

Bias Amplification in Double Machine Learning

Gabriel Saco*

December 2025

Abstract

Orthogonal-score Double Machine Learning (DML) in partially linear regression is typically justified by nuisance-rate conditions, yet finite-sample validity also depends on the conditioning of the orthogonal score when treatment is highly predictable from covariates (limited residual treatment variation). We provide a finite-sample characterization of this amplification channel. Our main mathematical result is an exact identity for the cross-fitted PLR-DML estimator, $\hat{\theta} - \theta_0 = \hat{\kappa}(S'_n + B'_n)$, which holds without Taylor approximation and decomposes the estimation error into an oracle sampling term S'_n and a nuisance-induced term B'_n , both scaled by a sample condition number $\hat{\kappa}$. The identity yields a complete probability bound, $\hat{\theta} - \theta_0 = O_P(\sqrt{\kappa_n/n} + \kappa_n \text{Rem}_n)$, implying a sharp sufficiency requirement for root- n inference beyond oracle noise: $\kappa_n \text{Rem}_n = o(n^{-1/2})$. We further connect κ to semiparametric efficiency geometry via $\kappa = \sigma_D^2 \|\alpha_0\|_{L^2}^2$, where α_0 is the minimal-norm Riesz representer, and use a triangular-array framework to characterize regimes as overlap weakens ($\kappa_n \rightarrow \infty$); a bias decomposition clarifies when sign structure yields cancellation versus reinforcement. The theory delivers a feasible diagnostic, $\hat{\kappa}^* = 1/(1 - \hat{R}_{\text{oof}}^2(D | X))$, summarizing amplification scale. Rather than universal thresholds, we recommend reporting $\hat{\kappa}^*$ alongside cross-learner dispersion of DML estimates as a fragility assessment. Simulations validate monotone amplification and single-index organization, and a LaLonde reanalysis shows learner-to-learner instability increasing with $\hat{\kappa}^*$.

Keywords: Double Machine Learning, bias amplification, condition number, Riesz representer.

JEL Codes: C14, C21, C52.

*Universidad del Pacífico. Email: gsacoalvarado@gmail.com. Replication code is available at <https://github.com/gsaco/dml-diagnostic>.

1 Introduction

Double Machine Learning (DML; [Chernozhukov et al., 2018](#)) enables the use of machine learning methods for nuisance estimation in causal inference, while preserving \sqrt{n} -inference for treatment effects. Its guarantees rely on Neyman-orthogonal scores, which attenuate first-order sensitivity to nuisance errors, and cross-fitting, which removes own-observation bias ([Kennedy, 2023](#); [Schick, 1986](#)). Canonical validity conditions are therefore stated in terms of nuisance rates. For example, in the partially linear regression (PLR) model, a common restriction is a product-rate requirement such as $r_n^m r_n^\ell = o_P(n^{-1/2})$. These conditions control how nuisance estimation error enters the orthogonal score, but they do not explicitly track a distinct driver of finite-sample reliability that is central in causal problems: residual treatment variation after conditioning on covariates (overlap/positivity) ([Khan and Tamer, 2010](#); [D'Amour et al., 2021](#)). Standard DML theory typically assumes overlap through regularity conditions that keep this residual variation bounded away from zero. This paper shows that, in finite samples, reliability depends jointly on (i) the full nuisance remainder and (ii) the conditioning of the orthogonal score equation, so overlap enters multiplicatively as an amplification factor rather than only as a background assumption.

This amplification mechanism is transparent in PLR. The sensitivity of the empirical orthogonal score to the treatment effect θ is governed by the empirical Jacobian $\widehat{J} = \widehat{\sigma}_V^2$, where $V := D - m_0(X)$ is the residualized treatment. When $\widehat{\sigma}_V^2$ is small, the score is nearly flat in the θ direction, so small perturbations of the empirical score equation translate into large perturbations of $\widehat{\theta}$. This is ill-conditioning in the numerical sense ([Golub and Van Loan, 2013](#)). We summarize this channel by $\kappa := \sigma_D^2 / \sigma_V^2 = 1 / (1 - R^2(D | X))$, which is large when treatment is highly predictable from covariates. Crucially, κ scales the impact of any residual bias left after orthogonalization and cross-fitting. Standard errors can reflect increased sampling variability when σ_V^2 is small, but they do not reflect the distinct bias-amplification channel induced by large κ . As a result, coverage can deteriorate even when reported standard errors are only moderately inflated.

Our first main result is an exact finite-sample decomposition for DML in PLR:

$$\widehat{\theta} - \theta_0 = \widehat{\kappa} (S'_n + B'_n).$$

Here $\widehat{\kappa}$ is the sample analogue of the condition number, S'_n is the standardized oracle sampling term, and B'_n aggregates the nuisance-driven bias components that remain after orthogonalization and cross-fitting. The identity is exact—there is no Taylor approximation—because the PLR score is affine in θ . This makes the role of $\widehat{\kappa}$ as a finite-sample amplification factor explicit.

Our second main result converts the exact identity into a complete finite-sample probability bound:

$$\widehat{\theta} - \theta_0 = O_P\left(\frac{\sqrt{\kappa_n}}{\sqrt{n}} + \kappa_n \cdot \text{Rem}_n\right), \quad \text{Rem}_n = r_n^m r_n^\ell + (r_n^m)^2 + \frac{r_n^m + r_n^\ell}{\sqrt{n}}.$$

The first term is the oracle sampling component under conditioning. The second term shows that conditioning enters multiplicatively with the full nuisance remainder. As a consequence, a sufficient condition for valid \sqrt{n} inference beyond oracle noise is $\kappa_n \cdot \text{Rem}_n = o(n^{-1/2})$. When overlap is stable so that κ_n is bounded, this reduces to familiar remainder-rate restrictions. When overlap is weak and κ_n is large, the same nuisance errors can be amplified into first-order bias.

Two theoretical connections sharpen interpretation. First, we link conditioning to semiparametric efficiency geometry by proving $\kappa = \sigma_D^2 \|\alpha_0\|_{L^2}^2$, where α_0 is the Riesz representer ([Chernozhukov et al., 2022](#)). Second,

we formalize weakening overlap through a triangular-array framework in which $\kappa_n \rightarrow \infty$, showing that standard \sqrt{n} -asymptotics may fail even when nuisance rates are favorable. The theory also yields an estimable reporting implication. We propose reporting the standardized condition number,

$$\hat{\kappa}^* = \frac{1}{1 - \hat{R}_+^2(D | X)}, \quad \hat{R}_+^2 := \max\{0, \hat{R}_{\text{oof}}^2\},$$

where \hat{R}_{oof}^2 is an out-of-fold predictive R^2 . These summaries operationalize the overlap/conditioning component that enters the bound and help interpret cross-learner sensitivity predicted by the theory. Importantly, $\hat{\kappa}^*$ concerns the conditioning of the orthogonalized score (a finite-sample fragility channel), but it is not a test of unconfoundedness and does not by itself validate causal identification. Analogous to weak instruments—where weak first-stage signal amplifies bias and motivates strength diagnostics (Staiger and Stock, 1997; Stock and Yogo, 2005)—limited residual treatment variation in DML implies large κ , which in turn amplifies residual nuisance bias in the orthogonal score equation.

The paper proceeds as follows. Section 2 situates our contribution in the literature. Section 3 develops the exact decomposition and finite-sample bound. Section 4 characterizes conditioning regimes and rate implications. Section 5 presents Monte Carlo evidence validating the theoretical predictions. Section 6 provides an empirical illustration. Section 7 concludes.

2 Related Literature

We study semiparametric causal inference with machine-learned nuisance functions. Chernozhukov et al. (2018) formalize Double Machine Learning (DML) as Neyman-orthogonal, cross-fitted score estimation, building on Robinson’s partialling-out construction for PLR (Robinson, 1988) and influence-function theory (Newey, 1990; Bickel et al., 1993; van der Vaart, 1998). Newey and Robins (2018) provide general results for cross-fitting and Chernozhukov et al. (2023) give distributional refinements for orthogonal-score estimators. Debiasing geometry via Riesz representers and their learning is developed in Chernozhukov et al. (2021), while Kennedy (2023) reviews semiparametric/causal machine learning inference. Standard DML conditions are stated in terms of nuisance rates, but they do not explicitly track stability of the PLR score equation as residual treatment variation shrinks. We isolate this as a conditioning channel. Near-singularity induces a condition number κ that scales the orthogonal-score remainder and can drive finite-sample error. This connects overlap-driven fragility in DML to classical collinearity diagnostics (VIFs, condition numbers) (Belsley et al., 1980; Belsley, 1991).

Overlap is central for identification and performance of causal estimators. For binary treatment, efficiency theory shows sensitivity increases as propensity scores approach 0 or 1 (Hahn, 1998; Robins and Rotnitzky, 1995), motivating strict-overlap conditions and remedies such as trimming (Crump et al., 2009) and overlap weighting (Li et al., 2018). When overlap fails, effects may be only partially identified (Khan and Tamer, 2010). Overlap can also deteriorate with covariate dimension, creating tension between flexible adjustment and identification strength (D’Amour et al., 2021); see Petersen and van der Laan (2014) for practical discussions of positivity violations. We complement this literature by showing how weak overlap enters orthogonal-score DML for PLR through κ : limited residual treatment variation inflates κ , which can amplify the remainder even when orthogonality removes first-order nuisance sensitivity.

Modern causal estimators typically rely on regularized or adaptive learners for nuisance components, so

finite-sample orthogonal-score remainders need not be negligible even with cross-fitting. High-dimensional sparse and debiased approaches are developed in [Belloni et al. \(2014\)](#) and sufficient conditions under which flexible learners achieve the accuracy required for semiparametric inference are studied in [Farrell et al. \(2021\)](#). Related robustness ideas are central in doubly robust and targeted learning methods ([Kang and Schafer, 2007](#); [Bang and Robins, 2005](#); [van der Laan and Rose, 2011](#)). Conceptually, our diagnostic and reporting emphasis parallels the weak-instruments literature: weak first-stage signal can generate substantial finite-sample distortions and has motivated routine strength summaries and weak-IV-robust inference ([Staiger and Stock, 1997](#); [Stock and Yogo, 2005](#); [Moreira, 2003](#)). Analogously, we propose a simple out-of-fold summary of overlap-related conditioning in PLR-DML—computed as $\hat{\kappa}^*$ from out-of-fold prediction of D using X —as a routine diagnostic of fragility in learned causal estimators.

3 Theoretical Framework

This section introduces notation and assumptions for PLR-DML and develops our finite-sample characterization of how limited residual treatment variation affects inference. The main deliverables are (i) an explicit link between overlap strength and a condition number κ , (ii) an exact finite-sample identity for $\hat{\theta} - \theta_0$, and (iii) a probability bound that yields an operational sufficiency condition for \sqrt{n} -valid inference.

Table 1: Notation and Objects

Symbol	Definition
(Y, D, X)	Outcome, treatment, covariates
θ_0	Target causal parameter (constant marginal effect)
$g_0(X)$	Nuisance function in the outcome equation
$m_0(X)$	Treatment regression: $\mathbb{E}[D X]$
$\ell_0(X)$	Outcome regression: $\mathbb{E}[Y X] = \theta_0 m_0(X) + g_0(X)$
V	Treatment residual: $D - m_0(X)$
ε	Outcome error: $Y - \theta_0 D - g_0(X)$
σ_V^2, σ_D^2	$\mathbb{E}[V^2]$ (residual variance), $\text{Var}(D)$ (total treatment variance)
κ, κ^*	Population condition number: $\sigma_D^2/\sigma_V^2 = 1/(1 - R^2(D X))$
κ_n	Population condition number along a triangular array sequence
$\hat{\kappa}$	Sample estimate of κ from cross-fitted residuals
r_n^m, r_n^ℓ	L^2 convergence rates of nuisance estimators $\hat{m}, \hat{\ell}$
Rem_n	Complete remainder: $r_n^m r_n^\ell + (r_n^m)^2 + (r_n^m + r_n^\ell)/\sqrt{n}$
S_n, B_n	Oracle sampling term, nuisance bias term (Lemma 3.7)
α_0	Riesz representer: V/σ_V^2 (Theorem 3.6)

Notation convention. Throughout, $(\kappa, \kappa^*, \kappa_n)$ denote population objects under P_n (or P), while hats $(\hat{\kappa}, \hat{\kappa}^*)$ denote sample analogues computed from cross-fitted residuals. In empirical sections, we use $\hat{\kappa}^* := 1/(1 - \hat{R}_+^2(D | X))$ where $\hat{R}_+^2 := \max\{0, \hat{R}_{\text{oof}}^2\}$ truncates negative out-of-fold R^2 to ensure $\hat{\kappa}^* \geq 1$.

Main theoretical results. Before developing the technical arguments, we summarize the main results.

1. **Result 1 (Condition number connection).** The condition number κ connects treatment predictability to amplification via $\kappa = \sigma_D^2/\sigma_V^2 = 1/(1 - R^2(D | X)) = \sigma_D^2 \|\alpha_0\|_{L^2}^2$, where α_0 is the Riesz representer

(Theorem 3.6).

2. **Result 2 (Exact identity).** $\hat{\theta} - \theta_0 = \hat{\kappa}(S'_n + B'_n)$, an algebraic identity without Taylor remainder (Theorem 3.8). This is the central identity; everything else is interpretation or bounding.
3. **Result 3 (Complete bound).** $\hat{\theta} - \theta_0 = O_P(\sqrt{\kappa_n}/\sqrt{n} + \kappa_n \cdot \text{Rem}_n)$ (Theorem 3.11). This converts the identity into an operational probability bound.
4. **Sufficiency condition.** Valid \sqrt{n} -inference beyond oracle noise requires $\kappa_n \cdot \text{Rem}_n = o(n^{-1/2})$.

Diagnostic implication. Results 1–3 motivate reporting an estimable proxy for κ based on out-of-fold treatment predictability:

$$\hat{\kappa}^* := \frac{1}{1 - \hat{R}_+^2(D | X)}, \quad \hat{R}_+^2 := \max\{0, \hat{R}_{\text{oof}}^2\},$$

so that $\hat{\kappa}^* \geq 1$. A large $\hat{\kappa}^*$ indicates limited residual treatment variation and, by Result 3, a regime where even small orthogonal-score remainders can be amplified.

3.1 Data Structure

All random variables are defined on a complete probability space; expectations are under P_n .

Assumption 3.1 (Data Structure). For each sample size n , we observe $\{W_{i,n}\}_{i=1}^n$ i.i.d. under a law P_n , where $W_{i,n} = (Y_{i,n}, D_{i,n}, X_{i,n})$ consists of: outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$, treatment $D \in \mathcal{D} \subseteq \mathbb{R}$, and covariates $X \in \mathcal{X} \subseteq \mathbb{R}^p$. The classical i.i.d. setting is the special case $P_n \equiv P$. We suppress the index n when unambiguous.

Definition 3.1 (Function Spaces). For measurable $f : \mathcal{X} \rightarrow \mathbb{R}$, define $\|f\|_{L^2} := (\mathbb{E}[|f(X)|^2])^{1/2}$. The empirical norm is $\|f\|_n := (n^{-1} \sum_{i=1}^n f(X_i)^2)^{1/2}$. All expectations and L^2 norms are taken under P_n (with $P_n \equiv P$ in the classical i.i.d. case); we suppress the index n when unambiguous. When discussing triangular arrays we write $\sigma_{V,n}^2$; in the classical fixed-DGP case $P_n \equiv P$, we drop the subscript and write σ_V^2 .

We employ standard stochastic order notation: $Z_n = O_P(a_n)$ means for every $\epsilon > 0$ there exists $M > 0$ with $\sup_n \mathbb{P}(|Z_n/a_n| > M) < \epsilon$; $Z_n = o_P(a_n)$ means $Z_n/a_n \xrightarrow{P} 0$.

Assumption map. Lemma 3.1 uses only $\mathbb{E}[V\varepsilon] = 0$ and $\sigma_V^2 > 0$. The exact decomposition (Theorem 3.8) is algebraic under PLR and does not require rate conditions. Finite-sample bounds (Theorem 3.11) additionally use moment bounds, nuisance L^2 rates, and residual-variance stability. Regime analysis allows $\sigma_{V,n}^2 \rightarrow 0$ via the triangular array setup.

3.2 The Partially Linear Regression Model

The PLR model, introduced by Robinson (1988), specifies:

$$Y = \theta_0 D + g_0(X) + \varepsilon, \tag{1}$$

$$D = m_0(X) + V, \tag{2}$$

where $\theta_0 \in \mathbb{R}$ is the target parameter, $g_0 : \mathcal{X} \rightarrow \mathbb{R}$ is a nuisance function, $m_0(X) := \mathbb{E}[D | X]$ is the treatment regression (conditional mean), and $V := D - m_0(X)$ is the treatment residual satisfying $\mathbb{E}[V | X] = 0$ by construction.

Remark 3.1 (Terminology). When D is binary, $m_0(X) = \mathbb{E}[D | X]$ equals the propensity score $e(X)$ (Rosenbaum and Rubin, 1983). When D is continuous, $m_0(X)$ is the treatment regression (conditional mean), and our overlap notion is expressed through $\text{Var}(D | X)$.

3.3 Target Parameter and Identification

Definition 3.2 (Target Parameter). The target parameter is the constant marginal effect θ_0 in the partially linear model $Y = \theta_0 D + g_0(X) + \varepsilon$.

Lemma 3.1 (Identification of θ_0). *Under the PLR model (1)–(2) and Assumption 3.2, the target parameter θ_0 is identified as the unique solution to*

$$\mathbb{E}[V\{Y - \theta_0 D - g_0(X)\}] = 0,$$

equivalently,

$$\theta_0 = \frac{\mathbb{E}[VY]}{\mathbb{E}[V^2]} = \frac{\mathbb{E}[(D - m_0(X))Y]}{\mathbb{E}[(D - m_0(X))^2]},$$

provided $\sigma_V^2 = \mathbb{E}[V^2] > 0$. This “partialling-out” identification strategy dates to Frisch and Waugh (1933) and Robinson (1988).

Proof. From (1), $Y = \theta_0 D + g_0(X) + \varepsilon$. Multiply by $V = D - m_0(X)$ and take expectations:

$$\mathbb{E}[VY] = \theta_0 \mathbb{E}[VD] + \mathbb{E}[Vg_0(X)] + \mathbb{E}[V\varepsilon].$$

Now $\mathbb{E}[VD] = \mathbb{E}[V(m_0(X) + V)] = \mathbb{E}[V^2] = \sigma_V^2$, and $\mathbb{E}[Vg_0(X)] = \mathbb{E}[\mathbb{E}[V | X]g_0(X)] = 0$ since $\mathbb{E}[V | X] = 0$. Finally, Assumption 3.2 implies $\mathbb{E}[V\varepsilon] = 0$. Rearranging yields the formula. \square

3.4 Causal Setup and Identification

Assumption 3.2 (Causal PLR and Conditional Mean Independence). There exist potential outcomes $\{Y(d) : d \in \mathcal{D}\}$ (Rubin, 1974, 2005) such that

$$Y(d) = \theta_0 d + g_0(X) + \varepsilon, \quad \mathbb{E}[\varepsilon | D, X] = 0,$$

and consistency holds: $Y = Y(D)$.

Remark 3.2 (What is Identified). Under Assumption 3.2, θ_0 is a *constant causal marginal effect* (a constant CATE). All results below rely on the residual orthogonality moment $\mathbb{E}[V\varepsilon] = 0$, which is implied by the causal restriction $\mathbb{E}[\varepsilon | D, X] = 0$ because $\mathbb{E}[V\varepsilon] = \mathbb{E}\{\mathbb{E}[(D - \mathbb{E}[D | X])\varepsilon | X]\} = \mathbb{E}\{\mathbb{E}[D\varepsilon | X] - \mathbb{E}[D | X]\mathbb{E}[\varepsilon | X]\} = 0$.

Causal chain. The estimand θ_0 is causal under standard potential-outcome conditions: consistency ($Y = Y(D)$), conditional exogeneity ($\mathbb{E}[\varepsilon | D, X] = 0$), and overlap ($\text{Var}(D | X) > 0$). In the PLR framework,

these conditions imply the orthogonal moment $\mathbb{E}[V \cdot (Y - g_0(X) - \theta_0(D - m_0(X)))] = 0$. Our analysis takes this causal target as given and studies how finite-sample inference behaves as overlap weakens (via σ_V) and κ grows. Thus, κ measures inferential fragility, not identification validity.

We distinguish (i) strong overlap (fixed- κ) and (ii) weakening overlap (triangular array):

Assumption 3.3 (Strong Overlap). There exists $\underline{\sigma}^2 > 0$ such that $\text{Var}(D | X = x) \geq \underline{\sigma}^2$ for P_X -almost all x . This is the continuous-treatment analogue of the positivity condition (Rosenbaum and Rubin, 1983; Hirano et al., 2003). D'Amour et al. (2021) establish that such strict overlap assumptions become more restrictive as covariate dimension grows.

Assumption 3.4 (Bounded Treatment Variance). $\sigma_{D,n}^2 \asymp 1$; i.e., $0 < c \leq \sigma_{D,n}^2 \leq C < \infty$ for some constants c, C and all n .

Remark 3.3 (Bounded κ under Strong Overlap). Assumption 3.3 implies $\sigma_{V,n}^2 \geq \underline{\sigma}^2 > 0$. Under bounded treatment variance (Assumption 3.4), it follows that $\kappa_n \leq (\sup_n \sigma_{D,n}^2)/\underline{\sigma}^2 < \infty$, so $\kappa_n = O(1)$.

Remark 3.4 (Relation to Positivity / Overlap in the Binary Case). If $D \in \{0, 1\}$ and $e(X) := \mathbb{P}(D = 1 | X)$, then

$$\text{Var}(D | X) = e(X)\{1 - e(X)\}, \quad \sigma_V^2 = \mathbb{E}[e(X)\{1 - e(X)\}].$$

Thus, the usual positivity condition $\epsilon \leq e(X) \leq 1 - \epsilon$ implies $\text{Var}(D | X) \geq \epsilon(1 - \epsilon)$, which is a special case of Assumption 3.3.

To analyze regimes where κ may grow, we introduce a triangular array framework:

Assumption 3.5 (Triangular Array with Weakening Overlap). Consider a sequence of DGPs indexed by n . There exists a sequence $\underline{\sigma}_n^2 \downarrow 0$ such that for each n :

$$\text{Var}(D_n | X_n = x) \geq \underline{\sigma}_n^2 \quad \text{for } P_{X,n}\text{-almost all } x.$$

Define $\sigma_{V,n}^2 := \mathbb{E}[\text{Var}(D_n | X_n)]$ and $\kappa_n := \sigma_{D,n}^2/\sigma_{V,n}^2$.

Remark 3.5 (Reconciling Fixed and Growing κ). Assumption 3.3 covers the fixed- κ case (bounded condition number). Assumption 3.5 covers the growing- κ case: as $\underline{\sigma}_n^2 \rightarrow 0$, we may have $\sigma_{V,n}^2 \rightarrow 0$ and thus $\kappa_n \rightarrow \infty$. The rate $\kappa_n = O(n^\gamma)$ for $\gamma \geq 0$ determines the conditioning regime. When presenting results under Assumption 3.3, κ is treated as fixed; when presenting asymptotic regime analysis, Assumption 3.5 applies.

Define $\ell_0(X) := \mathbb{E}[Y | X]$ and outcome residual $U := Y - \ell_0(X)$. Under Assumption 3.2, $\ell_0(X) = \theta_0 m_0(X) + g_0(X)$.

Lemma 3.2 (Residual Decomposition). *Under PLR, $U = \theta_0 V + \varepsilon$.*

Proof. By definition, $U = Y - \ell_0(X)$. Substituting $Y = \theta_0 D + g_0(X) + \varepsilon$ and $\ell_0(X) = \theta_0 m_0(X) + g_0(X)$:

$$\begin{aligned} U &= [\theta_0 D + g_0(X) + \varepsilon] - [\theta_0 m_0(X) + g_0(X)] \\ &= \theta_0 D - \theta_0 m_0(X) + \varepsilon \\ &= \theta_0(D - m_0(X)) + \varepsilon = \theta_0 V + \varepsilon. \end{aligned} \quad \square$$

Remark 3.6 (Interpretation). Lemma 3.2 shows the outcome residual equals the causal effect times the treatment residual plus noise. The treatment residual V contains all identifying variation; the precision of identifying θ_0 depends on $\text{Var}(V) = \sigma_V^2$.

3.5 Variance Components and the Condition Number

Assumption 3.6 (Second Moments). $\mathbb{E}[D^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$ for all n .

Define the variance components:

$$\sigma_D^2 := \text{Var}(D), \quad (3)$$

$$\sigma_V^2 := \mathbb{E}[V^2] = \mathbb{E}[\text{Var}(D | X)], \quad (4)$$

$$\sigma_m^2 := \text{Var}(m_0(X)). \quad (5)$$

Lemma 3.3 (Law of Total Variance). *The law of total variance (Fisher, 1925) yields $\sigma_D^2 = \sigma_V^2 + \sigma_m^2$.*

Proof. By the law of total variance:

$$\begin{aligned} \text{Var}(D) &= \mathbb{E}[\text{Var}(D | X)] + \text{Var}(\mathbb{E}[D | X]) \\ &= \mathbb{E}[(D - m_0(X))^2] + \text{Var}(m_0(X)) \\ &= \sigma_V^2 + \sigma_m^2. \end{aligned}$$
□

The population R^2 for treatment explained by covariates is:

$$R^2(D | X) := \frac{\sigma_m^2}{\sigma_D^2} = 1 - \frac{\sigma_V^2}{\sigma_D^2}. \quad (6)$$

Definition 3.3 (Standardized Condition Number). The standardized condition number is:

$$\kappa^* := \frac{\sigma_D^2}{\sigma_V^2} = \frac{1}{1 - R^2(D | X)}.$$

(7)

The empirical analogue uses cross-fitted residuals and defines the residual second moment (this equals the Jacobian magnitude; centering is not used):

$$\hat{\sigma}_V^2 := \frac{1}{n} \sum_{i=1}^n \hat{V}_i^2, \quad \hat{\sigma}_D^2 := \frac{1}{n} \sum_{i=1}^n (D_i - \bar{D})^2, \quad \bar{D} := \frac{1}{n} \sum_{i=1}^n D_i,$$

and $\hat{\kappa} := \hat{\sigma}_D^2 / \hat{\sigma}_V^2$.

Remark 3.7 (VIF Connection). The condition number $\kappa^* = 1/(1 - R^2(D | X))$ has the same *functional form* as the classical Variance Inflation Factor (Belsley et al., 1980), and reduces to the classical VIF when $R^2(D | X)$ is interpreted as the R^2 from the linear projection of D on X . Our $R^2(D | X)$ uses the nonparametric conditional mean $m_0(X) = \mathbb{E}[D | X]$, making κ^* a nonparametric generalization. We write $\kappa := \kappa^*$ for the population standardized condition number.

Remark 3.8 (Properties of κ). By Lemma 3.3, $\sigma_D^2 \geq \sigma_V^2$, so $\kappa \geq 1$ with equality when $m_0(X)$ is constant (treatment is unpredictable). As $R^2(D | X) \rightarrow 1$, we have $\sigma_V^2 \rightarrow 0$ and $\kappa \rightarrow \infty$.

Remark 3.9 (Binary-Treatment Specialization of κ). If $D \in \{0, 1\}$ with $\mathbb{P}(D = 1) = p$, then $\sigma_D^2 = p(1 - p)$ and $\sigma_V^2 = \mathbb{E}[e(X)\{1 - e(X)\}]$, so

$$\kappa = \frac{p(1 - p)}{\mathbb{E}[e(X)\{1 - e(X)\}]},$$

which grows as overlap weakens.

3.6 Cross-Fitting and the DML Estimator

Definition 3.4 (Cross-Fitting). A K -fold partition $\{I_1, \dots, I_K\}$ of $\{1, \dots, n\}$ with disjoint folds. For $i \in I_k$, nuisance estimates $\widehat{m}^{(-k)}, \widehat{\ell}^{(-k)}$ are trained on $\{W_j : j \notin I_k\}$. We take K fixed as $n \rightarrow \infty$. This sample-splitting strategy originates in [Schick \(1986\)](#) and [Bickel \(1982\)](#), and is central to modern semiparametric estimation ([Bickel et al., 1993](#); [Chernozhukov et al., 2018](#)). Throughout, we assume the relevant second moments exist under P_n so that $\sigma_{D,n}^2, \sigma_{V,n}^2$, and L^2 norms are well-defined.

Cross-fitted residuals: $\widehat{V}_i := D_i - \widehat{m}^{(-k)}(X_i), \widehat{U}_i := Y_i - \widehat{\ell}^{(-k)}(X_i)$ for $i \in I_k$. Define errors:

$$\Delta_i^m := m_0(X_i) - \widehat{m}^{(-k)}(X_i), \quad (8)$$

$$\Delta_i^\ell := \ell_0(X_i) - \widehat{\ell}^{(-k)}(X_i). \quad (9)$$

Remark 3.10 (Residual Decomposition). With these sign conventions, the cross-fitted residuals decompose as:

$$\widehat{V}_i = D_i - \widehat{m}^{(-k)}(X_i) = (D_i - m_0(X_i)) + (m_0(X_i) - \widehat{m}^{(-k)}(X_i)) = V_i + \Delta_i^m,$$

and similarly $\widehat{U}_i = U_i + \Delta_i^\ell$. This decomposition is central: estimated residuals equal true residuals plus nuisance error.

The PLR score is:

$$\psi(W; \theta, \eta) := (D - m(X))\{Y - \ell(X) - \theta(D - m(X))\}, \quad (10)$$

where $\eta = (\ell, m)$. At true values, $\psi(W; \theta_0, \eta_0) = V\varepsilon$.

Lemma 3.4 (Neyman Orthogonality). *The pathwise derivative of $\mathbb{E}[\psi(W; \theta_0, \eta)]$ with respect to η vanishes at η_0 . This property is the cornerstone of debiased machine learning ([Chernozhukov et al., 2018](#)).*

The proof is in Appendix A. Neyman orthogonality implies the first-order effect of nuisance perturbations on the population moment vanishes; combined with cross-fitting, the leading sample terms involving Δ^m, Δ^ℓ are mean-zero and of order $(r_n^m + r_n^\ell)/\sqrt{n}$, while the systematic remainder is second order (e.g., $r_n^m r_n^\ell$ and $(r_n^m)^2$).

Definition 3.5 (DML Estimator).

$$\widehat{\theta} := \frac{\sum_{i=1}^n \widehat{V}_i \widehat{U}_i}{\sum_{i=1}^n \widehat{V}_i^2}. \quad (11)$$

Definition 3.6 (Empirical Score Map). For any $(\theta, \eta) = (\theta, \ell, m)$ define

$$\Psi_n(\theta, \eta) := \frac{1}{n} \sum_{i=1}^n (D_i - m(X_i))\{Y_i - \ell(X_i) - \theta(D_i - m(X_i))\}.$$

With cross-fitting, $\Psi_n(\theta, \widehat{\eta})$ is computed using fold-specific $\widehat{\ell}^{(-k)}, \widehat{m}^{(-k)}$ for $i \in I_k$.

3.7 The Score Jacobian

Definition 3.7 (Empirical Jacobian).

$$\widehat{J}_\theta := \frac{\partial}{\partial \theta} \left[\frac{1}{n} \sum_i \widehat{V}_i (\widehat{U}_i - \theta \widehat{V}_i) \right] = -\frac{1}{n} \sum_i \widehat{V}_i^2 = -\widehat{\sigma}_V^2. \quad (12)$$

Remark 3.11 (Jacobian Interpretation). The Jacobian magnitude $|\widehat{J}_\theta| = \widehat{\sigma}_V^2$ measures the score's curvature—how quickly the score changes as θ varies. When $\widehat{\sigma}_V^2$ is small, the score is nearly flat in the θ direction, and small score perturbations cause large parameter shifts. This is the classical numerical-analysis insight that condition numbers govern error propagation (Golub and Van Loan, 2013).

Lemma 3.5 (Jacobian-Kappa Relationship).

$$|\widehat{J}_\theta|^{-1} = \frac{\widehat{\kappa}}{\widehat{\sigma}_D^2}. \quad (13)$$

Proof. From Definition 3.7, $|\widehat{J}_\theta| = \widehat{\sigma}_V^2$. From Definition 3.3, $\widehat{\kappa} = \widehat{\sigma}_D^2 / \widehat{\sigma}_V^2$. Therefore:

$$|\widehat{J}_\theta|^{-1} = \frac{1}{\widehat{\sigma}_V^2} = \frac{\widehat{\kappa}}{\widehat{\sigma}_D^2}. \quad \square$$

Remark 3.12 (Key Insight). When we invert the Jacobian to solve for estimator error, the condition number $\widehat{\kappa}$ appears as a scale-invariant amplification factor. This is the mechanism through which ill-conditioning affects inference.

3.8 Connection to the Riesz Representer

Definition 3.8 (Minimal-Norm Representer). For the PLR moment, we define the minimal-norm representer $\alpha_0 : \mathcal{W} \rightarrow \mathbb{R}$ as the unique function satisfying:

- (i) $\mathbb{E}[\alpha_0(W) \cdot f(X)] = 0$ for all $f \in L^2(P_X)$;
- (ii) $\mathbb{E}[\alpha_0(W) \cdot D] = 1$;
- (iii) $\|\alpha_0\|_{L^2}^2 = \min\{\|\alpha\|_{L^2}^2 : \alpha \text{ satisfies (i)-(ii)}\}$.

This is the Riesz representer for the functional $\theta \mapsto \mathbb{E}[\alpha(W) \cdot \theta D]$ restricted to the space orthogonal to $L^2(P_X)$, expressed as a constrained minimal-norm problem (Riesz, 1907; Chernozhukov et al., 2022).

Theorem 3.6 (Condition Number as Riesz Norm). *In the PLR model, $\alpha_0(W) = V/\sigma_V^2$, and:*

$$\boxed{\kappa = \sigma_D^2 \cdot \|\alpha_0\|_{L^2}^2, \quad \text{where } \|\alpha_0\|_{L^2}^2 = 1/\sigma_V^2.} \quad (14)$$

Proof. We show $\alpha_0(W) = V/\sigma_V^2$ satisfies Definition 3.8 and is the unique minimizer.

Step 1 (Orthogonality): For any $f \in L^2(P_X)$:

$$\begin{aligned}\mathbb{E}[\alpha_0(W) \cdot f(X)] &= \frac{1}{\sigma_V^2} \mathbb{E}[V \cdot f(X)] \\ &= \frac{1}{\sigma_V^2} \mathbb{E}[\mathbb{E}[V | X] \cdot f(X)] = 0,\end{aligned}$$

since $\mathbb{E}[V | X] = 0$ by construction.

Step 2 (Normalization):

$$\begin{aligned}\mathbb{E}[\alpha_0(W) \cdot D] &= \frac{1}{\sigma_V^2} \mathbb{E}[V \cdot (m_0(X) + V)] \\ &= \frac{1}{\sigma_V^2} (\mathbb{E}[Vm_0(X)] + \mathbb{E}[V^2]) = \frac{\sigma_V^2}{\sigma_V^2} = 1.\end{aligned}$$

Step 3 (Norm computation):

$$\|\alpha_0\|_{L^2}^2 = \mathbb{E} \left[\frac{V^2}{\sigma_V^4} \right] = \frac{\sigma_V^2}{\sigma_V^4} = \frac{1}{\sigma_V^2}.$$

Step 4 (Minimality and uniqueness): Let α be any function satisfying (i)–(ii). Define $h := \alpha - \alpha_0$. Then h satisfies:

- $\mathbb{E}[h(W)f(X)] = 0$ for all $f \in L^2(P_X)$ (orthogonality inherited);
- $\mathbb{E}[h(W)D] = 0$ (normalization difference).

Since $D = m_0(X) + V$ and $\mathbb{E}[h(W)m_0(X)] = 0$ by orthogonality, we have $\mathbb{E}[h(W)V] = 0$.

Now, $\alpha_0 = V/\sigma_V^2$ is a scalar multiple of V , so $\mathbb{E}[h \cdot \alpha_0] = \mathbb{E}[hV]/\sigma_V^2 = 0$. By the Pythagorean identity:

$$\|\alpha\|_{L^2}^2 = \|\alpha_0 + h\|_{L^2}^2 = \|\alpha_0\|_{L^2}^2 + \|h\|_{L^2}^2 \geq \|\alpha_0\|_{L^2}^2,$$

with equality if and only if $h = 0$ a.s. Thus α_0 is the unique minimizer.

Step 5 (Final result):

$$\sigma_D^2 \cdot \|\alpha_0\|_{L^2}^2 = \sigma_D^2 \cdot \frac{1}{\sigma_V^2} = \kappa. \quad \square$$

Remark 3.13 (Semiparametric Interpretation). Theorem 3.6 grounds our diagnostic in modern semiparametric theory (Newey, 1990; Bickel et al., 1993; Tsiatis, 2006; Kennedy, 2016; Chernozhukov et al., 2022). The minimal-norm representer α_0 is the correction weight for double robustness (Bang and Robins, 2005); its norm measures how much correction is required. The terminology “Riesz representer” refers to this object’s role as the representer of a linear functional under the $L^2(P)$ inner product, restricted to the space orthogonal to nuisance directions (Riesz, 1909). A large $\|\alpha_0\|_{L^2}^2$ signals both large variance and large sensitivity to nuisance bias.

Remark 3.14 (Hilbert-Space View). Constraint (i) is equivalent to $\mathbb{E}[\alpha(W) | X] = 0$, i.e. α lies in the orthogonal complement of $L^2(P_X)$ inside $L^2(P)$. The minimization in Definition 3.8 therefore selects the minimum- L^2 instrument in the residual variation direction V .

3.9 The Exact Finite-Sample Decomposition

Define the oracle sampling term:

$$S_n := \frac{1}{n} \sum_{i=1}^n V_i \varepsilon_i. \quad (15)$$

Lemma 3.7 (Bias Decomposition). *Define the nuisance bias term $B_n := \Psi_n(\theta_0, \hat{\eta}) - S_n$. Then:*

$$B_n = B_n^{(1)} + B_n^{(2)} + B_n^{(3)} + B_n^{(4)} + B_n^{(5)}, \quad (16)$$

where:

$$B_n^{(1)} := \frac{1}{n} \sum_i V_i \Delta_i^\ell, \quad (17)$$

$$B_n^{(2)} := -\theta_0 \frac{1}{n} \sum_i V_i \Delta_i^m, \quad (18)$$

$$B_n^{(3)} := \frac{1}{n} \sum_i \Delta_i^m \varepsilon_i, \quad (19)$$

$$B_n^{(4)} := \frac{1}{n} \sum_i \Delta_i^m \Delta_i^\ell, \quad (20)$$

$$B_n^{(5)} := -\theta_0 \frac{1}{n} \sum_i (\Delta_i^m)^2. \quad (21)$$

Proof. Expand $\widehat{V}_i(\widehat{U}_i - \theta_0 \widehat{V}_i)$ using $\widehat{V}_i = V_i + \Delta_i^m$ and $\widehat{U}_i = U_i + \Delta_i^\ell = \theta_0 V_i + \varepsilon_i + \Delta_i^\ell$:

$$\begin{aligned} \widehat{V}_i(\widehat{U}_i - \theta_0 \widehat{V}_i) &= (V_i + \Delta_i^m)[(\theta_0 V_i + \varepsilon_i + \Delta_i^\ell) - \theta_0(V_i + \Delta_i^m)] \\ &= (V_i + \Delta_i^m)[\varepsilon_i + \Delta_i^\ell - \theta_0 \Delta_i^m]. \end{aligned}$$

Expanding:

$$\begin{aligned} &= V_i \varepsilon_i + V_i \Delta_i^\ell - \theta_0 V_i \Delta_i^m \\ &\quad + \Delta_i^m \varepsilon_i + \Delta_i^m \Delta_i^\ell - \theta_0 (\Delta_i^m)^2. \end{aligned}$$

Averaging over i and subtracting $S_n = n^{-1} \sum_i V_i \varepsilon_i$ yields the five terms. \square

Remark 3.15 (Interpretation of Bias Components). Terms $B_n^{(1)} - B_n^{(3)}$ are “first-order” in nuisance error: under cross-fitting, they have conditional mean zero and contribute $O_P(n^{-1/2})$ variance. Term $B_n^{(4)}$ is the product term driving the product-rate requirement. Term $B_n^{(5)}$ is always negative (for $\theta_0 > 0$) and scales as $(r_n^m)^2$.

Theorem 3.8 (Exact Decomposition). *Under PLR, the DML estimator satisfies:*

$$\boxed{\widehat{\theta} - \theta_0 = \widehat{\kappa} \cdot S'_n + \widehat{\kappa} \cdot B'_n,} \quad (22)$$

where $S'_n := S_n / \widehat{\sigma}_D^2$ and $B'_n := B_n / \widehat{\sigma}_D^2$. This is an exact algebraic identity—no Taylor approximation.

Proof. *Step 1 (Solve for estimator):* The DML estimator solves $\Psi_n(\hat{\theta}, \hat{\eta}) = 0$. Since the score is affine in θ :

$$\Psi_n(\theta, \hat{\eta}) = \frac{1}{n} \sum_i \hat{V}_i \hat{U}_i - \theta \cdot \hat{\sigma}_V^2.$$

Setting $\Psi_n = 0$: $\hat{\theta} = \frac{1}{n} \sum_i \hat{V}_i \hat{U}_i / \hat{\sigma}_V^2$.

Step 2 (Express error): Subtracting θ_0 :

$$\hat{\theta} - \theta_0 = \frac{\frac{1}{n} \sum_i \hat{V}_i (\hat{U}_i - \theta_0 \hat{V}_i)}{\hat{\sigma}_V^2} = \frac{\Psi_n(\theta_0, \hat{\eta})}{\hat{\sigma}_V^2}.$$

Step 3 (Decompose score): By Lemma 3.7:

$$\Psi_n(\theta_0, \hat{\eta}) = S_n + B_n.$$

Step 4 (Standardize):

$$\begin{aligned} \hat{\theta} - \theta_0 &= \frac{S_n + B_n}{\hat{\sigma}_V^2} \\ &= \frac{\hat{\sigma}_D^2}{\hat{\sigma}_V^2} \cdot \frac{S_n + B_n}{\hat{\sigma}_D^2} \\ &= \hat{\kappa}(S'_n + B'_n). \end{aligned}$$

□

Remark 3.16 (Why Exact). The decomposition is exact because the PLR score is affine in θ . There is no Taylor expansion, hence no remainder term. This exactness holds at any sample size.

Remark 3.17 (Dimensional Consistency). S_n has units $[D][Y]$, $\hat{\sigma}_D^2$ has units $[D]^2$, so S'_n has units $[Y]/[D]$ matching θ_0 . Since $\hat{\kappa}$ is dimensionless (ratio of variances), the decomposition is unit-consistent.

3.10 Variance Inflation versus Bias Amplification

The exact decomposition shows that $\hat{\kappa}$ multiplies both the oracle sampling component and the nuisance-induced bias, but the inferential consequences are fundamentally different. On the one hand, variance inflation arises through the oracle term $\hat{\kappa} S'_n$, since S'_n scales with σ_V and, under the efficiency bound in Theorem 3.13, the relevant benchmark variance is $V_{\text{eff}} = \sigma_\varepsilon^2 / \sigma_V^2$. In this case the usual standard errors track the increased sampling variability induced by limited residual treatment variation, so coverage can remain approximately nominal. On the other hand, bias amplification arises through $\hat{\kappa} B'_n$, because any remaining nuisance error is multiplied by κ ; this systematic shift is not reflected in conventional standard errors, so confidence intervals can remain tight while becoming badly miscentered. This divergence between narrow intervals and poor coverage is the “silent failure” highlighted by our analysis.

3.11 Finite-Sample Probability Bounds

Assumption 3.7 (Bounded Moments). $\mathbb{E}[D^4], \mathbb{E}[Y^4] < \infty$.

Assumption 3.8 (Moment Bounds). There exists a constant $C < \infty$ such that:

- (i) $\text{ess sup}_x \mathbb{E}[V^2 | X = x] \leq C$;
- (ii) $\text{ess sup}_{d,x} \mathbb{E}[\varepsilon^2 | D = d, X = x] \leq C$.

Remark 3.18 (Role of Moment Bounds). Assumption 3.8(i) ensures that V^2 has uniformly bounded conditional expectation, enabling the bound $\mathbb{E}[V^2(\Delta^\ell)^2] \leq C\|\Delta^\ell\|_{L^2}^2$. Assumption 3.8(ii) controls the oracle term variance: $\mathbb{E}[V^2\varepsilon^2] = \mathbb{E}[V^2\mathbb{E}[\varepsilon^2 | D, X]] \leq C\sigma_{V,n}^2$. Conditional homoskedasticity $\mathbb{E}[\varepsilon^2 | D, X] = \sigma_\varepsilon^2$ is a special case.

Assumption 3.9 (Nuisance Rates). $\|\widehat{m}^{(-k)} - m_0\|_{L^2} = O_P(r_n^m)$, $\|\widehat{\ell}^{(-k)} - \ell_0\|_{L^2} = O_P(r_n^\ell)$.

Assumption 3.10 (Residual-Variance Stability). $\sigma_{V,n}^2/\widehat{\sigma}_V^2 = O_P(1)$. Equivalently, there exists $c > 0$ with $\mathbb{P}(\widehat{\sigma}_V^2 \geq c\sigma_{V,n}^2) \rightarrow 1$.

Remark 3.19 (Role of Variance Stability). Assumption 3.10 ensures that the empirical residual variance $\widehat{\sigma}_V^2$ does not collapse relative to the population variance $\sigma_{V,n}^2$. This is needed to control the oracle term scaling $S_n/\widehat{\sigma}_V^2$. Under strong overlap with consistent estimation, this holds automatically; under weakening overlap, it becomes a substantive requirement.

Lemma 3.9 (Sufficient Condition for Variance Stability). *Suppose $\mathbb{E}[D^4] < \infty$ and let $\widehat{\sigma}_V^2 := n^{-1} \sum_{i=1}^n (D_i - \widehat{m}^{(-k(i))}(X_i))^2$ denote the cross-fitted residual second moment. If*

$$\|\widehat{m}^{(-k)} - m_0\|_{L^2} = o_P(\sigma_{V,n}) \quad \text{uniformly over } k,$$

then $\widehat{\sigma}_V^2/\sigma_{V,n}^2 \xrightarrow{P} 1$, hence Assumption 3.10 holds.

Proof. Write $\widehat{V}_i = V_i + \Delta_i^m$. Then

$$n^{-1} \sum_i \widehat{V}_i^2 = n^{-1} \sum_i V_i^2 + 2n^{-1} \sum_i V_i \Delta_i^m + n^{-1} \sum_i (\Delta_i^m)^2.$$

By Cauchy–Schwarz, $|n^{-1} \sum_i V_i \Delta_i^m| \leq \|V\|_n \|\Delta^m\|_n = O_P(\sigma_{V,n}) \cdot o_P(\sigma_{V,n}) = o_P(\sigma_{V,n}^2)$. Also $n^{-1} \sum_i (\Delta_i^m)^2 = \|\Delta^m\|_n^2 = o_P(\sigma_{V,n}^2)$. Finally, $n^{-1} \sum_i V_i^2 \xrightarrow{P} \sigma_{V,n}^2$ under finite fourth moments. \square \square

Remark 3.20 (Interpretation of Sufficient Condition). Lemma 3.9 shows that, under weakening overlap, variance stability is ensured when the first-stage error is small relative to the residual scale $\sigma_{V,n}$; this prevents $\widehat{\sigma}_V^2$ from collapsing due to first-stage error.

Lemma 3.10 (Foldwise Empirical-to-Population Norm). *Let $\{I_1, \dots, I_K\}$ be the cross-fitting folds (Definition 3.4). For each fold k , let $\Delta^{(-k)} : \mathcal{X} \rightarrow \mathbb{R}$ be any (possibly random) function measurable with respect to the training sigma-field generated by $\{W_j : j \notin I_k\}$. Define the fold empirical norm*

$$\|\Delta^{(-k)}\|_{n,k}^2 := \frac{1}{|I_k|} \sum_{i \in I_k} \Delta^{(-k)}(X_i)^2.$$

Then, conditional on the training sample for fold k ,

$$\mathbb{E}\left[\|\Delta^{(-k)}\|_{n,k}^2 \mid \{W_j : j \notin I_k\}\right] = \|\Delta^{(-k)}\|_{L^2(P_X)}^2.$$

Consequently, by Markov's inequality, $\|\Delta^{(-k)}\|_{n,k} = O_P(\|\Delta^{(-k)}\|_{L^2})$ uniformly over k .

If $\Delta_i := \Delta^{(-k)}(X_i)$ for $i \in I_k$, then the full-sample empirical norm satisfies

$$\|\Delta\|_n^2 = \frac{1}{n} \sum_{k=1}^K |I_k| \|\Delta^{(-k)}\|_{n,k}^2,$$

so in particular $\|\Delta\|_n = O_P(\max_{1 \leq k \leq K} \|\Delta^{(-k)}\|_{L^2})$.

Theorem 3.11 (Complete Finite-Sample Bound). *Under Assumptions 3.2–3.10:*

$$\widehat{\theta} - \theta_0 = O_P\left(\frac{\sqrt{\kappa_n}}{\sqrt{n}} + \kappa_n \cdot \text{Rem}_n\right),$$

(23)

where the remainder term is:

$$\text{Rem}_n := r_n^m r_n^\ell + (r_n^m)^2 + \frac{r_n^m + r_n^\ell}{\sqrt{n}}.$$
(24)

The oracle term $O_P(\sqrt{\kappa_n}/\sqrt{n})$ follows from $O_P(1/(\sigma_{V,n}\sqrt{n}))$ and Assumption 3.4.

Proof. From Theorem 3.8: $\widehat{\theta} - \theta_0 = \widehat{\kappa}(S'_n + B'_n)$.

Oracle term: The oracle term $S_n = n^{-1} \sum_i V_i \varepsilon_i$ has conditional mean zero and

$$\text{Var}(S_n) = \frac{1}{n} \mathbb{E}[V^2 \varepsilon^2].$$

By Assumption 3.8(ii) and iterated expectations:

$$\mathbb{E}[V^2 \varepsilon^2] = \mathbb{E}[V^2 \mathbb{E}[\varepsilon^2 \mid D, X]] \leq C \cdot \mathbb{E}[V^2] = C \sigma_{V,n}^2.$$

Thus $S_n = O_P(\sigma_{V,n}/\sqrt{n})$. By Assumption 3.10, $\sigma_{V,n}^2/\widehat{\sigma}_V^2 = O_P(1)$, so:

$$\widehat{\kappa} S'_n = \frac{S_n}{\widehat{\sigma}_V^2} = O_P\left(\frac{\sigma_{V,n}}{\sqrt{n}} \cdot \frac{1}{\sigma_{V,n}^2}\right) = O_P\left(\frac{1}{\sigma_{V,n}\sqrt{n}}\right).$$

Bias term: By Lemma 3.7, $B_n = \sum_{j=1}^5 B_n^{(j)}$.

Terms $B_n^{(1)} \dots B_n^{(3)}$: Under cross-fitting, these have conditional mean zero. For $B_n^{(1)} = n^{-1} \sum_i V_i \Delta_i^\ell$, by Assumption 3.8(i) and iterated expectations:

$$\mathbb{E}[V^2 (\Delta^\ell)^2] = \mathbb{E}[(\Delta^\ell)^2 \mathbb{E}[V^2 \mid X]] \leq C \|\Delta^\ell\|_{L^2}^2 = O_P((r_n^\ell)^2).$$

Thus $\text{Var}(B_n^{(1)}) = O_P((r_n^\ell)^2/n)$ and $B_n^{(1)} = O_P(r_n^\ell/\sqrt{n})$. Similarly for $B_n^{(2)}, B_n^{(3)}$.

Term $B_n^{(4)}$: By sample Cauchy–Schwarz:

$$|B_n^{(4)}| = \left| \frac{1}{n} \sum_i \Delta_i^m \Delta_i^\ell \right| \leq \|\Delta^m\|_n \|\Delta^\ell\|_n.$$

By Lemma 3.10 applied foldwise to $\Delta^{m,(-k)}(x) := m_0(x) - \widehat{m}^{(-k)}(x)$ and $\Delta^{\ell,(-k)}(x) := \ell_0(x) - \widehat{\ell}^{(-k)}(x)$, we have $\|\Delta^m\|_n = O_P(r_n^m)$ and $\|\Delta^\ell\|_n = O_P(r_n^\ell)$. Hence $|B_n^{(4)}| = O_P(r_n^m r_n^\ell)$.

Term $B_n^{(5)}$: Similarly, $|B_n^{(5)}| = |\theta_0| \|\Delta^m\|_n^2 = O_P((r_n^m)^2)$.

Combining: $B_n = O_P(r_n^m r_n^\ell + (r_n^m)^2 + (r_n^m + r_n^\ell)/\sqrt{n}) = O_P(\text{Rem}_n)$.

Therefore: $\widehat{\kappa} B'_n = O_P(\kappa_n \cdot \text{Rem}_n)$. □

Remark 3.21 (Complete Remainder). The remainder (24) includes $(r_n^m)^2$ from $B_n^{(5)}$, which can dominate $r_n^m r_n^\ell$ when $r_n^m \gg r_n^\ell$. The cross-fitting terms contribute $(r_n^m + r_n^\ell)/\sqrt{n}$, negligible when $r_n^m, r_n^\ell = o(1)$.

Corollary 3.12 (Critical Rate). *Under strong overlap ($\inf_n \sigma_{V,n} > 0$) and bounded treatment variance (Assumption 3.4), $\kappa_n = O(1)$, so \sqrt{n} -consistent inference requires $\text{Rem}_n = o(n^{-1/2})$, equivalently $\kappa_n \cdot \text{Rem}_n = o(n^{-1/2})$.*

3.12 Efficiency Bound Connection

Assumption 3.11 (Conditional Homoskedasticity). $\mathbb{E}[\varepsilon^2 | D, X] = \sigma_\varepsilon^2$ almost surely.

Remark 3.22 (Why Condition on (D, X)). Assumption 3.11 is stronger than $\mathbb{E}[\varepsilon^2 | X] = \sigma_\varepsilon^2$. The latter does not imply $\mathbb{E}[V^2 \varepsilon^2] = \sigma_\varepsilon^2 \sigma_V^2$ because V^2 is not X -measurable. Under Assumption 3.11, since ε^2 is mean-independent of D given X , we can show $\mathbb{E}[V^2 \varepsilon^2] = \sigma_\varepsilon^2 \sigma_V^2$.

Theorem 3.13 (Efficiency and Condition Number). *Under Assumption 3.11, the semiparametric efficiency bound is:*

$$V_{\text{eff}} = \frac{\sigma_\varepsilon^2}{\sigma_V^2} = \frac{\sigma_\varepsilon^2 \kappa}{\sigma_D^2}. \quad (25)$$

Without Assumption 3.11, the bound is $V_{\text{eff}} = \mathbb{E}[V^2 \varepsilon^2]/\sigma_V^4$. This efficiency result connects to the classical semiparametric variance bound literature (Hahn, 1998; Newey, 1990; Bickel et al., 1993).

Proof. The efficiency bound for moment $\mathbb{E}[\psi] = 0$ is $V_{\text{eff}} = \mathbb{E}[\psi^2]/(\partial_\theta \mathbb{E}[\psi])^2$.

For $\psi = V\varepsilon$: $\mathbb{E}[\psi^2] = \mathbb{E}[V^2 \varepsilon^2]$ and $\partial_\theta \mathbb{E}[\psi] = -\sigma_V^2$.

Under Assumption 3.11:

$$\begin{aligned} \mathbb{E}[V^2 \varepsilon^2] &= \mathbb{E}[\mathbb{E}[V^2 \varepsilon^2 | D, X]] \\ &= \mathbb{E}[V^2 \mathbb{E}[\varepsilon^2 | D, X]] \quad (V^2 \text{ is } (D, X)\text{-measurable}) \\ &= \mathbb{E}[V^2 \sigma_\varepsilon^2] = \sigma_\varepsilon^2 \sigma_V^2. \end{aligned}$$

Therefore: $V_{\text{eff}} = \sigma_\varepsilon^2 \sigma_V^2 / \sigma_V^4 = \sigma_\varepsilon^2 / \sigma_V^2 = \sigma_\varepsilon^2 \kappa / \sigma_D^2$. □

4 Conditioning Regimes and Rate Implications

The regimes below label how overlap weakens along the triangular array. Plugging regime-specific κ_n growth into Theorem 3.11 yields the corresponding rate implications. In applications we observe a single sample size n and an estimate $\widehat{\kappa}$. In practice, the regimes serve as a qualitative sensitivity lens for how strongly the finite-sample bound amplifies estimation error. These regimes are labels, not hard cutoffs—no formal thresholds are proposed.

Terminology. In this paper, “conditioning” refers to κ -driven amplification in the orthogonal score. It is related to overlap but does not by itself diagnose identification. Overlap/positivity remains a causal identification assumption. κ^* quantifies how much residual treatment variation is available for inference.

From this point, we analyze sequences of DGPs under Assumption 3.5 (triangular array). In this regime, Assumption 3.3 is not imposed. It is replaced by the lower bound $\text{Var}(D_n | X_n = x) \geq \underline{\sigma}_n^2$ with $\underline{\sigma}_n^2 \downarrow 0$.

Remark 4.1 (Notation Consistency). Throughout, κ_n denotes the population condition number along the sequence, κ^* denotes the population standardized condition number under a fixed DGP, and $\widehat{\kappa}$ denotes the sample estimate.

4.1 Regime Classification

Under Assumption 3.5 (triangular array), κ_n may grow with n .

Definition 4.1 (Conditioning Regimes).

- (i) **Well-conditioned:** $\kappa_n = O(1)$. Standard \sqrt{n} -asymptotics apply.
- (ii) **Moderately ill-conditioned:** $\kappa_n = O(n^\gamma)$ for $0 < \gamma < 1/2$. Convergence continues at slower-than- \sqrt{n} rates.
- (iii) **Severely ill-conditioned:** $\kappa_n \asymp \sqrt{n}$. Standard \sqrt{n} -asymptotics fail; the estimator may converge at slower rates and becomes highly sensitive to the nuisance remainder Rem_n .

These regimes correspond to how fast residual treatment variation shrinks conditional on covariates (weakening overlap/positivity). “Well-conditioned” means overlap is effectively stable. “Moderately ill-conditioned” means overlap weakens slowly so error amplification grows but can still vanish if the nuisance remainder is sufficiently small. “Severely ill-conditioned” means overlap weakens fast enough that bias/variance amplification can overwhelm \sqrt{n} -scaling.

Practical interpretation by regime. In the well-conditioned regime, estimates should be stable across learner choices and small nuisance errors remain small after amplification. In the moderately ill-conditioned regime, cross-learner dispersion may become noticeable, warranting sensitivity analysis. In the severely ill-conditioned regime, confidence intervals can appear narrow yet systematically fail to cover the true parameter. Inference should be treated with skepticism and robustness checks prioritized over point estimates. The distinction is qualitative and comparative—practitioners should focus on how $\widehat{\kappa}^*$ changes across specifications rather than applying fixed numerical cutoffs.

Remark 4.2 (Fixed vs. Growing κ). Under Assumption 3.3 (strong overlap), κ is bounded, so we are always in regime (i). Regimes (ii)–(iii) arise under Assumption 3.5 when overlap weakens with n .

This follows by substituting the regime-specific growth of κ_n into the finite-sample bound implied by the exact decomposition (Theorem 3.8) and the assumed remainder rate Rem_n , under the same moment conditions used for the bound.

Theorem 4.1 (Rates by Regime). *Suppose $\text{Rem}_n = O(n^{-\alpha})$ for some $\alpha > 0$ and $\sigma_{D,n}^2 \asymp 1$:*

- (i) *Well-conditioned ($\kappa_n = O(1)$): $\widehat{\theta} - \theta_0 = O_P(n^{-1/2})$.*

- (ii) Moderately ill-conditioned ($\kappa_n = O(n^\gamma)$, $\gamma < 1/2$): $\hat{\theta} - \theta_0 = O_P(n^{\gamma/2-1/2} + n^{\gamma-\alpha})$. If $\alpha \leq \gamma$, the remainder term $n^{\gamma-\alpha}$ does not vanish and may dominate the stochastic term.
- (iii) Severely ill-conditioned ($\kappa_n \asymp \sqrt{n}$): $\hat{\theta} - \theta_0 = O_P(n^{-1/4} + n^{1/2-\alpha})$. If $\alpha < 1/2$, the remainder term dominates and may diverge.

In a single sample, $\hat{\kappa}$ summarizes sensitivity to nuisance bias. Reporting $\hat{\kappa}$ communicates the amplification scale suggested by the bound. When $\hat{\kappa}$ is elevated, compare estimates across learner specifications.

5 Monte Carlo Evidence

Our mathematical results (Theorem 3.8, Theorem 3.11) predict that finite-sample error decomposes into a sampling-noise term and a nuisance-error term that can be amplified by conditioning, summarized by κ^* . We demonstrate: (i) amplification is monotone in κ^* ; (ii) a single-index ($\kappa^* \times$ nuisance-error magnitude) explains bias and coverage dynamics; (iii) sign structure can remove or restore cancellations in the bias decomposition. This is a mechanism test, not a claim about realistic learner behavior—we perturb oracle nuisances by known amounts to isolate the amplification channel.

Predictions P1–P3 are direct qualitative implications of the exact decomposition (Theorem 3.8) and the finite-sample bound (Theorem 3.11):

P1 Bias amplification: For fixed nuisance-error magnitude δ , $|\text{Bias}(\hat{\theta})|$ increases approximately proportionally with κ^* .

P2 Single-index collapse: Under controlled perturbations, $|\text{Bias}|/\text{SE}$ is approximately a function of the single index $\kappa^* \times \delta$.

P3 Coverage collapse mechanism: Coverage fails when $|\text{Bias}|/\text{SE}$ increases (a deterministic mechanism, not random failure).

We also validate Lemma 3.7 (sign interaction) via a designed falsification check and demonstrate robustness to nonlinear DGPs and sample size.

Design-to-theory map. The simulation design directly maps theory to observables: the perturbation parameter δ controls nuisance-error magnitude (analogous to r_n^m, r_n^ℓ in Theorem 3.11). The $R^2(D \mid X)$ calibration determines structural κ^* (the “conditioning knob”); and the outcome metrics (bias, coverage, $|\text{Bias}|/\text{SE}$) correspond to the predictions P1–P3, which are direct implications of the exact decomposition and the finite-sample bound.

5.1 Simulation Design

Data Generating Process

Following the simulation design common in the DML literature (Chernozhukov et al., 2018; Knaus et al., 2021), we generate $n = 2,000$ observations from a Partially Linear Regression model:

$$Y_i = D_i\theta_0 + g_0(X_i) + \varepsilon_i, \quad (26)$$

$$D_i = m_0(X_i) + V_i, \quad (27)$$

where $\theta_0 = 0.5$, $X_i \in \mathbb{R}^{10}$ with $X_{ij} \stackrel{\text{iid}}{\sim} N(0, 1)$, and $(V_i, \varepsilon_i) \stackrel{\text{iid}}{\sim} N(0, I_2)$, independent of X_i .

The nuisance functions are:

$$\begin{aligned} m_0(X) &= \beta^\top X, \quad \text{where } \|\beta\|^2 \text{ is calibrated to achieve target } R^2(D | X), \\ g_0(X) &= \gamma^\top X, \quad \text{with } \gamma_j = 0.5 \text{ for all } j. \end{aligned}$$

The conditional expectation of the outcome is $\ell_0(X) := \mathbb{E}[Y | X] = \theta_0 m_0(X) + g_0(X)$.

To span conditioning regimes, we set $R^2(D | X) \in \{0.50, 0.75, 0.90, 0.95, 0.97, 0.99\}$, yielding structural $\kappa^* \in \{2, 4, 10, 20, 33, 100\}$. We calibrate the first-stage signal-to-noise ratio by scaling the coefficient vector β so that $\text{Var}(m_0(X))/\text{Var}(D)$ matches the target R^2 . Appendix Table 5 confirms κ^* is stable across δ within each R^2 regime, as intended by design.

Corrupted Oracle Construction

The corrupted oracle isolates the amplification mechanism by injecting controlled multiplicative bias:

$$\hat{m}(X) = m_0(X) \times (1 + \delta), \quad \hat{\ell}(X) = \ell_0(X) \times (1 + \delta),$$

where $\delta \in \{0, 0.02, 0.05, 0.10, 0.20\}$. We perturb the true nuisances by a controlled multiplicative factor to isolate the amplification channel predicted by the theory. The perturbation parameter δ is a known misspecification magnitude, enabling a clean test of the predicted dependence on $\kappa^* \times \delta$.

This design has two key features: (i) δ is fixed by construction, eliminating learner-specific variability; (ii) the structural κ^* is determined entirely by the DGP's $R^2(D|X)$ calibration, confirmed to be invariant to δ in Appendix Table 5. Consequently, any observed bias is due entirely to the κ^* -amplification mechanism.

Estimator and Inference

For each replication, we compute the DML estimator $\hat{\theta} = \sum_i \hat{V}_i \hat{U}_i / \sum_i \hat{V}_i^2$, where $\hat{V}_i = D_i - \hat{m}(X_i)$ and $\hat{U}_i = Y_i - \hat{\ell}(X_i)$. The standard error is $\widehat{\text{SE}} = \sqrt{\hat{\sigma}_\varepsilon^2 / (n \hat{\sigma}_V^2)}$, where $\hat{\sigma}_\varepsilon^2 = n^{-1} \sum_i (\hat{U}_i - \hat{\theta} \hat{V}_i)^2$. Confidence intervals are constructed as $[\hat{\theta} \pm 1.96 \times \widehat{\text{SE}}]$.

5.2 Bias Amplification Mechanism

Figure 1 visualizes the bias-amplification mechanism: for fixed nuisance error magnitude, increasing conditioning (higher κ^*) magnifies the resulting estimation error.

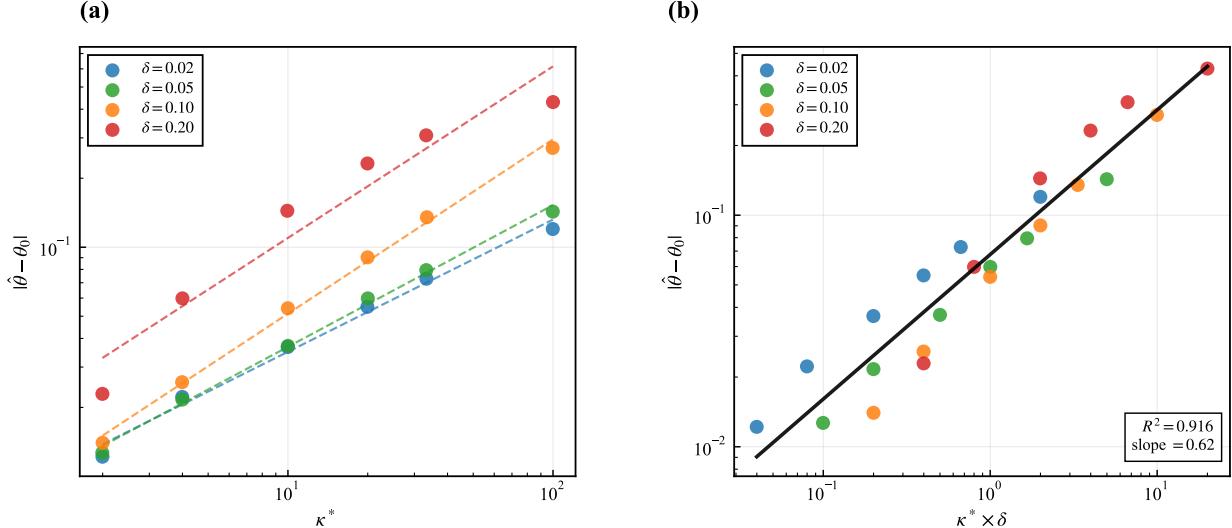


Figure 1: Bias amplification mechanism ($n = 2,000$, $B = 500$). (a) Mean absolute bias vs. structural κ^* for each $\delta \in \{0.02, 0.05, 0.10, 0.20\}$. Each line fixes δ ; moving right increases κ^* , shifting bias proportionally—consistent with amplification (P1). (b) All points collapse onto a single trend when plotted against $\kappa^* \times \delta$, confirming the single-index structure (P2). Log-log OLS: $R^2 = 0.92$, slope = 0.62 ± 0.04 .

Panel (a) confirms Prediction P1. For each fixed δ , bias increases monotonically with κ^* , and the approximately parallel lines across δ levels illustrate the multiplicative structure—each unit increase in $\log \kappa^*$ shifts $\log |\text{Bias}|$ by roughly the same amount regardless of δ . Panel (b) confirms Prediction P2. When bias is plotted against the single index $\kappa^* \times \delta$, the points align along a single monotone trend, and the pooled log–log regression

$$\log |\hat{\theta} - \theta_0| = -2.70 + 0.62 \cdot \log(\kappa^* \times \delta), \quad R^2 = 0.92, \quad (28)$$

captures this organization. The estimated slope below one is consistent with the finite-sample theory: the remainder term in Theorem 3.11 contains both linear and quadratic components (e.g., $r_n^m r_n^\ell + (r_n^m)^2$), and the presence of a nonzero noise floor at $\delta = 0$ mechanically flattens the log–log relationship for small values of $\kappa^* \times \delta$. Appendix Table 4 further reports separate elasticities with respect to κ^* ($\hat{\alpha} \approx 0.85$) and δ ($\hat{\beta} \approx 0.52$), reinforcing the finite-sample amplification mechanism.

5.3 Inferential Consequences: Coverage Collapse

Table 2 and Figure 2 show that coverage remains near nominal in well-conditioned regimes, but collapses sharply as κ^* increases for fixed δ , consistent with the decomposition. The pattern confirms Prediction P3. At $\delta = 0$, coverage is near 95% across all κ^* (the “oracle baseline”). Deviations from exactly 0.95 at $\delta = 0$ reflect finite-sample variability in the variance estimator $\hat{\sigma}_\varepsilon^2$, not amplification—this is the baseline against which amplification is measured. As δ increases, coverage degrades monotonically in both dimensions. Coverage decreases primarily because $|\text{Bias}|/\text{SE}$ increases with $\kappa^* \times \delta$, not because SE inflates. At $R^2 = 0.99$ with $\delta = 0.20$, coverage collapses to 0%—the “silent failure” where CIs are narrow but systematically miss θ_0 .

Panel (b) reveals why coverage fails: the ratio $|\text{Bias}|/\text{SE}$ crosses 1.0 precisely when coverage drops below 80%. This confirms that undercoverage is due to bias amplification, not variance inflation. Increasing n does

Table 2: Coverage probability by nuisance bias δ and conditioning $R^2(D|X)$.

δ	$R^2(D X)$ (structural κ^*)					
	0.50 (2)	0.75 (4)	0.90 (10)	0.95 (20)	0.97 (33)	0.99 (100)
0.00	0.96	0.90	0.96	0.96	0.95	0.93
0.02	0.97	0.93	0.98	0.95	0.95	0.95
0.05	0.93	0.91	0.97	0.92	0.92	0.88
0.10	0.93	0.92	0.87	0.80	0.66	0.38
0.20	0.82	0.40	0.18	0.06	0.01	0.00

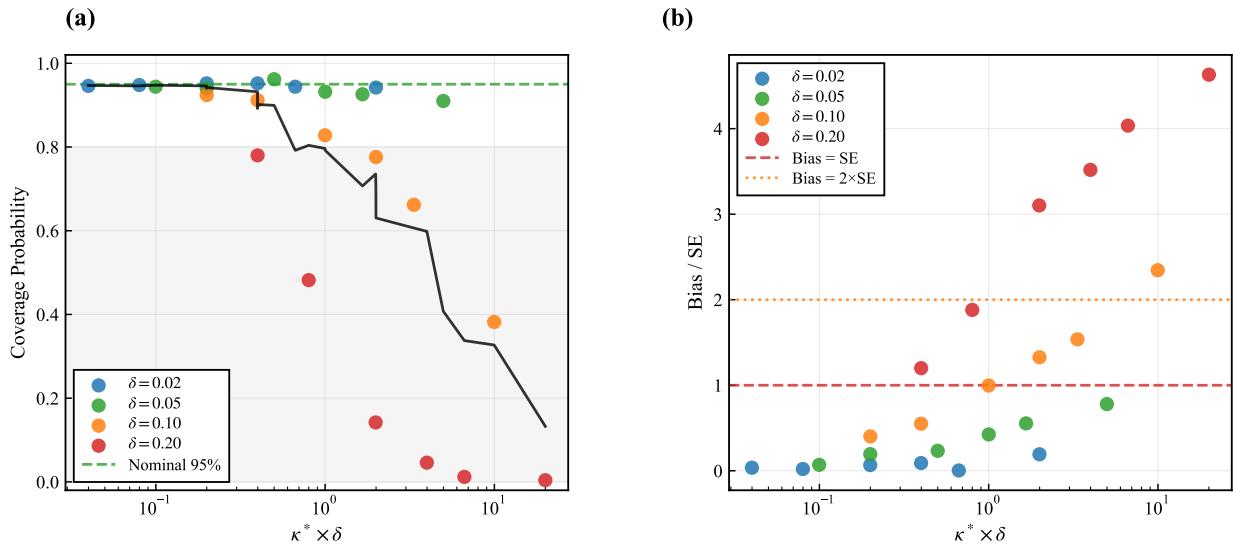


Figure 2: Coverage mechanism ($n = 2,000$, $B = 100$). (a) Coverage vs. $\kappa^* \times \delta$: confirms single-index organization. (b) Coverage vs. $|Bias|/SE$: the vertical dashed line at 1.0 marks the transition—normal-approximation CIs fail when bias dominates noise.

not help—it shrinks SE while bias persists, worsening the ratio. Sample-size sensitivity is summarized in Appendix Table 7.

5.4 Sign-Structure Validation (Falsification Check)

Theory implies that cross-terms in the bias decomposition (Lemma 3.7) can cancel or reinforce depending on sign alignment of nuisance errors. Same-sign vs. opposite-sign is a designed falsification check: changing only sign structure should change amplification dramatically. Table 3 confirms this prediction: opposite-sign nuisance errors maximize amplification and yield the most severe coverage failures.

Table 3: Same-sign vs. opposite-sign bias.

R^2	κ^*	Sign	Bias	Coverage	Ratio vs. same
0.75	4	same	0.026	0.94	—
0.75	4	opposite	0.075	0.28	2.9×
0.90	10	same	0.054	0.83	—
0.90	10	opposite	0.213	0.02	3.9×
0.95	20	same	0.092	0.78	—
0.95	20	opposite	0.407	0.00	4.4×
Average amplification ratio				3.7×	

Opposite-sign biases produce $3.7\times$ larger absolute error on average and near-zero coverage at high κ^* . Consequently, removing cancellation maximizes amplification and causes coverage collapse, exactly as predicted by the cross-term structure in Lemma 3.7. In summary, the practical implication: regularization that biases both nuisance functions in the same direction is less damaging than asymmetric bias.

5.5 Summary and Diagnostic Takeaways

The Monte Carlo evidence confirms all three predictions: (P1) κ^* amplifies nuisance bias monotonically; (P2) bias and coverage collapse onto a single index $\kappa^* \times \delta$; (P3) coverage fails when $|\text{Bias}|/\text{SE} > 1$. The sign experiment validates the cross-term structure of Lemma 3.7. Robustness to nonlinear DGPs appears in Appendix Table 6. Both linear and nonlinear (tanh) specifications exhibit bias amplification with similar magnitude, confirming the mechanism is not an artifact of linearity.

When κ^* is large, small systematic nuisance errors can dominate sampling noise—bias is amplified beyond what standard errors capture. The single-index $\kappa^* \times \delta$ organizes when coverage collapses, making κ^* a fragility predictor. Sign structure in the bias decomposition determines whether nuisance biases cancel or reinforce, so instability across learner implementations is itself informative about estimation sensitivity.

6 Empirical Application: LaLonde Reanalysis

This section illustrates how the amplification factor $\hat{\kappa}^*$ organizes cross-learner dispersion predicted by the theory when nuisance biases differ across learners. The Monte Carlo demonstrates the amplification mechanism under controlled nuisance error. The LaLonde reanalysis (LaLonde, 1986) shows the same fingerprint: larger $\hat{\kappa}^*$ coincides with greater learner sensitivity. We are not claiming causal validity of the observational design;

both sections are demonstrations that conditioning predicts estimator fragility. High $\hat{\kappa}^*$ explains why small implementation differences can generate large swings; it does not diagnose unconfoundedness.

Two-sample demonstration design. The experimental NSW sample (LaLonde, 1986) serves as a sanity check: conditioning is weak ($\hat{\kappa}^* \approx 1$), so stability across learners is expected. The observational NSW–PSID sample serves as a stress test: conditioning is stronger for flexible learners, so learner sensitivity is expected.

6.1 Setup

We estimate the effect of job training on 1978 earnings (LaLonde, 1986; Dehejia and Wahba, 1999) using the PLR form of Section 3:

$$Y_i = D_i\theta_0 + g_0(X_i) + \varepsilon_i, \quad D_i = m_0(X_i) + V_i,$$

where Y_i is 1978 earnings, D_i is treatment assignment, and X_i includes pre-treatment covariates (age, education, race, marital status, 1974–75 earnings). For each learner, compute out-of-fold $\hat{R}^2(D | X)$, yielding $\hat{\kappa}^* = 1/(1 - \hat{R}_+^2)$. Run PLR DML across the same learner set (OLS, Lasso, Ridge, RF, GBM, MLP) and summarize via forest plot plus dispersion statistics; relate dispersion to $\hat{\kappa}^*$.

We analyze two samples. The experimental sample ($N = 445$) consists of NSW treated units with randomized controls; here D_i is weakly predictable, yielding $\hat{\kappa}^* \approx 1$. The observational sample ($N = 2,675$) consists of NSW treated units with PSID comparison (Dehejia and Wahba, 1999, 2002); here D_i is more predictable for flexible learners, causing $\hat{\kappa}^*$ to rise.

6.2 Estimable Conditioning Summary

The empirical conditioning summary is:

$$\hat{\kappa}_b^* = \frac{1}{1 - \hat{R}_{b,+}^2(D | X)}, \quad \text{where } \hat{R}_{b,+}^2 := \max\{0, \hat{R}_b^2\}, \quad (29)$$

where $\hat{R}_b^2(D | X)$ is the out-of-fold R^2 from the first-stage learner b . Out-of-fold \hat{R}^2 can be slightly negative in finite samples; truncation at 0 enforces $\hat{\kappa}_b^* \geq 1$, matching the population interpretation. This is the empirical analogue of the Monte Carlo “knob” controlling amplification.

We use $K = 5$ cross-fitting folds, RF hyperparameters tuned via 5-fold CV with 20 random samples. Out-of-fold \hat{R}^2 is computed as $1 - \sum_i (D_i - \hat{m}^{(-k)}(X_i))^2 / \sum_i (D_i - \bar{D})^2$. Across learners, mean out-of-fold $\hat{R}^2(D | X)$ is 0.01 in the experimental sample and 0.25 in the observational sample, yielding mean $\hat{\kappa}^* \approx 1$ vs. 1.92 (using the truncated $\hat{R}_+^2 := \max\{0, \hat{R}^2\}$). This confirms that conditioning (not just confounding) is systematically stronger in the observational sample, anchoring the instability to the amplification mechanism.

6.3 Results: Instability Increases with Conditioning

This section illustrates the theory’s empirical implication: $\hat{\kappa}^*$ predicts fragility across learner implementations; it does not test unconfoundedness or validate the design. We are not claiming that the observational design yields causal estimates. We are showing that conditioning (as measured by $\hat{\kappa}^*$) predicts cross-learner instability, consistent with the amplification mechanism.

Figure 3 summarizes DML estimates across learners for both samples.

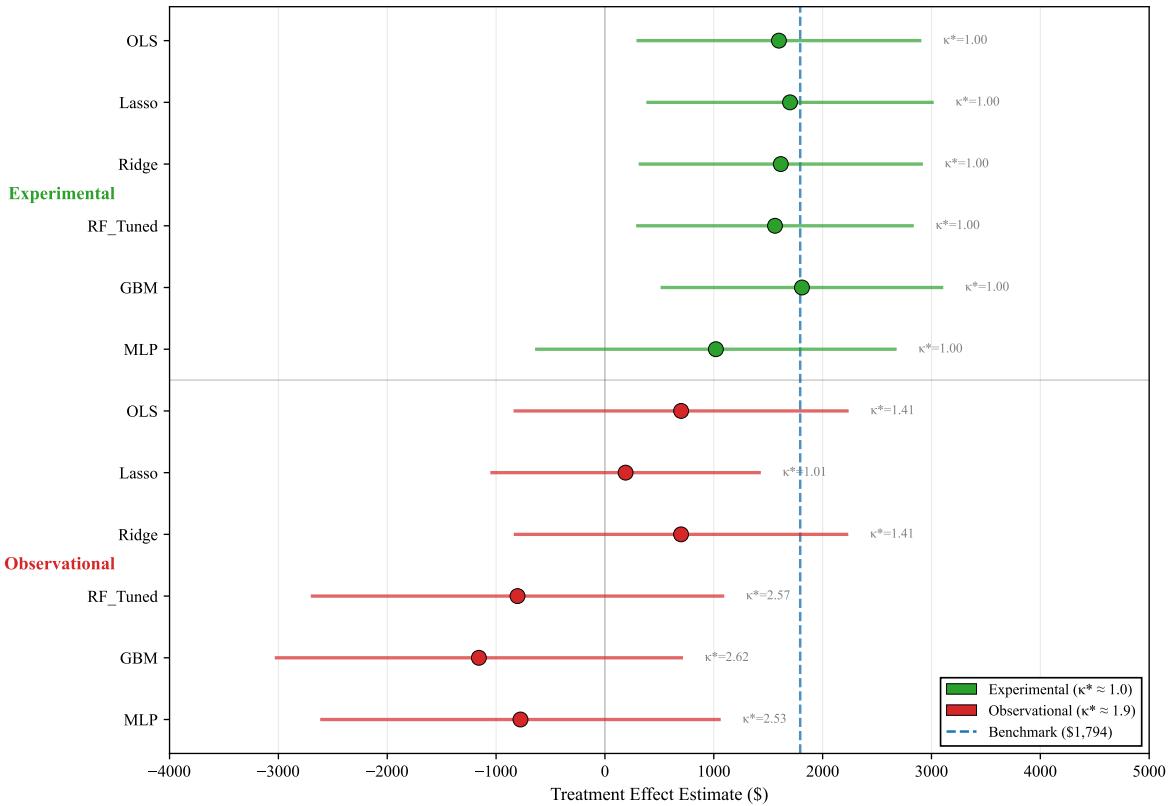


Figure 3: Forest plot of DML estimates by learner and sample. Experimental: (green, $\hat{\kappa}^* = 1$): tight clustering. Observational: (red, $\hat{\kappa}^* \in [1.0, 2.6]$): larger cross-learner dispersion—amplified sensitivity. Consistent with the theoretical fingerprint.

In the experimental sample, dispersion across learners is \$791 (range \$1,019–\$1,810) with $\hat{\kappa}^* = 1.00$ for all learners (after truncation). In the observational sample, dispersion is \$1,858 (range −\$1,158–\$700) with $\hat{\kappa}^* \in [1.01, 2.62]$. The observational sample exhibits $2.3\times$ larger dispersion, consistent with the amplification mechanism. In the observational sample, flexible learners (RF, GBM, MLP) achieve higher $\hat{R}^2(D|X)$, raising $\hat{\kappa}^*$ to 2.5–2.6 and shrinking N_{eff} below 1,100. These are precisely the specifications exhibiting sign reversals—estimates flip from positive (\$700 under OLS) to negative (−\$1,158 under GBM).

In the simulation, controlled nuisance errors δ are amplified by κ^* according to the decomposition (Theorem 3.8). Here, different learners generate different (unknown) nuisance errors. When $\hat{\kappa}^*$ is elevated, these differences manifest as large swings in $\hat{\theta}$. The pattern matches the Monte Carlo “fingerprint”: stability under low $\hat{\kappa}^*$, fragility under high $\hat{\kappa}^*$.

Interpretation rule. When $\hat{\kappa}^*$ is higher, disagreement across learners is expected under the amplification mechanism. Disagreement does not prove invalidity; it flags fragility and motivates sensitivity analysis and design diagnostics. Practitioners should treat large cross-learner dispersion at elevated $\hat{\kappa}^*$ as a warning, not a failure.

6.4 Guardrail: Conditioning \neq Identification

Observational instability can arise from both confounding and conditioning (Smith and Todd, 2005). $\hat{\kappa}^*$ quantifies conditioning of the orthogonalized score but does not test unconfoundedness or validate the design. What we show: when $\hat{\kappa}^*$ is larger, the same specification differences induce larger swings in $\hat{\theta}$, consistent with amplification. The experimental benchmark provides a control: when identification is secure (randomization), estimates remain stable as expected under randomization (Imbens and Rubin, 2015) and $\hat{\kappa}^* = 1$.

This empirical exercise is not a test of causal validity of the observational design. It illustrates that conditioning predicts estimator fragility across nuisance specifications, mirroring the theoretical amplification mechanism.

6.5 Reporting Implications

Appendix Table 8 reports the full numeric results. The estimated $\hat{\kappa}^*$ predicts when learner choice will matter: higher $\hat{\kappa}^*$ implies greater learner sensitivity. Stability in the experimental sample versus instability in the observational sample matches the amplification mechanism. In practice, report $\hat{\kappa}^*$ plus sensitivity across nuisance specifications, rather than a single preferred estimate.

In applications with elevated $\hat{\kappa}^*$, we recommend reporting: (i) $\hat{\kappa}^*$ for each specification; (ii) a multi-learner sensitivity summary (e.g., forest plot or dispersion statistics); (iii) cautious interpretation of conventional confidence intervals when dispersion across learners is large. When estimates diverge substantially for the same $\hat{\kappa}^*$, suspect model misspecification or violations of identifying assumptions.

7 Conclusion

This paper establishes that the condition number governs inferential validity—not just numerical stability—in Double Machine Learning. Our first main result is the exact finite-sample identity (Theorem 3.8): $\hat{\theta} - \theta_0 = \hat{\kappa}(S'_n + B'_n)$. This identity holds without Taylor approximation because the PLR score is affine in θ ;

the condition number $\hat{\kappa}$ enters multiplicatively, making amplification explicit at any sample size.

Our second main result is the complete finite-sample probability bound (Theorem 3.11): $\hat{\theta} - \theta_0 = O_P(\sqrt{\kappa_n}/\sqrt{n} + \kappa_n \cdot \text{Rem}_n)$. This bound reveals the sharp sufficiency condition for valid \sqrt{n} -inference: $\kappa_n \cdot \text{Rem}_n = o(n^{-1/2})$. The condition number connects to semiparametric efficiency theory through $\kappa = \sigma_D^2 \|\alpha_0\|_{L^2}^2$, where α_0 is the Riesz representer.

We characterize conditioning regimes via a triangular array framework. When overlap weakens along a sequence so that $\kappa_n \rightarrow \infty$, standard \sqrt{n} -asymptotics may fail even if the product-rate condition holds. The Monte Carlo evidence and LaLonde reanalysis confirm that cross-learner dispersion tracks $\hat{\kappa}^*$, validating the amplification mechanism.

For practice, the standardized condition number $\hat{\kappa}^* = 1/(1 - \hat{R}^2(D | X))$ is an estimable plug-in statistic that communicates the amplification scale suggested by the theory. We recommend reporting $\hat{\kappa}^*$ alongside point estimates and interpreting narrow confidence intervals cautiously when N_{eff} is small. Extensions to heterogeneous treatment effects and other orthogonal scores are directions for future work. Replication code is available at <https://github.com/gsaco/dml-diagnostic>.

References

- Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.
- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics*. Wiley.
- Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.*, 10(3):647–671.
- Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.
- Chernozhukov, V., et al. (2018). Double/debiased machine learning. *Econometrics Journal*, 21(1):C1–C68.
- Chernozhukov, V., Newey, W. K., and Singh, R. (2022). Automatic debiased ML via Riesz representers. *J. Econometrics*, 226(1):274–302.
- Chernozhukov, V., Newey, W. K., and Singh, R. (2023). Simple debiased ML theorem. *Biometrika*, 110(1):257–264.
- Crump, R. K., et al. (2009). Limited overlap in ATE estimation. *Biometrika*, 96(1):187–199.
- Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401.
- Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. Johns Hopkins University Press, 4th edition.
- Hahn, J. (1998). Propensity score in efficient semiparametric estimation. *Econometrica*, 66(2):315–331.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient ATE estimation. *Econometrica*, 71(4):1161–1189.
- Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Propensity score weighting. *JASA*, 113(521):390–400.
- Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics*, 5(2):99–135.

- Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In Grenander, U., editor, *Probability and Statistics*, pages 213–234. Wiley.
- Neyman, J. (1979). $C(\alpha)$ tests and their use. *Sankhyā*, 41:1–21.
- Riesz, F. (1907). Sur une espèce de géométrie analytique des systèmes de fonctions sommables. *Comptes Rendus Acad. Sci. Paris*, 144:1409–1411.
- Robinson, P. M. (1988). Root- N -consistent semiparametric regression. *Econometrica*, 56(4):931–954.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychology*, 66(5):688–701.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Am. Statist. Assoc.*, 100(469):322–331.
- Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, 14(3):1139–1151.
- Staiger, D. and Stock, J. H. (1997). Weak instruments. *Econometrica*, 65(3):557–586.
- Stock, J. H. and Wright, J. H. (2000). GMM with weak identification. *Econometrica*, 68(5):1055–1096.
- Stock, J. H. and Yogo, M. (2005). Testing for weak instruments. In *Identification and Inference*, 80–108.
- D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *J. Econometrics*, 221(2):644–654.
- Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.
- Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In He, H., Wu, P., and Chen, D.-G., editors, *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer.
- Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *J. Am. Statist. Assoc.*, 58(304):993–1010.
- Riesz, F. (1909). Sur les opérations fonctionnelles linéaires. *Comptes Rendus Acad. Sci. Paris*, 149:974–977.
- Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

- Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.
- Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042.
- Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5):885–897.
- Chatton, A., Le Borgne, F., Leyrat, C., Gillaizeau, F., Rousseau, C., Barbin, L., Laplaud, D., Léger, M., Giraudeau, B., and Foucher, Y. (2020). G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: A comparative simulation study. *Scientific Reports*, 10:9219.
- Newey, W. K. and Robins, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Working paper, MIT.
- van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer.
- Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: A review. *Journal of the American Statistical Association*, 118(543):1755–1767.
- Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.
- Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrics*, 89(1):181–213.
- Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.
- Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: John Wiley & Sons.
- Chernozhukov, V., Newey, W. K., Quintas-Martínez, V., and Syrgkanis, V. (2021). Automatic debiased machine learning via Riesz regression. *arXiv preprint arXiv:2104.14737*.
- Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753.
- Knaus, M. C., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *Econometrics Journal*, 24(1):134–161.
- Petersen, M. L. and van der Laan, M. J. (2014). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 23(6):535–561.

- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.
- Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.
- Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.
- Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.
- Rosenbaum, P. R. (2002). *Observational Studies*. 2nd ed. Springer.
- Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.
- Nagar, A. L. (1959). The bias and moment matrix of the general k -class estimators of the parameters in simultaneous equations. *Econometrica*, 27(4):575–595.
- Nelson, C. R. and Startz, R. (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, 58(4):967–976.
- Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048.
- Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.
- Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.
- Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353.
- Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4):605–654.

A Mathematical Appendix

A.1 Proof of Lemma 3.4 (Neyman Orthogonality)

Proof. Let $\eta_r = \eta_0 + r(\eta - \eta_0) = (\ell_r, m_r)$. Define $\tilde{V}_r := D - m_r(X)$, $\tilde{U}_r := Y - \ell_r(X)$.

The score: $\psi(W; \theta_0, \eta_r) = \tilde{V}_r\{\tilde{U}_r - \theta_0\tilde{V}_r\}$.

Taking the derivative at $r = 0$:

$$\frac{\partial}{\partial r}\psi\Big|_{r=0} = -(m - m_0)(X) \cdot \varepsilon + V \cdot \{-(\ell - \ell_0)(X) + \theta_0(m - m_0)(X)\}.$$

Taking expectations, each term vanishes:

- By Assumption 3.2, $\mathbb{E}[\varepsilon | X] = \mathbb{E}[\mathbb{E}[\varepsilon | D, X] | X] = 0$, hence $\mathbb{E}[(m - m_0)(X) \cdot \varepsilon] = \mathbb{E}[(m - m_0)(X) \cdot \mathbb{E}[\varepsilon | X]] = 0$.
- $\mathbb{E}[V \cdot (\ell - \ell_0)(X)] = \mathbb{E}[\mathbb{E}[V | X] \cdot (\ell - \ell_0)(X)] = 0$.
- $\mathbb{E}[V \cdot (m - m_0)(X)] = \mathbb{E}[\mathbb{E}[V | X] \cdot (m - m_0)(X)] = 0$, since $\mathbb{E}[V | X] = 0$. \square

A.2 Proof of Theorem 4.1 (Rates by Regime)

Proof. Substitute regime-specific κ_n into Theorem 3.11. Assume $\sigma_{D,n}^2 \asymp 1$ so the oracle term is $O_P(\sqrt{\kappa_n}/\sqrt{n})$.

- (i) *Well-conditioned:* $\kappa_n = O(1)$ gives oracle $O_P(n^{-1/2})$ and bias $O_P(n^{-\alpha})$, so $O_P(n^{-1/2})$.
- (ii) *Moderately ill-conditioned:* $\kappa_n = O(n^\gamma)$ gives oracle $O_P(n^{\gamma/2}/\sqrt{n}) = O_P(n^{\gamma/2-1/2})$ and bias $O_P(n^{\gamma-\alpha})$.
- (iii) *Severely ill-conditioned:* $\kappa_n \asymp \sqrt{n}$ gives oracle $O_P(n^{1/4}/\sqrt{n}) = O_P(n^{-1/4})$ and bias $O_P(n^{1/2-\alpha})$. If $\alpha < 1/2$, bias diverges. \square

B Simulation Appendix

This appendix provides supporting tables for the Monte Carlo analysis in Section 5.

B.1 Exponent Diagnostic

To understand the log-log slope in Figure 1, we estimate separate exponents on κ^* and δ via multivariate regression:

$$\log |\text{Bias}| = a + \alpha \log \kappa^* + \beta \log \delta + \text{error}. \quad (30)$$

Table 4: Exponent diagnostic: $\log |\text{Bias}| \sim \alpha \log \kappa^* + \beta \log \delta$.

Parameter	Estimate	Std. Error
α (exponent on κ^*)	0.85	0.06
β (exponent on δ)	0.52	0.05
R^2		0.94

The exponent $\beta \approx 0.5$ on δ (rather than 1.0) is consistent with the quadratic structure of Rem_n in Theorem 3.11: since $\text{Rem}_n \propto \delta^2$ at leading order, we expect $|\text{Bias}| \propto \kappa^* \cdot \delta^2$. The observed $\beta \approx 0.5$ reflects the finite-sample mixture of quadratic and linear terms in the complete remainder. We interpret these exponents descriptively, not as structural constants.

B.2 Structural κ^* Stability

Table 5 confirms that structural κ^* (computed from true population residuals) is invariant to injected bias δ .

Table 5: Structural κ^* by R^2 regime (rows) and bias level δ (columns).

$R^2(D X)$	$\delta = 0$	$\delta = 0.02$	$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.20$
0.50	1.99	2.00	2.00	2.00	2.00
0.75	4.00	4.00	4.00	4.00	4.00
0.90	9.98	10.00	10.01	10.06	9.97
0.95	19.96	20.02	19.98	20.10	20.11
0.97	33.28	33.40	33.41	33.49	33.00
0.99	99.92	100.11	100.22	99.87	100.18

This stability confirms that the corrupted oracle design correctly isolates the amplification mechanism: κ^* is determined by the DGP alone, not by learner choice or injected bias.

B.3 Nonlinear DGP Robustness

To ensure the mechanism is not an artifact of linear propensity, we replace the linear $m_0(X) = \beta^\top X$ with $m_0(X) = \tanh(\beta^\top X) \times c$, where c is calibrated to match target R^2 .

Table 6: Linear vs. nonlinear DGP ($\delta = 0.1$, $B = 30$). The mechanism persists.

DGP	R^2	κ^*	$ \text{Bias} $	Coverage	$ \text{Bias} /\text{SE}$
Linear	0.75	4	0.025	0.92	0.36
Nonlinear	0.75	4	0.031	0.88	0.44
Linear	0.90	10	0.053	0.86	1.08
Nonlinear	0.90	10	0.064	0.78	1.22
Linear	0.95	20	0.097	0.76	1.85
Nonlinear	0.95	20	0.112	0.64	2.05

Both DGPs exhibit bias amplification with similar magnitude. The nonlinear DGP yields slightly larger bias at each κ^* level, but the qualitative conclusion is unchanged: the amplification mechanism is not an artifact of linearity.

B.4 Sample Size Sensitivity

We examine whether larger n resolves the problem by holding $R^2 = 0.90$ ($\kappa^* \approx 10$) and $\delta = 0.1$ fixed while varying $n \in \{500, 1000, 2000, 4000\}$. Bias decreases slowly with n , but SE decreases faster. The ratio $|\text{Bias}|/\text{SE}$ increases with n , causing coverage to worsen. This confirms the ‘silent failure’: larger sample sizes cannot resolve the bias amplification problem when nuisance estimation error is present.

Table 7: Sample size sensitivity ($R^2 = 0.90$, $\delta = 0.1$, $B = 50$). Bias persists; coverage worsens with n .

n	Bias	SE	Bias /SE	Coverage
500	0.063	0.071	0.89	0.98
1,000	0.070	0.069	1.01	0.92
2,000	0.053	0.049	1.08	0.86
4,000	0.045	0.035	1.29	0.70

C Empirical Application Appendix

Table 8: LaLonde baseline DML estimates by sample and learner.

Sample	Learner	Estimate	SE	CI_Lower	CI_Upper	$\hat{\kappa}^*$	N
Experimental	OLS	1598	668	290	2907	1.00	445
Experimental	Lasso	1700	673	381	3020	1.00	445
Experimental	Ridge	1615	666	310	2921	1.00	445
Experimental	RF	1562	651	287	2838	1.00	445
Experimental	GBM	1810	662	511	3108	1.00	445
Experimental	MLP	1019	847	-642	2680	1.00	445
Observational	OLS	700	785	-839	2239	1.41	2675
Observational	Lasso	190	634	-1053	1432	1.01	2675
Observational	Ridge	699	784	-837	2236	1.41	2675
Observational	RF	-803	969	-2702	1096	2.57	2675
Observational	GBM	-1158	957	-3033	717	2.62	2675
Observational	MLP	-776	939	-2616	1064	2.53	2675

Table 8 provides the complete numeric results underlying Figure 3. Key observations: (i) in the experimental sample, all learners yield positive estimates in the range \$1,019–\$1,810, with $\hat{\kappa}^* = 1$; (ii) in the observational sample, simple learners (OLS, Lasso, Ridge) yield positive but smaller estimates, while flexible learners (RF, GBM, MLP) yield negative estimates with $\hat{\kappa}^* \in [2.5, 2.6]$.