs to take into account the influence from estimating η_ℓ. If η_ℓ corresponds to the solution to some other regression problem, i.e. $\mathrm{E}[Q\mid V]=\eta_\ell(V)$, with Q being an observed variable, then the influence function for θ_s becomes,

$$\psi_{\theta_s}(Z;\theta_s,g_s,\alpha_s,\ell) := m(W_s,g_s)+(Y-g_s(W_s))\alpha_s(W_s)+(Q-\eta_\ell(V))\,\mu(V)-\theta_s.$$

where $\mu(V) := \mathrm{E}[m^*(W_s,g_s)\,\partial_{\eta_\ell}\ell(W_s,\eta_\ell(V))\mid V]$. The ATT is a special case of this construction where $\eta_\ell=P(D=1)$, V is the empty set of variables, $\ell(W_s,\eta_\ell)=D/P(D=1)$, $Q=D$ and $\mu(V)=-\mathrm{E}[m^*(W_s,g_s)D/P(D=1)^2]=-\theta_s/P(D=1)$. Hence, the extra correction term takes the form $-\theta_s(D-P(D=1))/P(D=1)$, which can be combined with the term θ_s to yield the term $-\theta_sD/P(D=1)=-\theta_s\times\ell(W_s)$ that we provide in Equation (15).

Appendix B. Preliminaries

B.1. **Few Preliminaries.** To prove supporting lemmas we recall the following standard 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 $\langle f, g\rangle = \int fgdP$ 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-Nikodym Derivative. Consider a measure space $(\mathscr{X}, \mathscr{A})$ on which two σ-finite measures 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 $f : \mathscr{X} \to [0, \infty)$, such that for any measurable set $A \subseteq \mathscr{X}$, $\nu(A) = \int_A f\, d\mu$. The function f is conventionally denoted by $d\nu/d\mu$.

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

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

where S is the surface measure over the surface $\partial\mathscr{X}$ induced by V, and n is the outward normal vector.

Appendix C. Deferred Proofs and Examples

C.1. **Proof of Theorem 1.** The result follows from

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

using the fact that α_s is orthogonal to $g - g_s$ and g_s is orthogonal to $\alpha - \alpha_s$.

Moreover,

$$\mathrm{E}(g-g_s)(\alpha-\alpha_s) = \frac{\mathrm{E}(g-g_s)(\alpha-\alpha_s)}{\sqrt{\mathrm{E}(g-g_s)^2\mathrm{E}(\alpha-\alpha_s)^2}} \times \sqrt{\mathrm{E}(g-g_s)^2\mathrm{E}(\alpha-\alpha_s)^2} = \rho B,$$

where

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

We further note that B^2 factorizes as $B^2 = C_Y^2 C_D^2 S^2$, where $S^2 := \mathrm{E}(Y-g_s)^2\mathrm{E}\alpha_s^2$, and

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

and

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

where $\tilde{D} := D - \mathrm{E}[D \mid X, U]$, $\tilde{D}_s := D - \mathrm{E}[D \mid X]$, and $\tilde{U} = \mathrm{E}[D \mid X, U] - \mathrm{E}[D \mid X]$.

Here we used the observation that

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

holds because

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

In addition, we note

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

and,

$$R^2_{\tilde{D}_s\sim\tilde{U}} = \eta^2_{D\sim U|X}.$$

Finally, under correct specification,

$$\mathrm{E}(Y - g_s)^2 = \mathrm{E}[\mathrm{Var}(Y \mid D, X)], \quad R^2_{Y-g_s\sim g-g_s} = \eta^2_{Y\sim U|D,X}. \quad \square$$

C.2. **Relationship to the linear OVB formula.** The result of Theorem 1 generalizes the traditional OVB formula for linear models. To see how, without loss of generality, consider a scalar X and U. Now let,

$$g := \theta D + \beta X + \gamma U, \qquad g_s := \theta_s D + \beta_s X.$$

Here, as before, the models need not be correctly specified, and g, g_s can simply be the best linear approximations of the long and short CEF. Note the regression error is given by,

$$g - g_s = (\theta - \theta_s)D + (\beta - \beta_s)X + \gamma U,$$

and the error in the Riesz representer is given by,

$$\alpha - \alpha_s = \frac{\tilde{D}}{\mathrm{Var}(\tilde{D})} - \frac{\tilde{D}_s}{\mathrm{Var}(\tilde{D}_s)},$$

where here we write $\tilde{D}$ to denote D after partialling out X and U, and $\tilde{D}_s$ to denote D after partialling out X only. Now note $\mathrm{E}[D\tilde{D}] = \mathrm{Var}(\tilde{D})$ and $\mathrm{E}[D\tilde{D}_s] = \mathrm{Var}(\tilde{D}_s)$. Moreover, $\mathrm{E}[\tilde{D}X] = \mathrm{E}[\tilde{D}_s X] = 0$.

Thus, when taking the covariance of the two errors, all that is left out is the term containing the covariance of U with $\tilde{D}_s$:

$$\theta - \theta_s = \mathrm{E}[(g - g_s)(\alpha - \alpha_s)] = -\gamma \times \underbrace{\frac{\mathrm{E}[U\tilde{D}_s]}{\mathrm{Var}(\tilde{D}_s)}}_{\delta} = -\gamma\delta.$$

By the FWL theorem, δ is nothing but the regression coefficient of D on U, adjusting for X, thus,

$$\theta - \theta_s = -\gamma\delta$$

which equals the traditional omitted variable bias formula for linear models presented in textbooks (Goldberger, 1991; Angrist and Pischke, 2009; Wooldridge, 2010).

C.3. **Example of sharpness for PLM.** Suppose the observed data distribution, $P(Y,D,X)$ factorizes as,

$$
\begin{aligned}
X &\sim P(X), \\
D \mid X &\sim \mathcal{N}(\mu_d(X),\ \sigma_d^2), \\
Y \mid D,X &\sim \mathcal{N}(\theta_s\,(D-\mu_d(X)) + \mu_y(X),\ \sigma_y^2),
\end{aligned}
$$

where $\sigma_y^2 > 0$, $\sigma_d^2 > 0$, and $\mathcal{N}(\mu,\sigma^2)$ denotes the density of a standard normal with mean μ and variance σ^2.

We will show that we can always construct a full data law $P(Y,D,X,U)$ such that: (a) the long regression model is also partially linear on D; (b) the full data law marginalizes to the observed data law $P(Y,D,X)$; and (c) we achieve any value $\rho \in [-1,1]$, $\eta^2_{D\sim U|X} < 1$, $\eta^2_{Y\sim U|D,X} < 1$.

Our proof is constructive. Let the full data law $P(Y,D,X,U)$ be:

$$
\begin{aligned}
U_1 &\sim \mathcal{N}(0,1) \\
U_2 \mid U_1 &\sim \mathcal{N}(0,1) \\
X \mid U_2, U_1 &\sim P(X), \\
D \mid X, U_1, U_2 &\sim \mathcal{N}(\mu_d(X) + \delta U_1,\ (1-\eta_d^2)\times\sigma_d^2), \\
Y \mid D, X, U_1, U_2 &\sim \mathcal{N}(\theta\,(D-\mu_d(X)) + \mu_y(x) + \gamma_1 U_1 + \gamma_2 U_2,\ (1-\eta_y^2)\times\sigma_y^2),
\end{aligned}
$$

where,

$$
\theta = \theta_s + \rho\sqrt{\frac{\eta_y^2\eta_d^2}{1-\eta_d^2}\frac{\sigma_y^2}{\sigma_d^2}}, \quad \delta = \sqrt{\eta_d^2\sigma_d^2}, \quad \gamma_1 = -\rho\sqrt{\frac{\eta_y^2\sigma_y^2}{1-\eta_d^2}}, \quad \gamma_2 = \sqrt{(1-\rho^2)\eta_y^2\sigma_y^2}.
$$

We define $U = (U_1, U_2)$. We first show that this distributions marginalizes to the observed data. First, since U_1 and U_2 are jointly independent Gaussian, this implies $D \mid X$ and $Y \mid D, X$ are also Gaussian. It thus suffices to show that the marginal means and variances are preserved.

Starting with $D \mid X$:

$$\mathrm{E}[D \mid X] = \mathrm{E}[\mu_d(X) + \delta U_1 \mid X] = \mu_d(X) + \delta \mathrm{E}[U_1 \mid X] = \mu_d(X),$$

$$\mathrm{Var}(D \mid X) = \delta^2 \,\mathrm{Var}(U \mid X) + (1 - \eta_d^2)\sigma_d^2 = \eta_d^2\sigma_d^2 + (1 - \eta_d^2)\sigma_d^2 = \sigma_d^2.$$

That is the marginal distribution of $D \mid X$ matches that of the observed data. Now moving to $Y \mid D, X$, first note that:

$$U_1 \mid D, X \sim \mathcal{N}\left(\sqrt{\frac{\eta_d^2}{\sigma_d^2}}(D - \mu_d(X)),\ 1 - \eta_d^2\right)$$

and $U_2 \perp\!\!\!\perp \{U_1, D, X\}$. Thus, for the CEF we obtain,

$$\begin{aligned}
\mathrm{E}[Y \mid D, X] &= \mathrm{E}[\theta\,(D - \mu_d(X)) + \mu_y(x) + \gamma_1 U_1 + \gamma_2 U_2 \mid D, X] \\
&= \theta\,(D - \mu_d(X)) + \mu_y(x) + \gamma_1 \mathrm{E}[U_1 \mid D, X] + \gamma_2 \mathrm{E}[U_2 \mid D, X] \\
&= \theta\,(D - \mu_d(X)) + \mu_y(x) + \gamma_1 \times \sqrt{\frac{\eta_d^2}{\sigma_d^2}}(D - \mu_d(X)) \\
&= \left(\theta_s + \rho\sqrt{\frac{\eta_y^2\eta_d^2}{1-\eta_d^2}\frac{\sigma_y^2}{\sigma_d^2}}\right)(D - \mu_d(X)) + \mu_y(x) \\
&\qquad - \rho\sqrt{\frac{\eta_y^2\sigma_y^2}{1-\eta_d^2}} \times \sqrt{\frac{\eta_d^2}{\sigma_d^2}}(D - \mu_d(X)) \\
&= \theta_s\,(D - \mu_d(X)) + \mu_y(x).
\end{aligned}$$

And for the variance,

$$\begin{aligned}
\mathrm{Var}(Y \mid D, X) &= \gamma_1^2 \,\mathrm{Var}(U_1 \mid D, X) + \gamma_2^2 \,\mathrm{Var}(U_2 \mid D, X) + (1 - \eta_y^2) \times \sigma_y^2 \\
&= \gamma_1^2 \times (1 - \eta_d^2) + \gamma_2^2 + (1 - \eta_y^2) \times \sigma_y^2 \\
&= \rho^2 \frac{\eta_y^2\sigma_y^2}{1-\eta_d^2}(1 - \eta_d^2) + (1 - \rho^2)\eta_y^2\sigma_y^2 + (1 - \eta_y^2) \times \sigma_y^2
\end{aligned}$$

$$
\begin{aligned}
&= \rho^2\eta_y^2\sigma_y^2 + (1-\rho^2)\eta_y^2\sigma_y^2 + (1-\eta_y^2) \times \sigma_y^2 \\
&= \sigma_y^2.
\end{aligned}
$$

Therefore, the marginal distribution of $Y \mid D, X$ matches that of the observed data. We now show that we attain the desired sensitivity parameters.

First note that:

$$
\eta^2_{D\sim U|X} = 1 - \frac{\mathrm{E}[\mathrm{Var}(D \mid X, U)]}{\mathrm{E}[\mathrm{Var}(D \mid X)]} = 1 - \frac{(1-\eta_d^2)\sigma_d^2}{\sigma_d^2} = \eta_d^2.
$$

Also,

$$
\eta^2_{Y\sim U|D,X} = 1 - \frac{\mathrm{E}[\mathrm{Var}(Y \mid D, X, U)]}{\mathrm{E}[\mathrm{Var}(Y \mid D, X)]} = 1 - \frac{(1-\eta_y^2)\sigma_y^2}{\sigma_y^2} = \eta_y^2.
$$

Finally, since,

$$
\theta - \theta_s = \rho\sqrt{\frac{\eta_y^2\eta_d^2}{1-\eta_d^2}\frac{\sigma_y^2}{\sigma_d^2}},
$$

by Theorem 1 we must have $Cor(g - g_s, \alpha - \alpha_s) = \rho$. □

C.4. **Proof of Lemma 1.** We have from the Riesz-Frechet theory presented in Appendix B.1 and in particular by a direct application of the Riesz-Frechet representation theorem that given the assumptions of the Lemma, there exist unique functions $\alpha \in \Gamma$, $\alpha_s \in \Gamma_s$ such that:

$$
\forall \gamma \in \Gamma : \mathrm{E}m(W,\gamma) = \mathrm{E}\gamma(W)\alpha(W), \quad \forall \gamma_s \in \Gamma_s : \mathrm{E}m(W,\gamma_s) = \mathrm{E}m(W_s,\gamma_s) = \mathrm{E}\gamma_s(W)\alpha_s(W),
$$

Next we show that α_s is given by a projection of α onto Γ_s. Since $\Gamma_s \subseteq \Gamma$, we have that α represents the functional over Γ_s but it is not itself in Γ_s. Consider the decomposition of α in the orthogonal projection β_s of α onto Γ_s and the residual e:

$$
\alpha(W) = \beta_s(W_s) + e(W), \quad \mathrm{E}e(W)\gamma_s(W_s) = 0, \text{ for all } \gamma_s \text{ in } \Gamma_s.
$$

Then we have that for all $\gamma_s \in \Gamma_s$

$$
\mathrm{E}m(W,\gamma_s) = \mathrm{E}\gamma_s(W_s)\alpha(W) = \mathrm{E}\gamma_s(W_s)\beta_s(W_s) + \mathrm{E}\gamma_s(W_s)e(W) = \mathrm{E}\gamma_s(W_s)\beta_s(W_s).
$$

That is, the orthogonal projection β_s of α onto Γ_s is a RR of the functional over functions in Γ_s. Moreover, by the RF theory the RR is unique in Γ_s. By uniqueness of the RR over Γ_s, we must have that the projected β_s coincides with α_s and therefore α_s is the orthogonal projection of α onto Γ_s. The latter means that we can decompose:

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

where we used the fact that α_s is the orthogonal projection of α onto Γ_s and that $\alpha_s \in \Gamma_s$, which implies that $\mathrm{E}(\alpha-\alpha_s)\alpha_s = 0$. □

C.5. **Relationship of Restricted RRs of the ATE in PLM.** Note that, as per Lemma 1, α_s is the partially linear projection of α on W_s. Consider the minimization problem in the partially linear regression:

$$\min_{\beta,h} \mathrm{E}[(\alpha - \beta D - h(X))^2]$$

First, note that, for any β, the optimal function h must equal

$$h^*(X;\beta) = \mathrm{E}[\alpha - \beta D \mid X] = -\beta \mathrm{E}[D \mid X].$$

Plugging that back into the minimization problem we now have,

$$\min_{\beta} E\left[(\alpha - \beta(D - \mathrm{E}[D \mid X]))^2\right].$$

The solution for β is then,

$$\beta^* = \frac{\mathrm{E}[\alpha(D-\mathrm{E}[D \mid X])]}{\mathrm{E}[(D-\mathrm{E}[D \mid X])^2]} = \frac{1}{\mathrm{E}[(D-\mathrm{E}[D \mid X])^2]}.$$

Thus,

$$\beta^* D + h^*(X;\beta^*) = \frac{D}{E[(D-E[D \mid X])^2]} - \frac{E[D \mid X]}{E[(D-E[D \mid X])^2]} = \frac{D-E[D \mid X]}{E[(D-E[D \mid X])^2]} = \alpha_s. \quad \square$$

C.6. **Proof of Lemma 2.** We have from the Riesz-Frechet theory presented in Appendix B.1 and in particular by a direct application of the Riesz-Frechet representation theorem that given the assumptions of the Lemma, there exists a unique $\bar{\alpha} \in L^2(P_W)$, such that:

$$\forall f \in L^2(P_W) : \mathrm{E}m(W, f) = \mathrm{E}f(W)\bar{\alpha}(W),$$

Since, $\Gamma \subseteq L^2(P_W)$, the latter implies that:

$$\forall \gamma \in \Gamma : \mathrm{E}m(W, \gamma) = \mathrm{E}\gamma(W)\bar{\alpha}(W),$$

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

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

Then we have that for all $\gamma \in \Gamma$

$$\mathrm{E}m(W;,\gamma) = \mathrm{E}\gamma(W)\bar{\alpha}(W) = \mathrm{E}\gamma(W)\alpha(W) + \mathrm{E}\gamma(W)e(W) = \mathrm{E}\gamma(W)\alpha(W).$$

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

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

C.7. **Proof of Theorem 2.** We decompose Γ into $\Gamma_s \subseteq \Gamma \cap L^2(P_{W_s})$ and its orthocomplement $\Gamma_s^{\perp}$,

$$\Gamma = \Gamma_s \oplus \Gamma_s^{\perp},$$

so that any element $m_s \in \Gamma_s$ is orthogonal to any $e \in \Gamma_s^{\perp}$ in the sense that

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

By Lemma 1, we can write:

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

where $\alpha \in \Gamma$ and $\alpha_s \in \Gamma_s$ satisfy the properties stated in the Lemma 1. In particular, α_s is the orthogonal projection of α onto Γ_s. Therefore, by definition we have that $\alpha - \alpha_s \in \Gamma_s^{\perp}$, i.e.,

$$\forall \gamma_s \in \Gamma_s : \mathrm{E}(\alpha(W) - \alpha_s(W_s))\gamma_s(W_s) = 0.$$

Moreover, we can also argue that $g - g_s \in \Gamma_s^{\perp}$. By the definition of g and g_s, and the fact that both Γ and Γ_s are closed linear subspaces, we have that:

$$\forall \gamma \in \Gamma : \mathrm{E}(Y - g(W))\gamma(W) = 0 \tag{16}$$

$$\forall \gamma_s \in \Gamma_s : \mathrm{E}(Y - g_s(W_s))\gamma_s(W_s) = 0 \tag{17}$$

Since $\Gamma_s \subseteq \Gamma$, the first condition also implies that:

$$\forall \gamma_s \in \Gamma_s : \mathrm{E}(Y - g(W))\gamma_s(W_s) = 0 \tag{18}$$

Subtracting the last two conditions, we get:

$$\forall \gamma_s \in \Gamma_s : \mathrm{E}(g(W) - g_s(W_s))\gamma_s(W_s) = 0 \tag{19}$$

which by definition implies that $g - g_s \in \Gamma_s^{\perp}$. The claim of the theorem follows from

$$\begin{aligned}
\theta - \theta_s &= \mathrm{E}g\alpha - \mathrm{E}g_s\alpha_s \\
&= \mathrm{E}(g_s + g - g_s)(\alpha_s + \alpha - \alpha_s) - \mathrm{E}g_s\alpha_s \\
&= \mathrm{E}g_s(\alpha - \alpha_s) + \mathrm{E}\alpha_s(g - g_s) + \mathrm{E}(g - g_s)(\alpha - \alpha_s) \\
&= \mathrm{E}(g - g_s)(\alpha - \alpha_s),
\end{aligned}$$

using the fact that $\alpha_s \in \Gamma_s$ is orthogonal to $g - g_s \in \Gamma_s^{\perp}$ and $g_s \in \Gamma_s$ is orthogonal to $\alpha - \alpha_s \in \Gamma_s^{\perp}$. By the same argument as in the proof of Theorem 1, we also obtain,

$$\mathrm{E}(g - g_s)(\alpha - \alpha_s) = \frac{\mathrm{E}(g - g_s)(\alpha - \alpha_s)}{\sqrt{\mathrm{E}(g - g_s)^2\mathrm{E}(\alpha - \alpha_s)^2}} \times \sqrt{\mathrm{E}(g - g_s)^2\mathrm{E}(\alpha - \alpha_s)^2} = \rho B,$$

where

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

Then, using the fact that

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

and

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

we obtain the final expression of the OVB formula. □

C.8. **Proof of Theorem 3.** In this section we use the following shorthand notations: $\pi = P(D=1 \mid X,U)$, $\pi_s := P(D=1 \mid X)$, $O := \pi/(1-\pi)$, $O_s := \pi_s/(1-\pi_s)$, $V := \pi \times (1-\pi)$, $V_s := \pi_s \times (1-\pi_s)$, $p := P(D=1) = E[\pi_s]$. When conditioning on the event $X=x$, we use $\pi_s(x) := P(D=1 \mid X=x)$, $O_s(x) := \pi_s(x)/(1-\pi_s(x))$ and $V_s(x) := \pi_s(x) \times (1-\pi_s(x))$.

Before proving the theorem, the following lemmas will be useful.

Lemma 4 (Bounds on the long propensity score)**.** *Under the marginal sensitivity model, conditionally on $X=x$, π is a bounded random variable, with support on $[\pi_\ell(x), \pi_u(x)]$, where*

$$\pi_\ell(x) = \frac{\Lambda^{-1}O_s(x)}{1+\Lambda^{-1}O_s(x)}, \qquad \pi_u(x) = \frac{\Lambda O_s(x)}{1+\Lambda O_s(x)}.$$

Moreover, $\mathrm{E}[\pi \mid X=x] = \pi_s(x)$.

Proof. Recall the marginal sensitivity model imposes $\Lambda^{-1}O_s(x) \leq O(x,u) \leq \Lambda O_s(x)$. Use the fact that $O(x,u) = \pi(x,u)/(1-\pi(x,u))$ to obtain the lower and upper bounds π_ℓ and π_u. The law of iterated expectations gives $\mathrm{E}[\pi \mid X=x] = \pi_s(x)$. □

The next lemma appears in Madansky (1959), who credits Edmundson (1957) for first deriving it.

Lemma 5 (Edmundson-Madansky Bounds)**.** *Let Z be a bounded random variable with support on $[a,b]$ and mean* $\mathrm{E}Z = \mu$. *Then, for convex function f,*

$$\mathrm{E}[f(Z)] \leq \omega f(a) + (1-\omega) f(b), \quad \textit{with} \quad \omega = \frac{b-\mu}{b-a}.$$

Moreover, the bound is sharp. If f is concave, the inequality is reversed.

Proof. For any $z \in [a,b]$ we can write $z = \omega(z)a + (1-\omega(z))b$, where $\omega(z) = \frac{b-z}{b-a}$. Convexity of f gives $f(z) \leq \omega(z)f(a) + (1-\omega(z))f(b)$. Applying this to the random variable and taking the expectation yields the inequality. To verify that the bound is sharp, set Z to take values on $\{a,b\}$, with $P(Z=a) = \omega$, then note that $\mathrm{E}[f(Z)] = \omega f(a) + (1-\omega)f(b)$ and $\mathrm{E}Z = \mu$. For the concave case, apply the result to $-f$. □

We can now combine Lemmas 4 and 5 to obtain sharp bounds on any convex (or concave) function of the long propensity score, under the MSM.

Lemma 6 (Sharp bounds for convex f under MSM)**.** *Under the marginal sensitivity model, for any convex function f,*

$$\mathrm{E}[f(\pi) \mid X = x] \leq \omega(x)f(\pi_\ell(x)) + (1-\omega)f(\pi_u(x)),$$

where,

$$\omega(x) = \frac{\Lambda + O_s(x)}{(1+\Lambda)(1+O_s(x))} = \frac{\pi_s(x) + \Lambda(1-\pi_s(x))}{1+\Lambda}.$$

Moreover, the bound is sharp. If f is concave, the inequality is reversed.

Proof. By Lemma 4, $\pi \mid X = x$ is a bounded random variable with support on $[\pi_\ell(x), \pi_u(x)]$ and mean $\pi_s(x)$. Lemma 5 then yields the result. □

We are now ready to prove the theorem. To recall, we want to show that, under the MSM,

$$\begin{aligned}
1 - R^2_{\alpha_{ATT} \sim \alpha_{ATT,s}} &= 1 - \frac{\mathrm{E}O_s}{\mathrm{E}O} \leq \frac{(\Lambda-1)^2\mathrm{E}O_s - (\Lambda-1)^2 p}{((\Lambda-1)^2+\Lambda)\mathrm{E}O_s - (\Lambda-1)^2 p} \leq \frac{(\Lambda-1)^2}{(\Lambda-1)^2+\Lambda},\\
1 - R^2_{\alpha_{ATE} \sim \alpha_{ATE,s}} &= 1 - \frac{\mathrm{E}[1/V_s]}{\mathrm{E}[1/V]} \leq \frac{(\Lambda-1)^2\mathrm{E}[1/V_s] - 3(\Lambda-1)^2}{((\Lambda-1)^2+\Lambda)\mathrm{E}[1/V_s] - 3(\Lambda-1)^2} \leq \frac{(\Lambda-1)^2}{(\Lambda-1)^2+\Lambda},\\
1 - R^2_{\alpha_{PLM} \sim \alpha_{PLM,s}} &= 1 - \frac{\mathrm{E}V}{\mathrm{E}V_s} \leq 1 - \frac{\mathrm{E}\left[\frac{\Lambda V_s}{\Lambda + (\Lambda-1)^2 V_s}\right]}{\mathrm{E}V_s} \leq \frac{(\Lambda-1)^2}{(\Lambda-1)^2+4\Lambda}.
\end{aligned}$$

The first bound is sharp given Λ and observed data. The second bound is sharp given Λ alone.

Proof of Theorem 3. Our proof strategy is the following. We first use Lemma 6 to obtain sharp bounds that hold conditionally for any $X = x$. Then the marginal bounds follow from taking expectations. The data-independent version of the bound is obtained by taking the supremum over any possible realization of the observed data.

Gain in Odds. We start with the gain in average odds. Note the odds function is convex. Applying Lemma 6 yields the sharp bound,

$$\begin{aligned}
\mathrm{E}[O \mid X = x] &\leq \omega(x)\pi_\ell(x)/(1-\pi_\ell(x)) + (1-\omega(x))\pi_u(x)/(1-\pi_u(x)) \\
&= \omega(x)\Lambda^{-1}O_s(x) + (1-\omega(x))\Lambda O_s(x) \\
&= (\omega(x)\Lambda^{-1} + (1-\omega(x))\Lambda)O_s(x).
\end{aligned}$$

Using the identities $\omega(x)\Lambda^{-1} + (1-\omega(x))\Lambda = 1 + \left(\frac{(\Lambda-1)^2}{\Lambda}\right)\pi_s(x)$ and $O_s(x)\pi_s(x) = O_s(x) - \pi_s(x)$ gives us

$$\mathrm{E}[O \mid X = x] \leq \left(\frac{\Lambda + (\Lambda-1)^2}{\Lambda}\right)O_s(x) - \left(\frac{(\Lambda-1)^2}{\Lambda}\right)\pi_s(x).$$

Applying this to the random variable, taking expectations and rearranging yields the data-dependent bound. To obtain the data-independent bound, take $\mathrm{E}O_s \to \infty$.

Gain in Precision. We now move to the gains in average precision. First note we can write the inverse variance as

$$\frac{1}{V} = O + \frac{1}{O} + 2.$$

By symmetry, the previous bound holds for $\mathrm{E}[1/O \mid X = x]$, replacing $O_s(x)$ with $1/O_s(x)$ and π_s with $1 - \pi_s$. Then,

$$\mathrm{E}[1/V \mid X = x] \leq \left(\frac{\Lambda + (\Lambda-1)^2}{\Lambda}\right)O_s(x) + \left(\frac{\Lambda + (\Lambda-1)^2}{\Lambda}\right)1/O_s(x) - \frac{(\Lambda-1)^2}{\Lambda} + 2$$

Rearranging terms, and using again the identity $1/V_s(x) = O_s(x) + 1/O_s(x) + 2$ we can simplify the last expression to,

$$\mathrm{E}[1/V \mid X = x] \leq \left(\frac{\Lambda + (\Lambda - 1)^2}{\Lambda}\right)(1/V_s(x)) - 3\left(\frac{(\Lambda - 1)^2}{\Lambda}\right).$$

Applying this to the random variable, taking expectations and rearranging yields the data-dependent bound. To obtain the data-independent bound, take $\mathrm{E}[1/V_s] \to \infty$.

Gain in Variance Explained. Moving to the non-parametric partial R^2, note that the variance function is concave. We thus have the sharp *lower* bound,

$$\begin{aligned}\mathrm{E}[V \mid X = x] &\geq \omega(x) \times \pi_\ell(x) \times (1 - \pi_\ell(x)) + (1 - \omega(x)) \times \pi_u(x) \times (1 - \pi_u(x)) \\ &= \frac{\Lambda O_s(x)}{(\Lambda + O_s(x))(1 + \Lambda O_s(x))}.\end{aligned}$$

Divide the numerator and denominator by $(1 + O_s(x))^2$ and use the identities $V_s(x) = O_s(x)/(1 + O_s(x))^2$, $(\Lambda + O_s(x))(1 + \Lambda O_s(x)) = \Lambda(1 + O_s(x))^2 + O_s(x)(\Lambda - 1)^2$ to obtain,

$$\mathrm{E}[V \mid X = x] \geq \frac{\Lambda V_s(x)}{\Lambda + (\Lambda - 1)^2 V_s(x)}.$$

Applying this to the random variable, taking expectations and rearranging yields the data-dependent bound. To obtain the data-independent bound, take $V_s(x) \to 1/4$.

Bounds on C_D. Bounds on C_D can be obtained via the mapping $C_D^2 = (1 - R^2_{\alpha\sim\alpha_s})/R^2_{\alpha\sim\alpha_s}$. □

Remark 7. These bounds are in fact reached by the adversarial confounder of the construction in Dorn and Guo (2023), when proving sharpness of their MSM bounds for the average potential outcome, ATE and ATT.

C.9. **Rosenbaum's model and MSM can have unbounded parameters.** Recall that the marginal sensitivity model posits $\Lambda \geq 1$ such that

$$\frac{1}{\Lambda} \leq \frac{\mathrm{Odds}(D \mid X = x, U = u)}{\mathrm{Odds}(D \mid X = x)} \leq \Lambda, \qquad \forall x \in \mathscr{X}, u \in \mathscr{U}.$$

Similarly, Rosenbaum's model posits a $\Gamma \geq 1$ such that

$$\frac{1}{\Gamma} \leq \frac{\text{Odds}(D \mid X = x, U = u)}{\text{Odds}(D \mid X = x, U = u')} \leq \Gamma, \qquad \forall x \in \mathscr{X}, u \in \mathscr{U}.$$

We now show that these two parameters can be unbounded while the OVB is actually small. Let the unobserved confounder U be normally distributed and consider the case with no observed covariates X. Let the full propensity score be

$$P(D = 1 \mid U = u) = \frac{e^{\rho u}}{1 + e^{\rho u}}. \tag{20}$$

We then have that $\text{Odds}(D \mid X = x, U = u)/\,\text{Odds}(D \mid X = x, U = u') = e^{\rho(u-u')}$ and $\text{Odds}(D \mid X = x, U = u)/\,\text{Odds}(D \mid X = x) = e^{\rho u}$.

Thus, the true sensitivity parameters Γ and Λ are unbounded,

$$\Gamma = \Lambda = \infty,$$

once $\rho \neq 0$. In contrast, the true OVB parameter $1 - R^2_{\alpha \sim \alpha_s}$ (either for the ATE or ATT) converges to 0 as $\rho \to 0$. For example, with $\rho = .1$, the worst-case odds ratio is infinite, whereas the true OVB bound is very tight, since we have $1 - R^2_{\alpha_{ATE} \sim \alpha_{ATE,s}} \approx 0.0025$ and $1 - R^2_{\alpha_{ATT} \sim \alpha_{ATT,s}} \approx 0.005$.

In summary, in this example, the true sensitivity parameters translate into tight bounds on the ATE or ATT in our OVB-based approach, versus uninformative bounds in worst-case odds-ratio approaches. We note that similar discussion applies to approaches that constrain worst-case risk-ratios, such as those in Ding and VanderWeele (2016).

C.10. **Proof of Lemma 3 and Theorem 4.** The Lemma follows from the application of Theorem 3.1 and Theorem 3.2 in Chernozhukov et al. (2018a). Valid estimation of covariance follows similarly to the proof of Theorem 3.2 in Chernozhukov et al. (2018a). The first result of Theorem 4 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. □

C.11. **Proof of Theorem 5.** Here the argument is similar to Chernozhukov et al. (2018b), 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 $\mathrm{E}Y^2 < \infty$, we have $g \in L^2(P)$. Therefore, $|\mathrm{E}\alpha g| < \|\alpha\|_{P,2}\|g\|_{P,2} < \infty$ in any of the calculations below.

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

In Example 1, recall that $\bar{\ell}(X,U) := \mathrm{E}[\ell(W_s) \mid X,U]$. Then since $dP(d,x,u) = \sum_{j=0}^{1} 1(j = d)P[D = j \mid X = x, U = u]dP(x,u)$ by Bayes rule, we have

$$\mathrm{E}g(W)\alpha(W) = \int g(d,x,h)\frac{1(d=\bar{d})\bar{\ell}(x,u)}{P[D=\bar{d} \mid X=x, U=u]}dP(d,x,u) = \int g(\bar{d},x,u)\bar{\ell}(x,u)dP(x,u)$$

$$= \mathrm{E}g(\bar{d},X,U)\bar{\ell}(X,U) = \mathrm{E}g(\bar{d},X,U)\ell(W_s) = \theta(g),$$

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 $dP_{\tilde{W}}$ to dP_W, given the assumed absolutely continuity of the former with respect to the latter. Then we have

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

$$= \int \ell(W_s)(g(T(W_s),u) - g(W_s,u))dP_W(w) = \theta(g).$$

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

$$\begin{aligned}
\mathrm{E}g(W)\alpha(W) &= -\int\int g(w)\frac{\mathrm{div}_d(\ell(W_s)t(W_s)f(d \mid x,u))}{f(d \mid x,u)}f(d \mid x,u)\mathrm{d}d\mathrm{d}P(x,u) \\
&= -\int\int g(w)\mathrm{div}_d(\ell(W_s)t(W_s)f(d \mid x,u))\mathrm{d}d\mathrm{d}P(x,u) \\
&= -\int\int_{\partial\mathscr{D}_{u,x}} g(w)t(W_s)'\ell(W_s)f(d \mid x,u)n_{u,x}(u)\mathrm{d}S(d)\mathrm{d}P(x,u)
\end{aligned}$$

$$
\begin{aligned}
&+\int\int \partial_d g(w)' t(W_s)\ell(W_s) f(d \mid x,u) \mathrm{d}d\mathrm{d}P(x,u) \\
= \quad &\int\int \partial_d g(w)' t(W_s)\ell(W_s) f(d \mid x,u) \mathrm{d}d\mathrm{d}P(x,u) = \theta(g),
\end{aligned}
$$

where we used the integration by parts and that $\ell(W_s)t(W_s)f(d \mid x,u)$ vanishes for any d in the boundary of $\mathscr{D}_{x,u}$.

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

$$
\begin{aligned}
\mathrm{E}g(W)\alpha(W) &= \int g\ell\left(\frac{[dP_U \times d(F_1 - F_0)]}{dP_W}\right) dP_W \\
&= \int g(W_s,u)\ell(W_s) dP_U(u) d(F_1 - F_0)(W_s) = \theta(g).
\end{aligned}
$$

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 the scores given in Section 5.

Note that we do not 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. □

C.12. **Discussion on sharpness.** Here we note that certain model assumptions in combination with the distribution of observed data P can place restrictions on the admissible values of sensitivity parameters. Recall the omitted variable bias formula is,

$$
\theta - \theta_s = \rho \times \sqrt{R^2_{Y-g_s\sim g-g_s} \times \left(\frac{1 - R^2_{\alpha\sim\alpha_s}}{R^2_{\alpha\sim\alpha_s}}\right)} \times S.
$$

Given plausibility judgments on the maximum values of the components of the bias formula, i.e., $|\rho| \leq \rho^{\max}$, $R^2_{Y-g_s\sim g-g_s} \leq R_Y^{2\max}$ and $1 - R^2_{\alpha\sim\alpha_s} \leq R_D^{2\max}$, the OVB formula immediately implies the bound,

$$
|\theta - \theta_s| \leq \rho^{\max} \times \sqrt{R_Y^{2\max} \times \left(\frac{R_D^{2\max}}{1 - R_D^{2\max}}\right)} \times S.
$$

Formally we say this bound is sharp given model assumptions $\mathscr{M}$, and observed data $\mathscr{P}$, if, for any triplet $(\rho^{\max}, R_Y^{2\max}, R_D^{2\max}) \in [0,1)^3$, we can construct a full data law such that: (a) it is compatible with $\mathscr{M}$ and $\mathscr{P}$; and, (b) achieves (or gets arbitrarily close to) $|\rho| = \rho^{\max}$, $R^2_{Y-g_s \sim g-g_s} = R_Y^{2\max}$ and $1 - R^2_{\alpha \sim \alpha_s} = R_D^{2\max}$. In general, one takes $\mathscr{P}$ to be the full joint distribution of the observed data.

It is easy to see that the OVB bound may not be sharp in certain cases, depending on $\mathscr{M}$ and $\mathscr{P}$. One example is the case of the ATE when the outcome is bounded. If, say $0 \leq Y \leq 1$, the bias $|\theta - \theta_s|$ itself is also naturally bounded. This means that the sensitivity parameters are not variation independent. When that is the case, some combinations of the sensitivity parameters (especially when $R_D^{2\max}$ is high enough), may not be compatible with the observed data, meaning that the bound is conservative. However, note the bound could still be sharp for most plausible combinations of the sensitivity parameters, even in such scenarios. Having said that, a simple solution to tighten the bounds in such scenarios is to cap the OVB formula at the “no assumptions” bounds, whenever these are known (e.g. Manski’s bounds). This approach always yields valid bounds that are at least as informative as the unconstrained OVB formula.

More generally, our OVB framework generates bias formulas for a wide range of target parameters and model specifications. Given this breadth, it is difficult to establish general sharpness claims that apply broadly across many functionals, models, and data constraints. We believe specialized results and case-by-case analyses will be needed as demand for such results grows. The example in Appendix C.3 illustrates how one can derive such results, and we leave a broader characterization of sharpness to future work. Finally, we note that these characterizations do not preclude the usefulness of the bounds. Indeed, many widely used sensitivity analysis methods lacked general sharp results for years, or even decades, after their initial proposals. Indeed the sharpness results of Dorn and Guo (2023) and Dorn et al. (2025) were established almost two decades after Tan (2006).

Appendix D. Extended Literature Review

We now provide a more extended discussion of the related literature on sensitivity analysis. We focus the discussion on recent methods, and on how they differ from our proposal. We refer readers to Liu et al. (2013), Richardson et al. (2014), Cinelli and Hazlett (2020), and Scharfstein et al. (2021) for further details.

In contrast to our approach, many of the earlier works on sensitivity analyses demand from users a rather extensive specification, or parameterization, of the nature of unobserved confounders. This could range from positing the marginal (or conditional) distribution of these latent variables, along with specifying how such confounders would enter the outcome or treatment equations (e.g., entering linearly). Among such proposals, with varying degrees of requirements and parametric assumptions, we can find, e.g., Rosenbaum and Rubin (1983b), Imbens (2003), Vanderweele and Arah (2011), Dorie et al. (2016), Altonji et al. (2005), Cinelli et al. (2019),Zhang et al. (2021).

Another branch of the sensitivity literature requires users to specify instead a "tilting," "selection," or "bias" function, directly parameterizing the difference between the conditional distribution of the outcome under treatment (control) between treated and control units; or, when the target parameter is the ATE, just parameterizing the difference in conditional means. Earlier work on this area goes back to Robins (1999), Brumback et al. (2004), Blackwell (2013), and Luedtke et al. (2015), with more recent works from Franks et al. (2020) and Scharfstein et al. (2021), the latter with a special focus on binary treatments, and flexible semi-parametric estimation procedures. Our proposal differs from this literature in that we do not model the bias directly, instead we impose constraints on the maximum explanatory power of confounders.

Continuing with binary treatments, many sensitivity proposals focus on this special case. They differ mainly on how to parameterize departures from random assignment. For instance, Masten and Poirier (2018) places bounds on the *difference* between the treatment assignment distribution, conditioning and not conditioning on potential outcomes, whereas Rosenbaum (1987) and more recently Tan (2006); Zhao et al. (2019); Yadlowsky et al. (2022) place bounds on the worst-case *odds* of such distributions. Bonvini and Kennedy (2021), on the other hand, propose a contamination model approach, placing restriction on the *proportion of confounded units*. Our approach is different from all these approaches in at least two main ways. First, we do not restrict our analyses to the binary treatment case. Second, even in the important case of a binary treatment, we parameterize violations of ignorability via the exact OVB formula for each estimand (see discussion in Section 3.4), resulting in different sensitivity parameters, such as the *gains in average precision* for the ATE, or *gains in average odds* for the ATT. Our sensitivity parameters are thus different from these approaches (we provide a numerical example in Appendix C.9, which demonstrates practical and theoretical value of the new parameterization).

Other sensitivity results, while allowing for general confounders, treatments and outcomes, restrict their attention to specific target parameters. For example, Ding and VanderWeele (2016) derive general bounds for the risk-ratio, with sensitivity parameters also in terms of risk-ratios. Our approach is thus different both in terms of target parameters (continuous linear functionals of the CEF), and in terms of sensitivity parameters (R^2 based sensitivity parameters). Cinelli and Hazlett (2020) derive bounds for linear regression coefficients. Their result is a special case of ours when the target functional is the coefficient of a linear projection. Their approach does *not* cover nonlinear regression and the causal parameters that we study here (e.g., it does not cover the ATE in the nonparametric model with a binary treatment). Finally, Detommaso et al. (2021) provide an alternative expression for omitted variable bias of average causal derivatives.

Appendix E. Benchmarking Analysis

Here we describe our new approach to benchmarking in nonparametric models. Our analysis is partly inspired by benchmarking analyses previously proposed in Imbens (2003), Altonji et al. (2005), Oster (2019), and more recently Cinelli and Hazlett (2020). In particular our proposal is closest in nature to the latter reference for linear regression, by postulating that the gains in explanatory power due to latent variables is similar to the gains in explanatory power of observed variables.

E.1. **Notation.** We start by setting notation. For a given observed covariate X_j, let X_{-j} denote the set of all other observed covariates X except X_j. Let $g_{s,-j}$ and $\alpha_{s,-j}$ denote the short regression function and the short RR *excluding* covariate X_j. We define the gain in the explanatory power of X_j with the RR as:

$$1-R^2_{\alpha_s\sim\alpha_{s,-j}} = \frac{\mathrm{E}\alpha_s^2 - \mathrm{E}\alpha_{s,-j}^2}{\mathrm{E}\alpha_s^2}.$$

We also define the change in estimate, $\Delta\theta_{s,j} := \mathrm{E}m(W, g_{s,-j}) - \mathrm{E}m(W, g_s)$, and the correlation, $\rho_j := \mathrm{Cor}(g_{s,-j} - g_s, \alpha_s - \alpha_{s,-j})$.

E.2. **Relative bounds on** $1-R^2_{\alpha\sim\alpha_s}$**.** Note we can write $1-R^2_{\alpha\sim\alpha_s}$ as,

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

Now dividing and multiplying the fraction by $\mathrm{E}\alpha_{s,-j}^2$ we obtain the following decomposition:

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

The numerator stands for the additive gain in variation that the latent variables U create in the RR, in addition to what X_j creates. We can now define the following measure $k_{D,j}$ of *relative* strength of U, which captures how much U adds in terms of variation explained of the RR, as compared to the observed gains due to X_j,

$$k_{D,j} := \frac{R^2_{\alpha_s\sim\alpha_{s,-j}} - R^2_{\alpha\sim\alpha_{s,-j}}}{1-R^2_{\alpha_s\sim\alpha_{s,-j}}}. \tag{21}$$

This allows us to rewrite the sensitivity parameter $1-R^2_{\alpha\sim\alpha_s}$ in terms of relative measure $k_{D,j}$ and the observed strength of X_j:

$$1-R^2_{\alpha\sim\alpha_s} = k_{D,j}\left(\frac{1-R^2_{\alpha_s\sim\alpha_{s,-j}}}{R^2_{\alpha_s\sim\alpha_{s,-j}}}\right). \tag{22}$$

In a partially linear model, the above reparameterization corresponds to the following result:

$$1-R^2_{\alpha\sim\alpha_s}=\eta^2_{D\sim U|X}=k_{D,j}\left(\frac{\eta^2_{D\sim X_j|X_{-j}}}{1-\eta^2_{D\sim X_j|X_{-j}}}\right). \tag{23}$$

The usefulness of equations (22) and (23) is that they allow researchers to bound the bias by making claims of relative importance of U, as compared to X_j. For example, setting $k_{D,j}\leq 1$ is equivalent to claiming that the additive gains in explanatory power due to latent confounders is no greater than the observed gains in explanatory power due to X_j.

E.3. **Relative bounds on $\eta^2_{Y\sim U|DX}$.** Here we follow a similar strategy as in the previous section. First note we can write $\eta^2_{Y\sim U|DX}$ as,

$$\eta^2_{Y\sim U|DX}=\frac{\eta^2_{Y\sim UX_j|DX_{-j}}-\eta^2_{Y\sim X_j|DX_{-j}}}{1-\eta^2_{Y\sim X_j|DX_{-j}}}. \tag{24}$$

Now define the measure of relative strength $k_{Y,j}$,

$$k_{Y,j}:=\frac{\eta^2_{Y\sim UX_j|DX_{-j}}-\eta^2_{Y\sim X_j|DX_{-j}}}{\eta^2_{Y\sim X_j|DX_{-j}}}. \tag{25}$$

Note $k_{Y,j}$ stands for how much variation is explained by adding U to the regression equation, as compared to the observed gains in explanatory power due to X_j. This allows us to rewrite $\eta^2_{Y\sim U|DX}$ as a function of the relative strength $k_{Y,j}$ and the observed strength of X_j, as in

$$\eta^2_{Y\sim U|DX}=k_{Y,j}\left(\frac{\eta^2_{Y\sim X_j|DX_{-j}}}{1-\eta^2_{Y\sim X_j|DX_{-j}}}\right). \tag{26}$$

E.4. **Benchmarking $|\rho|$.** The correlation ρ is not a measure of strength or explanatory power of the latent variables. Rather, it measures how much errors in the outcome equation are systematically related to errors in the Riesz representer. That is, for a confounder to create bias, this confounder not only needs to be strongly associated with the treatment and the outcome, but also the functional form of these associations needs to be "similar" in both equations, in order to create systematic biases. For instance, consider the (extreme) example discussed in the text with structural equations:

$$D=U^2 \tag{27}$$

$$Y = \theta D + U \tag{28}$$

where $U \sim N(0,1)$. Here, even though the latent variable U (nonparametrically) explains 100% of the residual variation in both the treatment and the outcome equations, the nonlinearity of the confounding model attenuates this bias, making it effectively zero, because U^2 is uncorrelated with U.

Therefore, plausibility judgments on the magnitude of $|\rho|$ will depend on how much we expect the functional form of the latent confounder in the treatment and outcome equations to be similar. In order to calibrate this judgment from empirical data, we propose using as a reference the observed Pearson's correlation of the outcome and RR errors induced by X_j, as given by ρ_j.

E.5. **Statistical inference of benchmark components.** We have the following measure of strength of association of the confounders with the outcome:

$$\eta^2_{Y\sim U|DX} = k_{Y,j}\left(\frac{\eta^2_{Y\sim X_j|DX_{-j}}}{1-\eta^2_{Y\sim X_j|DX_{-j}}}\right) =: k_{Y,j}G_{Y,j} \tag{29}$$

We also have the following measure of strength of association of the confounders with the RR:

$$1-R^2_{\alpha\sim\alpha_s} = k_{D,j}\left(\frac{1-R^2_{\alpha_s\sim\alpha_{s,-j}}}{R^2_{\alpha_s\sim\alpha_{s,-j}}}\right) =: k_{D,j}G_{D,j}. \tag{30}$$

The latter metric, in a partially linear model, corresponds to:

$$1-R^2_{\alpha\sim\alpha_s} = \eta^2_{D\sim U|X} = k_{D,j}\left(\frac{\eta^2_{D\sim X_j|X_{-j}}}{1-\eta^2_{D\sim X_j|X_{-j}}}\right).$$

We call the estimable components $G_{Y,j}$ and $G_{D,j}$ above the "gain" metrics. They measure gains in the explanatory power of observed covariates and, under the stated hypotheses of $k_{Y,j}$ and $k_{D,j}$, serve as proxies for the sensitivity parameters $\eta^2_{Y\sim U|DX}$ and $1-R^2_{\alpha\sim\alpha_s}$. These quantities also immediately pin-down $C_Y^2 = \eta^2_{Y\sim U|DX}$ and $C_D^2 = (1-R^2_{\alpha\sim\alpha_s})/R^2_{\alpha\sim\alpha_s}$ that enter the bias formulas. Since these components need to be estimated from the data, we use the following debiased representations which we now discuss.

Remark 8 (Debiased Representations)**.** We use the following Neyman orthogonal representations for the gain metrics,

$$G_{Y,j} = \frac{\sigma^2_{s,-j}}{\sigma^2_s} - 1, \qquad G_{D,j} = \frac{\nu^2_s}{\nu^2_{s,-j}} - 1,$$

where the variance of the residuals, $\sigma^2_s := \mathrm{E}(Y - g_s)^2$ and $\sigma^2_{s,-j} := \mathrm{E}(Y - g_{s,-j})^2$, are already in debiased form, and

$$\nu^2_s := 2\mathrm{E}m(W, \alpha_s) - \mathrm{E}\alpha^2_s \text{ and } \nu^2_{s,-j} := 2\mathrm{E}m(W, \alpha_{s,-j}) - \mathrm{E}\alpha^2_{s,-j}$$

are the debiased forms for $\mathrm{E}\alpha^2_s$ and $\mathrm{E}\alpha^2_{s,-j}$.

As for ρ_j, we first define the debiased form of the change in estimates,

$$\Delta\theta_{s,j} = \mathrm{E}m(W, g_s) + \mathrm{E}(Y - g_s)\alpha_s - \mathrm{E}m(W, g_{s,-j}) - \mathrm{E}(Y - g_{s,-j})\alpha_{s,-j}.$$

This gives us the debiased representation for the correlation,

$$\rho_j = \frac{\Delta\theta_{s,j}}{\sqrt{\sigma^2_{s,-j} - \sigma^2_s}\sqrt{\nu^2_s - \nu^2_{s,-j}}}.$$

The debiasedness (Neyman orthogonality) of the above expressions follows from the chain rule for functional calculus (e.g., van der Vaart and Wellner (1996)), exploiting the fact that each representation is a smooth transformation of debiased representations. Note the components of $G_{D,j}$, $G_{Y,j}$, and ρ_j are the same as those already covered by Lemma 3, with the minor change of dropping covariate X_j where appropriate. Given DML estimators of the components, we then obtain the following influence functions for the plug-in estimators of $G_{D,j}$, $G_{Y,j}$, and ρ_j:

$$\psi^o_{G_{Y,j}}(Z) = \frac{\sigma^2_s \psi^o_{\sigma^2_{s,-j}}(Z) - \sigma^2_{s,-j}\psi^o_{\sigma^2_s}(Z)}{\sigma^4_s} \qquad \psi^o_{G_{D,j}}(Z) = \frac{\nu^2_{s,-j}\psi^o_{\nu^2_s}(Z) - \nu^2_s \psi^o_{\nu^2_{s,-j}}(Z)}{\nu^4_{s,-j}}$$

$$\begin{aligned}\psi^o_{\rho_j}(Z) = {} & \frac{\psi^o_{\theta_s}(Z) - \psi^o_{\theta_{s,-j}}(Z)}{(\sigma^2_{s,-j} - \sigma^2_s)^{1/2}(\nu^2_s - \nu^2_{s,-j})^{1/2}} - \frac{\Delta\theta_{s,j}(\psi^o_{\sigma^2_{s,-j}}(Z) - \psi^o_{\sigma^2_s}(Z))}{2(\sigma^2_{s,-j} - \sigma^2_s)^{3/2}(\nu^2_s - \nu^2_{s,-j})^{1/2}} \\ & - \frac{\Delta\theta_{s,j}(\psi^o_{\nu^2_s}(Z) - \psi^o_{\nu^2_{s,-j}}(Z))}{2(\sigma^2_{s,-j} - \sigma^2_s)^{1/2}(\nu^2 - \nu^2_{-j})^{3/2}}.\end{aligned} \tag{31}$$

E.6. **Statistical inference for the bounds.** Let $\mathrm{BF} = \rho C_Y C_D$ denote the "bias factor" of the bounds, that is, the part of the bounds that depends on quantities related to unobserved confounder U. When we perform benchmarking, this bias factor is estimated based on data, and the new plug-in estimator of the bounds is given by:

$$\widehat{\theta}_{\pm} := \widehat{\theta}_s \pm \underbrace{\left(\widehat{\rho}_j \times \sqrt{k_{Y,j}\widehat{G}_{Y,j} \times \frac{k_{D,j}\widehat{G}_{D,j}}{1 - k_{D,j}\widehat{G}_{D,j}}} \right)}_{\widehat{\mathrm{BF}}} \times \widehat{S} = \widehat{\theta}_s \pm \widehat{\mathrm{BF}} \times \widehat{S}$$

where the components $\widehat{\rho}_j$, $\widehat{G}_{D,j}$ and $\widehat{G}_{Y,j}$ are estimated using the DML algorithm and the debiased representations given above. Estimation errors on the estimable components of BF can thus be easily propagated via the delta method.

In particular, given the product form of the bias, $\widehat{\mathrm{BF}} \times \widehat{S}$, the adjustment in the influence function of Theorem 4 takes a simple form—it suffices to add an extra term due to the estimation error of BF, scaled by S. That is,

$$\varphi^o_{\pm}(Z) = \underbrace{\psi^o_{\theta_s}(Z) \pm \mathrm{BF} \times \psi^o_S(Z)}_{\text{Influence function with fixed BF}} \pm \underbrace{\psi^o_{\mathrm{BF}}(Z) \times S}_{\text{Addition due to estimated BF}},$$

where here for simplicity we denote by $\psi^o_S(Z)$ the influence function of the scaling factor S as described in the main text,

$$\psi^o_S(Z) := \frac{1}{2S}\left(\sigma^2_s \psi^o_{\nu^2_s}(Z) + \nu^2_s \psi^o_{\sigma^2_s}(Z)\right),$$

and $\psi^o_{\mathrm{BF}}(Z)$ is the influence function of the bias factor BF. This last term can be computed using the previous results:

$$\begin{aligned} \psi^o_{\mathrm{BF}}(Z) := &\left(\frac{k_{D,j}G_{D,j}}{1 - k_{D,j}G_{D,j}}\right)^{1/2} \times (k_{Y,j}G_{Y,j})^{1/2} \times \psi^o_{\rho_j}(Z) \\ &+ \rho_j \times \left(\frac{k_{D,j}G_{D,j}}{1 - k_{D,j}G_{D,j}}\right)^{1/2} \times \frac{k_{Y,j}\psi^o_{G_{Y,j}}(Z)}{2(k_{Y,j}G_{Y,j})^{1/2}} \\ &+ \rho_j \times (k_{Y,j}G_{Y,j})^{1/2} \times \left(\frac{k_{D,j}\psi^o_{G_{D,j}}(Z)}{2(k_{D,j}G_{D,j})^{1/2}(1 - k_{D,j}G_{D,j})^{3/2}}\right). \end{aligned} \tag{32}$$

	Gain Metrics		Correlation	Change in estimate
Observed covariate	$G_{Y,j}$	$G_{D,j}$	ρ_j	$\Delta\widehat{\theta}_{s,j}$
inc	0.145	0.047	0.34	3,349
pira	0.038	0.003	0.21	188
twoearn	0.021	0.007	-0.25	-621

TABLE 3. Explanatory power of observed covariates in the Partially Linear Model. All estimates are debiased and cross-fitted.

For simplicity, we do not propagate uncertainty of these metrics into the bounds in the main text.

E.7. **Empirical Benchmarking Results for 401(k) Example.** Using the formulas described above, we obtain the following empirical results for the 401(k) example. Table 3 shows the results for the partially linear model and Table 4 shows the results for the nonparametric model.[18] These gain metrics are the ones used in the contour plots of the main text.

	Gain Metrics		Correlation	Change in estimate
Observed covariate	$G_{Y,j}$	$G_{D,j}$	ρ_j	$\Delta\widehat{\theta}_{s,j}$
inc	0.129	0.143	0.23	3,767
pira	0.032	0.006	0.19	449
twoearn	0.015	0.011	-0.07	-661

TABLE 4. Explanatory power of observed covariates in NPM Model. All estimates are debiased and cross-fitted.

[18] All metrics are estimated using the same procedure described in footnote 10.

Appendix F. Deferred Empirical Example: price elasticity of gasoline demand

F.1. **Estimates under conditional ignorability.** An important part of estimating the welfare consequences of price changes is to identify the price elasticity of demand. Here we re-analyze the data on gasoline demand from the 2001 *National Household Travel Survey* (NHTS) (Blundell et al., 2012, 2017; Chetverikov and Wilhelm, 2017). This is a household level survey conducted by telephone and complemented by travel diaries and odometer readings (see Blundell et al. (2012) and ORNL for details). Important variables in the survey include household income, gasoline price, and annual gasoline consumption (as inferred by odometer readings and fuel efficiency of vehicles). Income data corresponds to the median of the income bracket of the household, with 15 income brackets equally spaced apart in the logarithmic scale. The survey also contains 24 covariates related to population density, urbanization, demographics and US Census region indicators.[19]

	Short Results		Robustness Values	
Model	Short Estimate	Std. Error	$\mathrm{RV}_{\theta=-1.5,a=0.05}$	$\mathrm{RV}_{\theta=0,a=.05}$
Partially linear	-0.701	0.257	0.026	0.019
nonparametric	-0.761	0.360	0.010	0.011

Note: $\rho^2 = 1$; Significance level of 5%. Standard errors in parenthesis.

Table 5. Minimal sensitivity reporting, gasoline demand.

Under the assumption of conditional ignorability, we estimate the average causal derivative of log price on log demand, adjusting for the 24 observed covariates.[20] We consider both a partially linear

[19] The data is available on the npiv STATA package (Chetverikov et al., 2018). The full data contains $3,640$ observations. After applying the same filters suggested by Blundell et al. (2017) and Chetverikov et al. (2018), the final data contains $3,466$ observations.

[20] This can be interpreted as the average price elasticity of demand. We approximate the derivative numerically using a finite difference (e.g., $f'(x) \approx (f(x+0.01) - f(x-0.01))/0.01$).

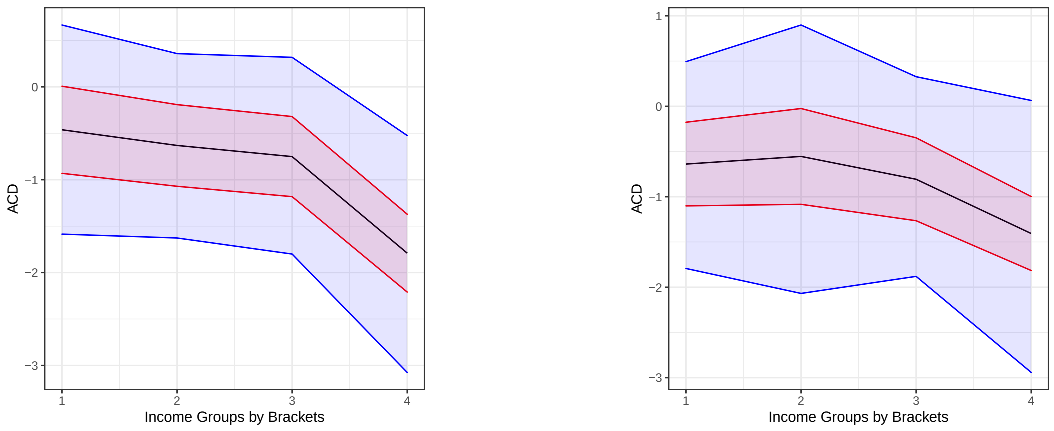
model, and a fully nonparametric model[21]. The results are shown in the first column of Table 5. In both models, we obtain estimates similar to the ones obtained in prior literature, with an estimated price elasticity of approximately -0.7.

F.2. **Sensitivity analysis.** Despite having a large number of control variables, there are several reasons why one should worry about the assumption of no unobserved confounders in this setting. For instance, as was argued in Blundell et al. (2017), prices vary at the local market level, and unobserved factors that affect consumer preferences could act as unobserved confounders. Another potential source of endogeneity is the fact that we only observe the median of the income bracket of each household, and not the actual income. Since these brackets correspond to large income intervals, the remnant variation in the true income could be another major source of unobserved confounding. This is exacerbated in the larger income brackets, which correspond to larger intervals (and explains the reason why these larger income brackets were not included in prior work).[22]

[21] For the partially linear specification we use DML with a cross-validated generic machine learning regression to residualize the outcome and the treatment. For the fully nonparametric specification, we use a generic machine learning approach to estimate both the regression function and the Riesz Representer. In both cases, the regression estimator uses $5-$fold cross-validation to select the best among: (i) lasso models with feature expansions; (ii) random forests; and, (iii) local polynomial forests. The Riesz representer is estimated based on the loss outlined in Remark 6. We again use 5-fold cross-validation to choose the best model among a penalized linear Riesz representation with expanded features and a combination of ℓ_1 and ℓ_2 penalty (Chernozhukov et al., 2021, 2022c), and a random forest representation (ForestRiesz) (Chernozhukov et al., 2022b). In both analyses, in order to reduce the variance that stems from sample splitting for cross-validation and for cross-fitting, we repeat the experiment for 5 random partitions of the data and average the final estimate, incorporating variation across experiments into the standard error, as described in Chernozhukov et al. (2018a). Moreover, since samples are highly correlated within states, we perform grouped cross-validation, where samples of the same state are always in the same fold and we stratify the folds by the census region variable.

[22] Prior work has also analyzed this data via instrumental variable (IV) approaches (Blundell et al., 2017; Chetverikov and Wilhelm, 2017), using the distance to the closest major oil platform as an instrument. They find that IV estimates are close to the ones based on unconfoundedness (Chetverikov and Wilhelm, 2017). Further, note that the above

We thus applied our sensitivity analysis tools to assess the sensitivity of the previous estimates to unobserved confounding.



(A) Partially Linear Model.

(B) Nonparametric Model.

FIGURE 4. One-Sided Confidence Bounds for the ACD by Income Brackets.

Note: Estimate (black), bounds (red), and confidence bounds (blue) for the ACD. Confounding scenario: $\rho^2 = 1$; $C_Y^2 = 0.03$; $C_D^2 \approx 0.03$. Significance level of 5%.

The second part of Table 5 reports the robustness values for price elasticity, such that the sensitivity bounds would contain a target value θ. Here we consider $\theta = -1.5$ (very elastic) and $\theta = 0$ (perfectly inelastic). We find that, at the 5% confidence level, these robustness values are at around 2% (PLM) and 1% (NPM). These results show that, unless researchers are able to rule out confounding that explains at about 2% of the residual variation of gasoline price and gasoline consumption, the evidence provided by the data is not strong enough to distinguish between extremes such as a "very elastic," or a "perfectly inelastic" demand function. To put this number in context, our coarse measure of income (median of the income bracket) explains around 15% of the residual variation of gasoline price and 7% of the residual variation of gasoline demand. It is thus not implausible that remnant variation in the true income could overturn these results.

threats to conditional ignorability are also credible threats to the validity of this proposed instrument. Extensions of our sensitivity results to IV is left to future work.

Finally, we explore how price elasticity varies with income under a specific confounding scenario. We consider three overlapping income groups defined as observations with income within $\pm .5$ in log-scale around the income points \$$42,500$, \$$57,500$ and \$$72,500$, as well as a fourth high income group of all units with income above 11.6 on the log scale ($\approx$ \$$110,000$). To illustrate, we consider a confounding scenario of approximately 3% for both sensitivity parameters, and repeat our nonparametric and partially linear estimation and sensitivity analysis for each sub-group. Point-estimates, bounds and confidence bounds are reported in Figure 4. Note that, under this scenario, the evidence for effect heterogeneity is substantially weakened, especially when using a fully nonparametric model.