# On Double Robustness in Double Machine Learning

**Simon Valentin**[1] **Gianluca Detommaso**[1,*] **Yikuan Li**[2]

**Manfred Opper**[2,*] **Michael Brückner**[2]
[1]AWS AI Labs [2]Amazon
{simval, yikuanli, brueckm}@amazon.com detommaso.gianluca@gmail.com
manfred.opper@tu-berlin.de

## Abstract

Double Machine Learning (DML) is widely used for causal estimation from observational data and is often assumed to be doubly robust. While this holds for the Z-estimator proposed by Chernozhukov et al., many practical implementations rely on the Robinson estimator, which crucially depends on correct treatment model specification. This misunderstanding has important implications as many practitioners incorrectly assume robustness to misspecification. We provide analyses clarifying when double robustness holds for popular DML estimators. Based on these insights, we develop a maximum likelihood estimator that achieves double robustness, providing a likelihood-based alternative to the Z-estimator.

## 1 Introduction

Treatment effect estimation from observational data is a fundamental challenge in science, with applications ranging from public health to social science and economics. Double Machine Learning (DML) [Chernozhukov et al., 2018] is one standard tool for such causal inference tasks combining flexible machine learning methods with classical parametric estimation to enable valid statistical inference in a semi-parametric setting. Popular software packages like DoubleML [Bach et al., 2022, 2024] and EconML [Battocchi et al., 2019] implement DML as one of their default estimators.

Double robustness, which predates DML, was introduced in the context of missing data imputation by Scharfstein et al. [1999], who showed that certain estimators remain consistent when either the propensity score or the outcome model is correctly specified. Comprehensive overviews of doubly robust methods in both missing data and causal inference contexts are provided by Bang and Robins [2005] and Tsiatis [2007]. Since then, doubly robust principles have been extended to more complex settings, particularly for continuous treatments Kennedy et al. [2017], Colangelo and Lee [2020].

However, a crucial misunderstanding has emerged in both theory and practice. DML is frequently described and implemented under the assumption that it possesses double robustness properties, that is, it converges to the true parameter when either the treatment or outcome model is correctly specified. This belief is, however, only correct for the original Z-estimator [Chernozhukov et al., 2018], but not the commonly implemented variant based on maximum likelihood estimation by Robinson [1988]. The misconception might be due to a misunderstanding of the implications of Neyman orthogonality (see Appendix A for a discussion). This finding has important implications for practice. As commonly applied, DML is not doubly robust – even when the model predicting the outcome from the confounders is perfectly specified, the unbiasedness of the causal treatment effect rests on the correctness of the treatment model.

[*] Work done while at Amazon.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CauScien.

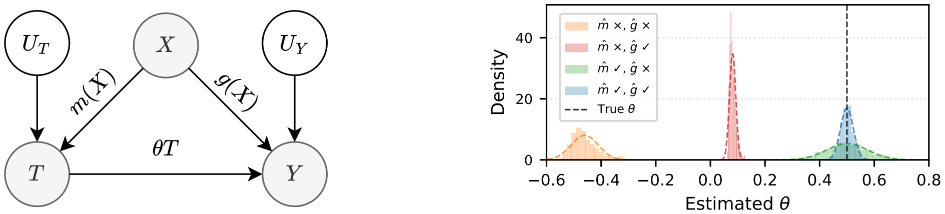

Figure 1: Left: Causal diagram for the partially linear model. Right: Sampling distributions of DML estimates under various model specifications. When the treatment model is misspecified (orange and red), estimates are biased regardless of the outcome model specification. Well-specified treatment models (green and blue) yield consistent estimation around the true value $\theta = 0.5$ (dashed line). Legend indicates whether the nuisance parameter models $\hat{m}$ and $\hat{g}$ are misspecified (×) or well-specified (✓). See Appendix B for details.

The remainder of the paper is organized as follows. Section 2 introduces the problem setting. Section 3, 4, and 5 analyzes standard DML and proves its lack of double robustness in the case of the likelihood-based estimator, then goes on to show how we can achieve double robustness. Section 6 discusses implications for theory and practice, as well as future directions for this line of work.

## 2 Problem Setup

We consider estimating the causal treatment effect from observational data where we assume that some treatment $T \in \mathcal{T} \subseteq \mathbb{R}$ affects outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$, and both are causally affected by confounding variable $X \in \mathcal{X}$ from a potentially high-dimensional space.

Let $(x_i, t_i, y_i)_{i=1}^n$ be $n$ independent observations generated according to the following partially linear structural causal model (see also Figure 1):

$$Y = \theta T + g(X) + U_Y, \quad T = m(X) + U_T, \tag{1}$$

where (1) outcome $Y$ depends on covariates $X$ through the unknown function $g$ and on treatment $T$ through the parameter $\theta$, (2) treatment $T$ depends on $X$ through the unknown function $m$, and where $T$ can be continuous ($T \in \mathbb{R}$) or binary ($T \in \{0, 1\}$), (3) error terms $U_Y$ and $U_T$ have mean zero and are independent of each other and of $X$. Under this model, variables $Y$ and $T$ are endogenous random variables deterministically derived from the exogenous random variables $X$, $U_Y$, and $U_T$.

For the special case of binary treatments, $\mathcal{T} := \{0, 1\}$, the parameter of interest, $\theta$, equals the average treatment effect Imbens and Rubin [2015], which can be expressed either through potential outcomes as $\mathbb{E}[Y(1) - Y(0)]$ or using the do-operator as $\mathbb{E}[Y \mid \mathrm{do}(T = 1)] - \mathbb{E}[Y \mid \mathrm{do}(T = 0)]$ Pearl [2009]. For continuous treatments, $\theta$ represents the average partial effect Rothenhäusler and Yu [2019], expressed as $\mathbb{E}[\partial_t Y(t)]$ or equivalently $\mathbb{E}[\partial_t Y \mid \mathrm{do}(T = t)]$. Both capture how the expected outcome changes in response to a change in the treatment level.

The identification of $\theta$ as the causal effect $\mathbb{E}[\partial_t Y(t)]$ (or $\mathbb{E}[Y(1) - Y(0)]$ Imbens and Rubin [2015] for binary treatments) relies on three standard assumptions: (1) exogenous zero-mean noise, (2) unconfoundedness, and (3) overlap. See Appendix C for further details.

## 3 Double Robustness

A key question in causal inference is whether estimators remain valid when either the treatment or the outcome model is biased. This property, known as double robustness, provides protection against model misspecification Tsiatis [2007]. For the binary treatment case, augmented inverse propensity weighting (AIPW) is a classical approach to achieving double robustness (see Appendix D). Here, we focus on the more general and challenging setting of continuous treatments. We refer to double robustness as follows:

**Definition 3.1** (Double Robustness)**.**

$$\mathbb{E}[(\hat{\theta}_n - \theta)^2] \xrightarrow{n} 0, \text{ if either } \mathbb{E}[(\hat{m}_n(x) - m(x))^2] \xrightarrow{n} 0, \text{ or } \mathbb{E}[(\hat{g}_n(x) - g(x))^2] \xrightarrow{n} 0.$$

By this definition, an estimator $\hat{\theta}_n$ is doubly robust if its mean squared error converges to zero when either the treatment model converges in mean square error: $\mathbb{E}[(\hat{m}_n(x) - m(x))^2] \xrightarrow{n} 0$, or the outcome model converges in mean square error: $\mathbb{E}[(\hat{g}_n(x) - g(x))^2] \xrightarrow{n} 0$. This ensures that both, the bias and variance of the estimator, vanish as $n \to \infty$ if at least one of the models is well-specified and converges.

## 4 Unpacking Double Machine Learning

Double Machine Learning (DML) Chernozhukov et al. [2018] is a two-stage approach that allows for flexible machine learning estimation of nuisance functions while maintaining valid inference for the causal parameter. The two stages are explained in the following.

**First Stage: Nuisance Functions Estimation.** The first stage estimates two nuisance functions:

$$\hat{m}(x) \approx m(x) = \mathbb{E}[T \mid X = x] \quad \hat{\ell}(x) \approx \ell(x) := \mathbb{E}[Y \mid X = x]. \tag{2}$$

In DML, these respective estimators for the nuisance functions are obtained via arbitrary machine learning methods and subsequently used in a plug-in fashion.

**Second Stage: Causal Effect Estimation.** The causal effect is estimated by regressing outcome residuals on treatment residuals. Two different estimators are commonly used, both discussed in the original work on DML Chernozhukov et al. [2018] and implemented in popular software packages, as mentioned above. We present the the *partialing-out* estimator Robinson [1988] and Z-estimator as below respectively,

$$\hat{\theta}_{\mathrm{R}} = \frac{\sum_{i=1}^{n}(y_i - \hat{\ell}(x_i))(t_i - \hat{m}(x_i))}{\sum_{i=1}^{n}(t_i - \hat{m}(x_i))^2}, \quad \hat{\theta}_{\mathrm{Z}} = \frac{\sum_{i=1}^{n}(y_i - \hat{g}(x_i))(t_i - \hat{m}(x_i))}{\sum_{i=1}^{n}(t_i - \hat{m}(x_i))t_i}, \tag{3}$$

where $\hat{g}(x)$ is an estimate of $g(x)$ in (1), which is typically approximated with $\hat{\ell}(x)$.

### 4.1 Where Does the Bias Come From?

The choice of approximating $g(x)$ with $\hat{\ell}(x)$ is problematic, as $\ell(x)$ systematically differs from $g(x)$ in the structural equation model. To see the disconnect, let us enter (1) in (2). Under Assumption C.1, we have

$$\ell(x) = \mathbb{E}[Y \mid X = x] = \mathbb{E}[\theta(m(X) + U_T) + g(X) + U_Y] = \theta m(x) + g(x) \neq g(x). \tag{4}$$

Function $l$ captures the *total* effect of the confounders $X$ on the outcome $Y$, whereas $g$ only explains the *direct* effect of $X$ on $Y$. Hence, setting $\hat{g}(x) := \hat{\ell}(x)$ introduces a systematic bias, which has significant implications for the estimator's robustness properties.

In practice, these estimators are often used interchangeably or in hybrid forms, creating confusion about their respective robustness properties. In the following sections, we analyze how first-stage estimation affects the downstream estimates and clarify the conditions under which each second-stage estimator achieves consistency. For further explanation from a graphical perspective, see Appendix E.

## 5 Robustness Analysis

### 5.1 Analyzing $\hat{\theta}_{\mathrm{R}}$

To analyze the robustness properties of the Robinson estimator as commonly implemented in DML, we examine its mean squared error (MSE). Consider the expected prediction error:

$$\mathcal{L}(\hat{\theta}) := \mathbb{E}[(Y - \hat{Y})^2]. \tag{5}$$

Solving for the estimation error $\hat{\theta} - \theta$ gives (see Appendix F for derivation):

$$\hat{\theta}_{\mathrm{R}} - \theta = \frac{\mathbb{E}[\hat{m}(X)(m(X) - \hat{m}(X))]}{\mathbb{E}[(m(X) - \hat{m}(X))^2] + \sigma_T^2}\theta + \frac{\mathbb{E}[(g(X) - \hat{\ell}(X))(m(X) - \hat{m}(X))]}{\mathbb{E}[(m(X) - \hat{m}(X))^2] + \sigma_T^2}. \tag{6}$$

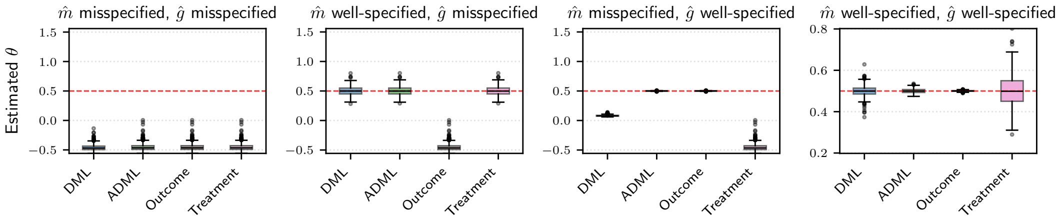

Figure 2: Estimator comparison under different model specifications. Results from 1,000 simulations ($n = 10,000$) with exponential treatment and cubic outcome equations. True effect $\theta = 0.5$ (dashed red line). Here, DML indicates the standard implementation estimating $g$ as $\ell$. See Appendix H for details and Appendix I for additional experiments

This result reveals two potential sources of bias. The first term represents bias from treatment model misspecification, while the second term captures the interaction between treatment and outcome model errors. For the estimator to be consistent, both terms must vanish as the sample size increases.

When $\hat{m}(X) \xrightarrow{n} m(X)$ (correctly specified treatment model), both terms converge to zero regardless of the outcome model specification. However, when only $\hat{\ell}(X) \xrightarrow{n} \ell(X)$, the second term retains the bias through $\ell$, since $\ell(X) = g(X) + \theta m(X)$, even with a perfectly estimated $\hat{\ell}(X)$. This bias does not vanish unless $\hat{m}(X)$ also converges to $m(X)$, and so the robustness of the estimator depends solely on the correctness of $\hat{m}(X)$.

### 5.2 Analyzing $\hat{\theta}_{\mathrm{Z}}$

Similarly, solving for the estimation errors gives ,

$$\hat{\theta}_{\mathrm{Z}} - \hat{\theta} = \frac{\mathbb{E}[(T\theta + U_Y)(T - \hat{m}(X))]}{\mathbb{E}[(T - \hat{m}(X))T]} = \frac{\mathbb{E}[(T - \hat{m}(X))T]}{\mathbb{E}[(T - \hat{m}(X))T]}\theta + \frac{\mathbb{E}[(m(X) - \hat{m}(X) + U_T)U_Y]}{\mathbb{E}[(T - \hat{m}(X))T]}$$

where, again, the first term equals $\theta$ and the second term vanishes as long as the outcome noise $U_Y$ is uncorrelated with estimation error, $\hat{m}(X) - m(X)$. Unlike $\hat{\theta}_{\mathrm{R}}$, estimator $\hat{\theta}_{\mathrm{Z}}$ achieves double robustness; though, only if we estimate $g(X)$ rather than $\ell(X)$; the direct effect of $X$ on $Y$.

### 5.3 Augmented Double Machine Learning

Our analysis shows that DML's consistency depends critically on the correct specification of the treatment model. We now develop a modified likelihood-based estimator that achieves double robustness.

We introduce the Augmented DML (ADML) estimator that achieves double robustness through a modified model structure:

$$\hat{Y} = (T - \hat{m}(X))\hat{\theta} + \hat{m}(X)\hat{\phi} + \hat{g}(X), \tag{7}$$

where $\hat{\phi}$ is a nuisance parameter that adjusts for potential misspecification in $\hat{m}(X)$. When minimizing the squared error between $Y$ and $\hat{Y}$, that is, maximizing the Gaussian likelihood of the model parameters, this formulation leads to the estimator (see Appendix G for details):

$$\hat{\theta}_{\mathrm{ADML}} = \frac{\sum_{i=1}^{n}(y_i - \hat{g}(x_i))(t_i - \tau\hat{m}(x_i))}{\sum_{i=1}^{n}(t_i - \tau\hat{m}(x_i))t_i}, \quad \tau = \frac{\sum_{i=1}^{n}\hat{m}(x_i)t_i}{\sum_{i=1}^{n}\hat{m}^2(x_i)}. \tag{8}$$

The parameter $\tau$ measures the alignment between $\hat{m}(X)$ and $T$, adaptively determining the degree of treatment residualization. In the special case where $\hat{m}(X) \xrightarrow{n} m(X)$, we have $\tau \xrightarrow{n} 1$ and ADML reduces to the standard DML estimator. In contrast, when $\hat{m}(X)$ is completely misspecified so that $\mathbb{E}[T\hat{m}(X)] = 0$, we have $\tau \xrightarrow{n} 0$ and the estimator relies only on the outcome model. Thus, ADML achieves consistency if either the treatment model or the outcome model is correctly specified. We illustrate this through simulations with non-linear confounding, see Figure 2.

# 6 Conclusion

Our analysis clarifies a fundamental misconception in DML: contrary to common belief, the widely implemented Robinson estimator is not doubly robust but critically depends on correctly specifying the treatment model. This has significant implications for practitioners who may incorrectly assume protection against model misspecification. We have demonstrated why this bias occurs and introduced Augmented DML (ADML), which achieves double robustness while maintaining a likelihood-based estimation framework. Moving forward, implementations should either adopt the $Z$-estimator, ensure treatment models are correctly specified via least squares estimation, or implement our proposed ADML estimator to guarantee consistency when either model is well-specified.

# References

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1–68, 2018.

Philipp Bach, Victor Chernozhukov, Malte S. Kurz, and Martin Spindler. DoubleML – An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

Philipp Bach, Malte S. Kurz, Victor Chernozhukov, Martin Spindler, and Sven Klaassen. DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3):1–56, 2024.

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML, 2019.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94 (448):1096–1120, 1999.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

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

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

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

Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2025.

Russell Davidson, James G MacKinnon, et al. *Estimation and inference in econometrics*, volume 63. Oxford New York, 1993.

## A Neyman Orthogonality and Convergence Properties

DML's theoretical properties are often misunderstood as implying double robustness. However, a careful examination of Neyman orthogonality reveals why correct specification of the treatment model remains crucial despite DML's apparent robustness to estimation error.

Neyman orthogonality is a local property concerning how estimation errors in nuisance functions affect the target parameter. A moment condition $\psi(w; \theta, \eta)$ is Neyman orthogonal if:

$$\partial_\eta \mathbb{E}[\psi(W; \theta, \eta)][\hat{\eta} - \eta] = 0 \tag{9}$$

where $\eta$ represents the true nuisance functions $m$ and $g$. This property ensures that small deviations from the true nuisance functions have no first-order effect on the estimation of $\theta$.

In our partially linear model, DML's moment condition takes the form:

$$\mathbb{E}[(Y - \ell(X) - \theta T)(T - m(X))] = 0 \tag{10}$$

This condition exhibits Neyman orthogonality, meaning that if both $\hat{\ell}(x)$ and $\hat{m}(x)$ are "close enough" to their true values, estimation errors have minimal impact on $\hat{\theta}_n$. However, this local robustness is fundamentally different from double robustness: Neyman orthogonality only provides protection against small deviations around the true nuisance functions. If $\hat{m}(x)$ is systematically misspecified and converges to something other than $m(x)$, the resulting bias in $\hat{\theta}_n$ can be substantial.

This explains why DML can handle noisy estimation of correctly specified models but not fundamental misspecification of the treatment mechanism. The method's robustness is local rather than global – it provides protection against estimation error but not model misspecification.

## B Empirical Illustration

To demonstrate the importance of correct treatment model specification in DML, we conduct a simulation study with non-linear confounding. We consider a data generating process where the treatment equation is exponential and the outcome equation is cubic:

$$T = m_0 + m_1 \exp(m_2 X) + U_T \tag{11}$$

$$Y = g_0 X + g_1 X^2 + g_2 X^3 + \theta T + U_Y \tag{12}$$

where $U_T, U_Y$ are independent standard normal errors, $\theta = 0.5$, and the confounder $X$ follows a bimodal distribution combining $\mathcal{N}(-2, 1)$ and $\mathcal{N}(2, 1)$.

We implement DML with four specifications varying in model correctness. Results are obtained across 1,000 simulation runs ($n = 10,000$). When $m(x)$ is misspecified, the estimator exhibits substantial bias regardless of $g(x)$ specification. With correct $m(x)$, the estimator centers on $\theta = 0.5$, achieving better precision when $g(x)$ is also correct.

## C Assumptions

We make the following standard assumptions, where $g$ and $m$ are nuisance functions whose estimation is not of direct interest but is necessary for identifying the causal effect $\theta$:

**Assumption C.1** (Exogenous zero-mean noise)**.** The error terms satisfy:

$$\mathbb{E}[U_T|X] = 0 \text{ and } \mathbb{E}[U_Y|X, T] = 0 \tag{13}$$

**Assumption C.2** (Unconfoundedness)**.** Potential outcomes are independent of treatment assignment conditional on covariates:

$$Y(t) \perp T \mid X \text{ for all } t \in \mathcal{T} \tag{14}$$

**Assumption C.3** (Overlap)**.** For all $x \in \mathcal{X}$ and all $t \in \mathcal{T}$:

$$p_{T|X}(t|x) > 0 \tag{15}$$

# D Augmented Inverse Probability Weighting

To understand how to achieve double robustness, it is instructive to examine the Augmented Inverse Probability Weighting (AIPW) estimator for binary treatments. Let $t \in \mathcal{T} := \{0, 1\}$ denote treatment values and $\hat{m}(x) = p_{T|X}(1|x)$ denote the propensity score. The function $\hat{h}(t, x)$ used in the AIPW estimator estimates the conditional expectation $\mathbb{E}[Y|T = t, X = x]$. Under the partially linear model, this equals $\theta t + g(x)$. The AIPW estimator combines outcome modeling with inverse probability weighting:

$$\hat{\theta}_{\text{AIPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{h}(1, x_i) - \hat{h}(0, x_i) + \frac{t_i(y_i - \hat{h}(1, x_i))}{\hat{m}(x_i)} - \frac{(1 - t_i)(y_i - \hat{h}(0, x_i))}{1 - \hat{m}(x_i)} \right] \tag{16}$$

The expected estimation error, $\mathbb{E}[\theta - \hat{\theta}_{\text{AIPW}}]$, can be written as

$$\mathbb{E}\left[ \frac{(\hat{m}(X) - m(X))(\hat{h}(1, X) - h(1, X))}{m(X)} \right] + \mathbb{E}\left[ \frac{(\hat{m}(X) - m(X))(\hat{h}(0, X) - h(0, X))}{1 - m(X)} \right] . \tag{17}$$

As the individual estimation errors appear as products, it follows that both terms vanish if either of the estimators, $\hat{h}$ and $\hat{m}$, converges to the true function. We refer to Murphy [2025] (Section 36.4.2.3) for a more detailed discussion. However, the AIPW estimator is designed for binary treatments. For continuous treatments, inverse probability weights become ill-defined as one cannot divide by the probability of observing a certain treatment value, which is zero.

# E Graphical Model Perspective

Consider the causal graph in Figure 1 (left). The challenge in identifying the causal effect $\theta$ arises from confounding: $X$ affects both treatment and outcome, creating spurious correlation between $T$ and $Y$. Any observed association between $T$ and $Y$ combines both the causal effect we want to estimate ($T \to Y$) and the spurious correlation through $X$, ($T \leftarrow X \to Y$).

DML addresses this through two residualization steps. The treatment residual, $U_T = T - \hat{m}(X)$, serves as an instrument, while the outcome residual, $U_Y = Y - \hat{\ell}(X)$, isolates the outcome noise by removing confounding information from the outcome. When $\hat{m}$ converges to $m$, the treatment residual $U_T$ becomes a perfect instrument, that is, it becomes independent of the confounder, $U_T \perp\!\!\!\perp X$, and it preserves the variation necessary for identifying $\theta$. In this case, the causal effect from $U_T$ on $Y$ is identical to the causal effect from $T$ on $Y$. However, the first inference problem is simpler as $U_T$ is by construction unconfounded and we not required to condition on $X$ (or $U_Y$) to remove spurious correlation. Still, it is advisable to do so to improve estimation efficiency, as further conditioning on $U_Y$ is expected to reduce the variance of the estimation error. As $U_Y$ can be viewed as a means to improve estimation efficiency rather than to remove causal bias, we are not required to obtain an unbiased estimate of $g$ in (1). Hence, in practice, we often estimate,

$$\ell(x) \; := \; \mathbb{E}[Y|X = x] \; = \; \underbrace{\theta m(x)}_{\text{indirect effect}} \; + \; \underbrace{g(x)}_{\text{direct effect}} \; , \tag{18}$$

which models the total effect of confounders $X$ on outcome $Y$, whereas $g$ in (1) represents only the direct effect, $X \to Y$. As long the treatment model is correctly specified, this procedure is valid as identification holds regardless of how well we estimate $g$ or whether we consider $X$ at all. However, when the treatment model is misspecified, the residual retains dependence on $X$, preventing identification when not controlling for confounder $X$ and $U_Y$, respectively. In this case, a correct specification and unbiased estimation of $g$ is crucial.

**Remark:** The asymmetric role of treatment and outcome models mirrors the classical Frisch-Waugh-Lovell (FWL) theorem from linear regression [Davidson et al., 1993]. Just as FWL requires correct specification of linear projections on controls, a correct specification of $m$ is sufficient for identification in DML.

# F Derivation of DML Robustness

Let us define the expected squared error as a function of $\hat{\theta}$, given plugin estimators of $m$ and $g$ denoted by $\hat{m}$ and $\hat{g}$, respectively:

$$\mathcal{L}(\hat{\theta}) := \mathbb{E}[(Y - \hat{Y})^2],$$

with

$$Y = (m(X) + U_T)\theta + g(X) + U_Y \quad \text{and} \quad \hat{Y} = (m(X) + U_T - \hat{m}(X))\hat{\theta} + \hat{\ell}(X) + U_Y,$$

where the true outcome follows the partially linear model in Equation 1, and the predicted outcome uses the partialing-out approach. This gives

$$\mathcal{L}(\hat{\theta}) = \mathbb{E}\big[\big((m(X) + U_T)\theta + g(X) - (m(X) + U_T - \hat{m}(X))\hat{\theta} - \hat{\ell}(X)\big)^2\big],$$

where the expectation is over exogenous random variables $X$ and $U_T$. Setting the derivative of $\mathcal{L}$ with respect to $\hat{\theta}$ to zero gives,

$$\begin{aligned}
0 &= \mathbb{E}\big[((m(X) + U_T)\theta + g(X) - (m(X) + U_T - \hat{m}(X))\hat{\theta} - \hat{\ell}(X)\big)(m(X) + U_T - \hat{m}(X))\big] \\
&= \mathbb{E}\big[(m(X)\theta + (g(X) - \hat{\ell}(X)) - (m(X) - \hat{m}(X))\hat{\theta}\big)(m(X) - \hat{m}(X)) + U_T^2(\theta - \hat{\theta})\big] \\
&= \mathbb{E}[(m(X) - \hat{m}(X))^2](\theta - \hat{\theta}) + \mathbb{E}[(g(X) - \hat{\ell}(X))(m(X) - \hat{m}(X))] \\
&\quad + \mathbb{E}[\hat{m}(X)(m(X) - \hat{m}(X))]\theta + \mathbb{E}[U_T^2](\theta - \hat{\theta})
\end{aligned}$$

and, consequently,

$$\hat{\theta}_{\text{R}} - \theta = \frac{\mathbb{E}[\hat{m}(X)(m(X) - \hat{m}(X))]}{\mathbb{E}[(m(X) - \hat{m}(X))^2] + \sigma_T^2}\theta + \frac{\mathbb{E}[(g(X) - \hat{\ell}(X))(m(X) - \hat{m}(X))]}{\mathbb{E}[(m(X) - \hat{m}(X))^2] + \sigma_T^2}$$

where the expectations are over $X$ and $\sigma_T^2 = \mathbb{E}[U_T^2]$ denotes the variance of the treatment noise.

# G Derivation of the ADML Estimator

Let us consider the limiting case $n \to \infty$, where

$$\hat{\theta}_{\text{ADML}} \xrightarrow{n} \hat{\theta} = \frac{\mathbb{E}[(Y - \hat{g}(X))(T - \tau\hat{m}(X))]}{\mathbb{E}[(T - \tau\hat{m}(X))T]} \quad \text{with} \quad \tau = \frac{\mathbb{E}[\hat{m}(X)T]}{\mathbb{E}[\hat{m}^2(X)]}. \tag{19}$$

When $\hat{m}(X) \xrightarrow{n} m(X)$, we have $\tau = 1$, so that $\hat{\theta}_{\text{ADML}}$ equals $\hat{\theta}_{\text{Z}}$, which yields a consistent estimator of $\theta$ (cf. Section 5.2). Now, let us consider $\hat{g}(X) \xrightarrow{n} g(X)$, which gives

$$\hat{\theta} = \frac{\mathbb{E}[(T\theta + U_Y)(T - \tau\hat{m}(X))]}{\mathbb{E}[(T - \tau\hat{m}(X))T]} = \frac{\mathbb{E}[(T - \tau\hat{m}(X))T]}{\mathbb{E}[(T - \tau\hat{m}(X))T]}\theta + \frac{\mathbb{E}[(m(X) - \tau\hat{m}(X) + U_T)U_Y]}{\mathbb{E}[(T - \tau\hat{m}(X))T]}$$

where the first term equals $\theta$ and the second term vanishes as long as the outcome noise $U_Y$ is uncorrelated with $\hat{m}(X)$.

Let us define the expected squared error as a function of $\hat{\theta}$ and $\hat{\phi}$, given plugin estimators of $m$ and $g$ denoted by $\hat{m}$ and $\hat{g}$, respectively:

$$\begin{aligned}
\mathcal{L}(\hat{\theta}, \hat{\phi}) &:= \mathbb{E}[(Y - \hat{Y})^2] = \mathbb{E}[(Y - (T - \hat{m}(X))\hat{\theta} - \hat{m}(X)\hat{\phi} - \hat{g}(X))^2] \\
\partial_{\hat{\phi}}\mathcal{L}(\hat{\theta}, \hat{\phi}) &= -2\mathbb{E}[(Y - (T - \hat{m}(X))\hat{\theta} - \hat{m}(X)\hat{\phi} - \hat{g}(X))\hat{m}(X)] \\
\partial_{\hat{\theta}}\mathcal{L}(\hat{\theta}, \hat{\phi}) &= -2\mathbb{E}[(Y - (T - \hat{m}(X))\hat{\theta} - \hat{m}(X)\hat{\phi} - \hat{g}(X))(T - \hat{m}(X))]
\end{aligned}$$

When setting the partial derivative w.r.t. $\hat{\phi}$ to zero and rearranging terms, we obtain

$$\mathbb{E}[\hat{m}^2(X)](\hat{\theta} - \hat{\phi}) = \mathbb{E}[\hat{m}(X)T]\hat{\theta} + \mathbb{E}[(\hat{g}(X) - Y)\hat{m}(X)]. \tag{20}$$

Similarly, when setting the sum of both partial derivatives to zero and rearranging terms, we obtain

$$\mathbb{E}[\hat{m}(X)T](\hat{\theta} - \hat{\phi}) = \mathbb{E}[T^2]\hat{\theta} + \mathbb{E}[(\hat{g}(X) - Y)T]. \tag{21}$$

Let us define

$$\tau = \frac{\mathbb{E}[\hat{m}(X)T]}{\mathbb{E}[\hat{m}^2(X)]}.$$

so that (20) reduces to

$$(\hat{\theta} - \hat{\phi}) = \tau\hat{\theta} + \frac{\mathbb{E}[(\hat{g}(X) - Y)\hat{m}(X)]}{\mathbb{E}[\hat{m}^2(X)]}.$$

Entering this expression in (21) gives

$$\tau\mathbb{E}[\hat{m}(X)T]\hat{\theta} + \tau\mathbb{E}[(\hat{g}(X) - Y)\hat{m}(X)] = \mathbb{E}[T^2]\hat{\theta} + \mathbb{E}[(\hat{g}(X) - Y)T],$$

and when solving for $\hat{\theta}$, we obtain

$$\hat{\theta} = \frac{\mathbb{E}[(Y - \hat{g}(X))(T - \tau\hat{m}(X))]}{\mathbb{E}[(T - \tau\hat{m}(X))T]}. \tag{22}$$

# H Experiment Setup

To evaluate our theoretical findings and compare estimator performance, we conduct extensive simulation studies. Our main scenario features an exponential treatment equation and cubic outcome equation, providing a clear setting where linear models are misspecified:

$$\begin{aligned}
&X_i \sim \mathcal{N}(-2,1) \text{ for } i \leq n/2, \quad X_i \sim \mathcal{N}(2,1) \text{ for } i > n/2 \\
&T_i = m(X_i) + U_{T,i} = -\exp(X_i) + U_{T,i} \\
&Y_i = g(X_i) + \theta T_i + U_{Y,i} = (-X_i + X_i^2 + X_i^3) + 0.5T_i + U_{Y,i}
\end{aligned} \tag{23}$$

where $U_{T,i}, U_{Y,i} \overset{iid}{\sim} \mathcal{N}(0,1)$ and the true treatment effect is $\theta = 0.5$. The confounder $X$ is drawn from a mixture of two normal distributions to ensure sufficient variation across its support. The exact parameters used are $m_0 = 0$, $m_1 = -1$, $m_2 = 1$ for the treatment equation and $g_0 = -1$, $g_1 = 1$, $g_2 = 1$ for the outcome equation.

# I Additional Empirical Results

Experimental results of models with different functional forms are presented in Figure 3.

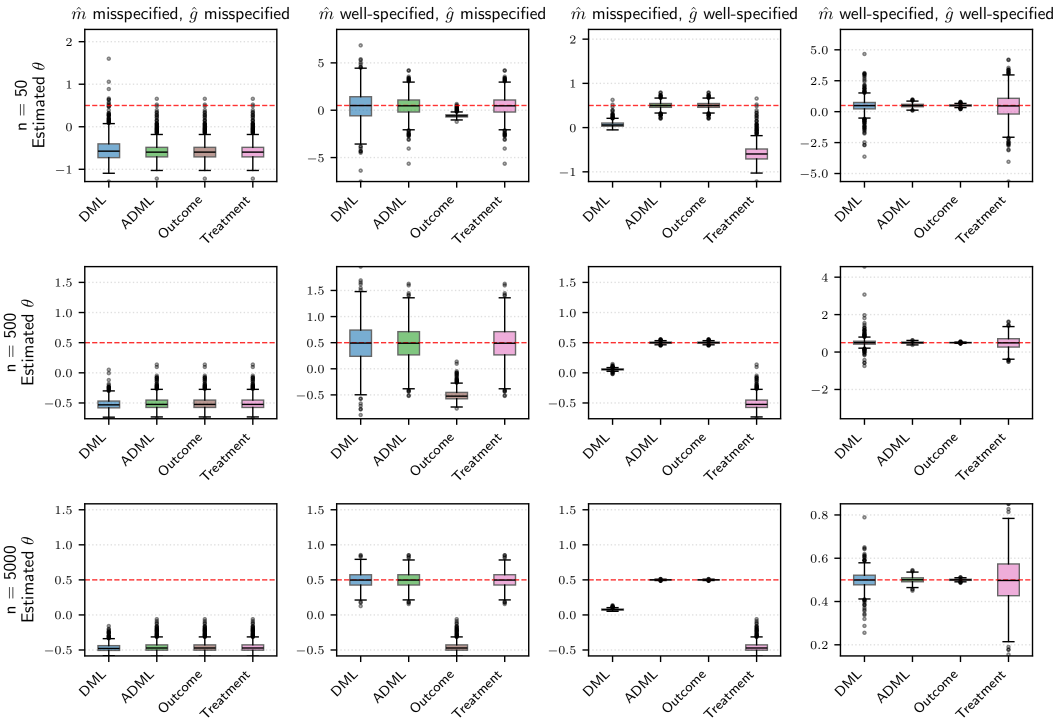

Figure 3: Performance of estimators across different sample sizes ($n = 50$, $n = 500$, and $n = 5000$) and model specifications. Each row represents a different sample size, with columns showing different combinations of model specifications. The data generating process features an exponential treatment equation and cubic outcome equation, with true causal effect $\theta = 0.5$ (dashed red line). Notably, misspecification of the treatment model $m$ leads to bias regardless of the sample size, while correct specification of $m$ yields consistent estimation with variance decreasing in sample size.