Ill-Conditioned Orthogonal Scores in Double Machine Learning

Gabriel Saco*

January 2, 2026

Abstract

Double Machine Learning is often justified by nuisance-rate conditions, yet finite-sample reliability also depends on the well-conditioning of the orthogonal scores. However, this is only assumed but not tracked. Our main result is an exact identity for the cross-fitted PLR-DML estimator, which holds without Taylor approximation. From the exact identity, we derive a stochastic-order bound, implying a sufficiency requirement for $\sqrt{n}$-inference involving both oracle noise and an amplified nuisance remainder. We further connect the terms in the bound to semiparametric efficiency geometry via the Riesz representer and use a triangular-array framework to characterize regimes as residual treatment variation weakens. These results motivate a simple diagnostic summarizing the amplification scale. Rather than propose universal thresholds, we recommend routinely reporting the diagnostic alongside cross-learner dispersion of DML estimates as a fragility assessment.

Keywords: Double machine learning; orthogonal scores; weak overlap; finite-sample inference; Riesz representer.

JEL Codes: C14, C21, C55.

*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) permits the application of machine learning methods for nuisance estimation, while preserving $\sqrt{n}$-inference for low-dimensional treatment effects. This holds because Neyman-orthogonal scores reduce first-order sensitivity to nuisance estimation errors and cross-fitting avoids own-observation bias (Chernozhukov et al., 2018; Kennedy, 2024). As a result, the corresponding sufficient conditions are typically stated as rate conditions, rather than as parametric smoothness restrictions. In the canonical partially linear regression (PLR) model, for instance, the classical product-rate requirement is $r_n^m r_n^\ell = o_P(n^{-1/2})$, where r_n^m, r_n^ℓ denote L^2 nuisance-estimation rates (Chernozhukov et al., 2018).

These conditions control the entry of nuisance estimation error into the orthogonal score. However, these rate conditions do not track how residual treatment variation after conditioning (the PLR analogue of overlap/positivity) directly impacts finite-sample reliability (Khan and Tamer, 2010; D'Amour et al., 2021). Standard DML theory often relies on regularity conditions to keep this residual variation bounded away from zero. We instead make the finite-sample sensitivity induced by low residual treatment variation explicit in the canonical PLR-DML setting through an exact identity and a stochastic bound.[1] A simple reportable out-of-fold diagnostic follows as a reporting implication.

An amplification mechanism is transparent in PLR. In this case, the sensitivity of the empirical orthogonal score to the treatment effect θ depends on the empirical Jacobian $\widehat{J}_\theta = -\widehat{\sigma}_V^2$, where V is the residualized treatment. If $\widehat{\sigma}_V^2$ is small, the score is nearly flat in the θ direction. Small perturbations of the empirical score equation therefore create large perturbations of $\widehat{\theta}$. This corresponds to ill-conditioning in the numerical sense (Golub and Van Loan, 2013). We capture this channel through $\kappa := \sigma_D^2/\sigma_V^2 = 1/(1 - R^2(D \mid X))$. It scales the impact of any residual bias left after orthogonalization and cross-fitting. While standard errors reflect increased sampling variability when σ_V^2 is small, they do not consider the bias-amplification induced by large κ. Coverage can fail as a result, even when reported standard errors are only moderately inflated.

[1] Our main results are for PLR-DML, where the score is affine in θ and the identity is exact. Appendix A.3 shows the generic orthogonal-score expansion for nonlinear scores, where a Taylor/linearization step is required.

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. In other words, 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 stochastic-order bound:

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

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 the usual $\sqrt{n}$-approximation (oracle term dominates the remainder) is $\kappa_n \cdot \mathrm{Rem}_n = o(n^{-1/2})$. When overlap is stable so that κ_n is bounded, this reduces to familiar remainder-rate restrictions. If overlap is weak and κ_n is large, the same nuisance errors can be amplified into first-order bias.

Two theoretical connections improve 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 $\kappa_n \to \infty$. This shows 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 condition number,

$$\widehat{\kappa}_{\mathrm{oof}} = \frac{1}{1 - \widehat{R}_+^2(D \mid X)}, \qquad \widehat{R}_+^2 := \max\{0, \widehat{R}_{\mathrm{oof}}^2\},$$

a scale-invariant measure of residual treatment variation, where $\widehat{R}_{\mathrm{oof}}^2$ is an out-of-fold predictive R^2. When multiple first-stage learners are compared, $\widehat{\kappa}_{\mathrm{oof}}$ is computed learner-by-learner.[2] This operationalizes the overlap/conditioning component that enters the bound and

[2]Out-of-fold R^2 avoids in-sample overfitting bias and aligns the conditioning diagnostic with the cross-fitting used in DML. In PLR, the amplification factor in the bound satisfies $\kappa = \sigma_D^2/\sigma_V^2 = 1/\{1 - R^2(D \mid X)\}$, so $\widehat{\kappa}_{\mathrm{oof}} = 1/\{1 - \widehat{R}_+^2\}$ is a plug-in proxy for the same conditioning object.

helps interpret cross-learner sensitivity predicted by the theory. Importantly, $\widehat{\kappa}_{\text{oof}}$ concerns the conditioning of the orthogonalized score, but it is not a test of unconfoundedness and does not by itself validate causal identification. Analogous to weak instruments, where weak first-stage amplifies bias and motivates strength diagnostics (Staiger and Stock, 1997; Stock and Yogo, 2005), limited residual treatment variation in DML implies large κ, which 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 stochastic-order 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 Work

We study semiparametric causal inference with machine-learned nuisance functions. Chernozhukov et al. (2018) formalize 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). Subsequent work develops distributional refinements for orthogonal-score estimators (Chernozhukov et al., 2023). Complementing these refinements, Kennedy (2024) reviews influence-function–based semiparametric inference (including cross-fitted one-step/TMLE constructions and doubly robust structure) and Chernozhukov et al. (2021) develops a geometric debiasing view via Riesz representers. Our focus is on an additional, separate stability component of the score equation that is typically maintained as a background regularity.

Standard orthogonal-score DML analyses pair nuisance-rate conditions with a standing nondegeneracy requirement ensuring the score equation is well posed. In PLR, this amounts to requiring nontrivial residualized treatment variation so the score Jacobian is not close to singular (Chernozhukov et al., 2018). When this curvature deteriorates, inference becomes a weak-identification/ill-conditioning problem. Estimating equations flatten and small singular values act as error amplifiers (Kaji, 2021; Han and McCloskey, 2019; Breunig et al., 2020). Yet in applied DML this stability component is rarely tracked or reported alongside the final estimate. We make conditioning an explicit channel in PLR-DML by isolating a

condition-number factor κ that governs amplification of the orthogonal-score remainder as residual treatment variation shrinks. This links overlap-driven fragility in DML to classical collinearity diagnostics (Belsley et al., 1980).

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), 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). In particular, Petersen et al. (2012) discusses positivity violations and provides different approaches for practitioners. We complement this literature by showing that, in PLR-DML, weak overlap operates through residualized treatment variation and thus directly inflates κ, potentially turning otherwise second-order remainder terms into first-order finite-sample distortions.

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). Conceptually, our diagnostic 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). In the same spirit, we propose routinely reporting $\widehat{\kappa}_{\text{oof}}$, computed from out-of-fold prediction of D using X, as a simple diagnostic of fragility in learned causal estimators.

3 Theoretical Results

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 $\widehat{\theta} - \theta_0$, and (iii) a stochastic-order 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: $\mathbb{E}[D \mid X]$
$\ell_0(X)$	Outcome regression: $\mathbb{E}[Y \mid X] = \theta_0 m_0(X) + g_0(X)$
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), $\mathrm{Var}(D)$ (total treatment variance)
κ	Condition number: $\kappa = \sigma_D^2/\sigma_V^2 = 1/(1 - R^2(D \mid X))$
$\widehat{\kappa}_{\text{oof}}$	Out-of-fold diagnostic: $\widehat{\kappa}_{\text{oof}} := 1/(1 - \widehat{R}^2_+(D \mid X))$
κ_n	Population condition number along a triangular array sequence
$\widehat{\kappa}$	Sample estimate of κ from cross-fitted residuals
r_n^m, r_n^ℓ	L^2 convergence rates of nuisance estimators $\widehat{m}, \widehat{\ell}$
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)
S_n', B_n'	Standardized versions: $S_n' := S_n/\widehat{\sigma}_D^2$ and $B_n' := B_n/\widehat{\sigma}_D^2$
α_0	Riesz representer: V/σ_V^2 (Theorem 3.6)

Notation convention. We write κ for the population condition number under a fixed data-generating law P. When considering a triangular-array sequence $(P_n)_{n\geq 1}$ (used to formalize weakening overlap), we write κ_n for the corresponding population condition number under P_n. Throughout, P_n denotes the data-generating law at sample size n (not the empirical measure). Empirical averages are written explicitly as $n^{-1}\sum_{i=1}^n(\cdot)$.

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 $\kappa = \sigma_D^2/\sigma_V^2 = 1/(1-R^2(D \mid X)) = \sigma_D^2\|\alpha_0\|_{L^2}^2$, where α_0 is the Riesz representer (Theorem 3.6).
2. **Result 2 (Exact identity).** $\widehat{\theta} - \theta_0 = \widehat{\kappa}(S_n' + B_n')$, an algebraic identity without Taylor remainder (Theorem 3.8).

3. **Result 3 (Stochastic bound).** $\widehat{\theta} - \theta_0 = O_P(\sqrt{\kappa_n}/\sqrt{n} + \kappa_n \cdot \mathrm{Rem}_n)$ (Theorem 3.11). This converts the identity into an operational stochastic-order bound.[3]
4. **Sufficiency condition.** A sufficient condition for valid $\sqrt{n}$-inference beyond oracle noise is $\kappa_n \cdot \mathrm{Rem}_n = o(n^{-1/2})$.

Diagnostic implication. Results 1–4 motivate reporting an estimable proxy for κ based on out-of-fold treatment predictability:

$$\widehat{\kappa}_{\mathrm{oof}} := \frac{1}{1 - \widehat{R}^2_+(D \mid X)}, \qquad \widehat{R}^2_+ := \max\{0, \widehat{R}^2_{\mathrm{oof}}\},$$

so that $\widehat{\kappa}_{\mathrm{oof}} \geq 1$. A large $\widehat{\kappa}_{\mathrm{oof}}$ indicates limited residual treatment variation and, by Result 3, a regime where even small orthogonal-score remainders can be amplified.[4]

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 $f : \mathcal{X} \to \mathbb{R}$, 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 $\sigma^2_{V,n}$. In the classical fixed-DGP case $P_n \equiv P$, we drop the subscript and write σ^2_V.

We employ standard stochastic order notation under P_n (or P in the fixed-DGP case). We write $Z_n = O_P(a_n)$ if for every $\epsilon > 0$ there exists $M < \infty$ such that

$$\limsup_{n \to \infty} \mathbb{P}\big(|Z_n/a_n| > M\big) \leq \epsilon,$$

[3] We keep the main bound in stochastic-order form because fully nonasymptotic Gaussian approximation requires additional constants and concentration tools, such as finite-sample analyses for sample-splitting/DML-type estimators (Quintas-Martinez, 2022).

[4] Out-of-sample R^2 can be negative because it is defined relative to the sample mean benchmark; truncation avoids reporting $\kappa_{\mathrm{oof}} < 1$ due to finite-sample noise (Kvalseth, 1985).

and we write $Z_n = o_P(a_n)$ if $Z_n/a_n \xrightarrow{p} 0$.

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 $\sigma_{V,n}^2 \to 0$ via the triangular array setup.

3.2 The Partially Linear Regression Model

The PLR model (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, $g_0 : \mathcal{X} \to \mathbb{R}$ is a nuisance function, $m_0(X) := \mathbb{E}[D \mid X]$ is the treatment regression (conditional mean), and $V := D - m_0(X)$ is the treatment residual satisfying $\mathbb{E}[V \mid X] = 0$ by construction.

Remark 3.1 (Terminology). When D is binary, $m_0(X) = \mathbb{E}[D \mid X]$ 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 $\mathrm{Var}(D \mid X)$.

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*

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

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:

$$\mathbb{E}[VY] = \theta_0 \mathbb{E}[VD] + \mathbb{E}[V g_0(X)] + \mathbb{E}[V\varepsilon].$$

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

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

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

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

Remark 3.2 (Causal interpretation and orthogonality). 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 $\mathbb{E}[\varepsilon \mid D, X] = 0$ because $\mathbb{E}[V\varepsilon] = \mathbb{E}\{\mathbb{E}[(D - \mathbb{E}[D \mid X])\varepsilon \mid X]\} = \mathbb{E}\{\mathbb{E}[D\varepsilon \mid X] - \mathbb{E}[D \mid X]\mathbb{E}[\varepsilon \mid X]\} = 0$.

Causal interpretation. The estimand θ_0 is causal under standard potential-outcome conditions: consistency ($Y = Y(D)$), conditional exogeneity ($\mathbb{E}[\varepsilon \mid D, X] = 0$), and overlap ($\mathrm{Var}(D \mid X) > 0$). In the PLR framework, these conditions imply the orthogonal moment $\mathbb{E}[V \cdot (Y - g_0(X) - \theta_0(D - m_0(X)))] = 0$ (Chernozhukov et al., 2018). 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 $\mathrm{Var}(D \mid X = x) \geq \underline{\sigma}^2$ 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)**.** $\sigma^2_{D,n} \asymp 1$; i.e., $0 < c \leq \sigma^2_{D,n} \leq C < \infty$ for some constants c, C and all n.

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

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

$$\mathrm{Var}(D \mid X) = e(X)\{1 - e(X)\}, \qquad \sigma^2_V = \mathbb{E}[e(X)\{1 - e(X)\}].$$

Thus, the usual positivity condition $\epsilon \leq e(X) \leq 1 - \epsilon$ implies $\mathrm{Var}(D \mid X) \geq \epsilon(1 - \epsilon)$, 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 $\underline{\sigma}^2_n \downarrow 0$ such that for each n:

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

Define $\sigma^2_{V,n} := \mathbb{E}[\mathrm{Var}(D_n \mid X_n)]$ and $\kappa_n := \sigma^2_{D,n}/\sigma^2_{V,n}$.

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 $\underline{\sigma}^2_n \to 0$, we may have $\sigma^2_{V,n} \to 0$ and thus $\kappa_n \to \infty$. 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 $\ell_0(X) := \mathbb{E}[Y \mid X]$ 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 $\ell_0(X) =$

$\theta_0 m_0(X) + g_0(X)$:

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

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:

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

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

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

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

Proof. By the law of total variance:

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

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

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

Definition 3.3 (Condition Number and R^2 Representation)**.** The condition number is:

$$\kappa := \frac{\sigma_D^2}{\sigma_V^2} = \frac{1}{1 - R^2(D \mid X)}. \tag{7}$$

The empirical analogue uses cross-fitted residuals and defines the sample ratio:

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

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

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

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 $R^2(D \mid X) \to 1$, we have $\sigma_V^2 \to 0$ and $\kappa \to \infty$.

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 $\{I_1, \ldots, I_K\}$ of $\{1, \ldots, n\}$ 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 $n \to \infty$. This sample-splitting strategy originates in Schick (1986) 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:

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

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

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:

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

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)**.**

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

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)**.**

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

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)**.**

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

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:

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

Remark 3.12. 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 $\alpha_0 : \mathcal{W} \to \mathbb{R}$ 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,* $\alpha_0(W) = V/\sigma_V^2$ *(Ichimura and Newey, 2022). Also:*

$$\kappa = \sigma_D^2 \|\alpha_0\|_{L^2}^2, \qquad \|\alpha_0\|_{L^2}^2 = \sigma_V^{-2}. \tag{14}$$

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

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

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

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

Step 2 (Normalization):

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

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):

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

Remark 3.13 (Semiparametric Interpretation). Theorem 3.6 grounds our diagnostic in modern semiparametric theory (Newey, 1990; Bickel et al., 1993; 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. "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 $\mathbb{E}[\alpha(W) \mid X] = 0$, 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:

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

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

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

where:

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

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

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

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

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

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. □

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:*

$$\widehat{\theta} - \theta_0 = \widehat{\kappa}\,(S_n' + B_n'). \tag{22}$$

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 $\Psi_n(\widehat{\theta}, \widehat{\eta}) = 0$. Since the

score is affine in θ:

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

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

Step 2 (Express error): Subtracting θ_0:

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

Step 3 (Decompose score): By Lemma 3.7:

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

Step 4 (Standardize):

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

□

Remark 3.16 (Exactness of the decomposition). 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. Appendix A.3 records the corresponding generic identity for cross-fitted score estimators where the score need not be affine in θ; the same "Jacobian amplification" channel appears, though controlling remainders requires additional arguments.

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

3.10 Variance Inflation versus Bias Amplification

The exact decomposition shows that $\widehat{\kappa}$ 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 $\widehat{\kappa}\, S_n'$, 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 $\widehat{\kappa}\, B_n'$, 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 Stochastic-Order Implications of the Finite-Sample Identity

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) $\operatorname{ess\,sup}_x \mathbb{E}[V^2 \mid X = x] \le C$;

(ii) $\operatorname{ess\,sup}_{d,x} \mathbb{E}[\varepsilon^2 \mid D = d, X = x] \le C$.

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

Assumption 3.9 (Nuisance Rates)**.** $\|\widehat{m}^{(-k)} - m_0\|_{L^2} = O_P(r_n^m)$, $\|\widehat{\ell}^{(-k)} - \ell_0\|_{L^2} = O_P(r_n^\ell)$. Here r_n^m and r_n^ℓ are deterministic rate sequences (so Rem_n in Table 1 is deterministic).

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

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*

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

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, $|n^{-1}\sum_i V_i\Delta_i^m| \le \|V\|_n\|\Delta^m\|_n = O_P(\sigma_{V,n}) \cdot o_P(\sigma_{V,n}) = o_P(\sigma_{V,n}^2)$. 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. □ □

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* $\Delta^{(-k)} : \mathcal{X} \to \mathbb{R}$ *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*,*

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

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 $\|\Delta\|_n = O_P\Big(\max_{1\le k\le K}\|\Delta^{(-k)}\|_{L^2}\Big)$.

Theorem 3.11 (Stochastic-order bound)**.** *Under Assumptions 3.2, 3.7, 3.8, 3.9, 3.10, and 3.4:*

$$\widehat{\theta} - \theta_0 = O_P\left(\frac{\sqrt{\kappa_n}}{\sqrt{n}} + \kappa_n \cdot \mathrm{Rem}_n\right). \tag{23}$$

where the remainder term is:

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

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

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

By Assumption 3.8(ii) and iterated expectations:

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

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 $B_n^{(1)}$–$B_n^{(3)}$: 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:

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

Thus $\mathrm{Var}(B_n^{(1)}) = O_P((r_n^\ell)^2/n)$ 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:

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

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 Assumption 3.3 (strong overlap) and Assumption 3.4, a sufficient condition for valid $\sqrt{n}$-inference beyond oracle noise is* $\text{Rem}_n = o(n^{-1/2})$*, equivalently* $\kappa_n \cdot \text{Rem}_n = o(n^{-1/2})$.

3.12 Efficiency Bound Connection

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

Remark 3.22 (Why Condition on (D, X)). Assumption 3.11 (conditional homoskedasticity) is stronger than the second-moment condition $\mathbb{E}[\varepsilon^2 \mid X] < \infty$. It is imposed here only to obtain a clean link between conditioning and the efficiency bound, since it delivers the identity $\mathbb{E}[V^2\varepsilon^2] = \sigma_\varepsilon^2 \sigma_V^2$.

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

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

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:

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

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 stochastic-order bound amplifies estimation error. These regimes are labels, not hard cutoffs. Here, no formal thresholds are proposed.

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 $\mathrm{Var}(D_n \mid X_n = x) \geq \underline{\sigma}_n^2$ 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 $\widehat{\kappa}_{\text{oof}}$ 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 stochastic-order 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* $\sigma^2_{D,n} \asymp 1$*:*

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

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

(iii) Severely ill-conditioned ($\kappa_n \asymp \sqrt{n}$*):* $\widehat{\theta} - \theta_0 = O_P(n^{-1/4} + n^{1/2-\alpha})$*. If* $\alpha < 1/2$*, the remainder term dominates and may diverge.*

In a single sample, $\widehat{\kappa}$ summarizes sensitivity to nuisance bias. Reporting $\widehat{\kappa}$ communicates the amplification scale suggested by the bound. When $\widehat{\kappa}$ 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 stochastic-order bound (Theorem 3.11):

- **P1 Bias amplification:** For fixed nuisance-error magnitude δ, $|\text{Bias}(\widehat{\theta})|$ increases approximately proportionally with κ.
- **P2 Single-index collapse:** Under controlled perturbations, |Bias|/SE is approximately a function of the single index $\kappa \times \delta$.
- **P3 Coverage collapse mechanism:** Coverage fails when |Bias|/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 and the outcome metrics (bias, coverage, |Bias|/SE) correspond to the predictions P1–P3, which are direct implications of the exact decomposition and the stochastic-order bound.

5.1 Simulation Design

Data Generating Process

Following the simulation design common in the DML literature, we generate $n = 2{,}000$ observations from a Partially Linear Regression model:

$$Y_i = D_i\theta_0 + g_0(X_i) + \varepsilon_i, \tag{26}$$

$$D_i = m_0(X_i) + V_i, \tag{27}$$

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

The nuisance functions are:

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

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

To span conditioning regimes, we set $R^2(D \mid X) \in \{0.50, 0.75, 0.90, 0.95, 0.97, 0.99\}$, 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:

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

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.

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.

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.

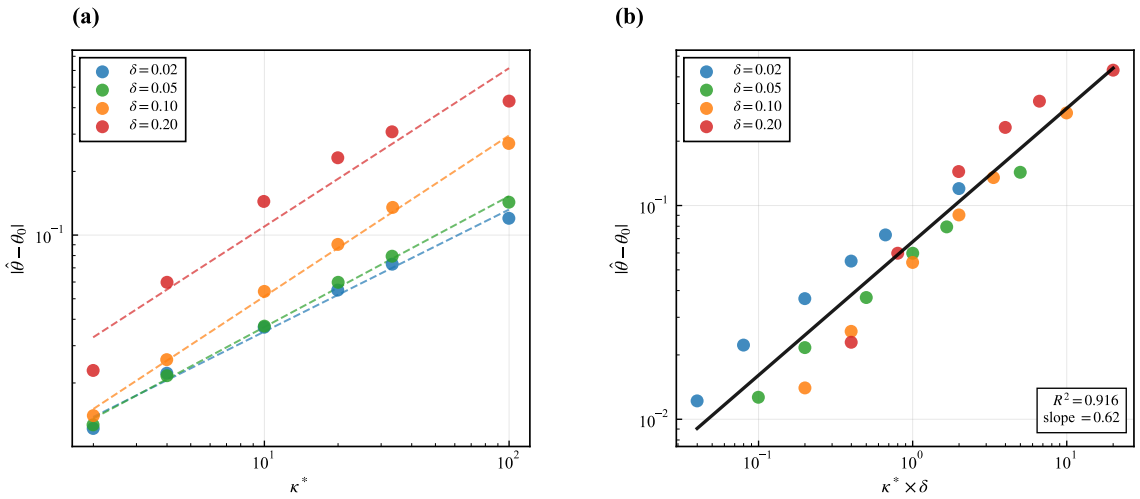

Figure 1: Bias amplification mechanism ($n = 2{,}000$, $B = 500$). (a) Mean absolute bias vs. structural κ for each $\delta \in \{0.02, 0.05, 0.10, 0.20\}$. Each line fixes δ; moving right increases κ, shifting bias proportionally, consistent with amplification (P1). (b) All points collapse onto a single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

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

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

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

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 close to nominal but not exactly 0.95 in finite samples; in our Monte Carlo grid it ranges from 0.90 to 0.96 across κ (Table 2). We treat the $\delta = 0$ row as the oracle baseline: departures from 0.95 there reflect finite-

sample and Monte Carlo variability, not amplification. Amplification is assessed by how coverage deteriorates as $\delta > 0$ interacts with increasing κ. As δ increases, coverage degrades monotonically in both dimensions. Coverage decreases primarily because |Bias|/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.[5]

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

	$R^2(D\|X)$ (structural κ)					
δ	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**

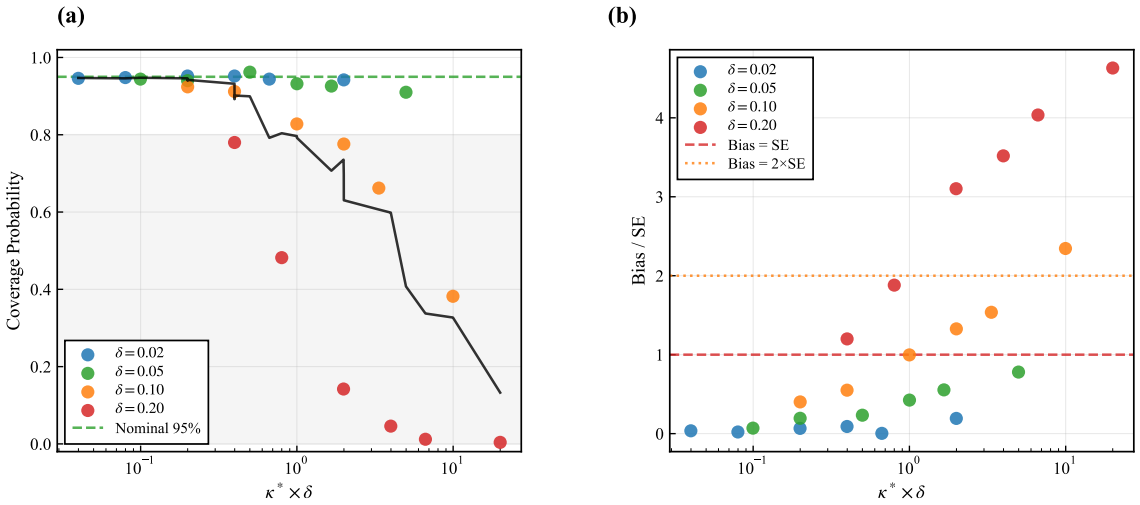

Figure 2: Coverage mechanism ($n = 2{,}000$, $B = 500$). (a) Coverage vs. $\kappa \times \delta$: confirms single-index organization. (b) Coverage vs. |Bias|/SE: the vertical dashed line at 1.0 marks the transition—normal-approximation CIs fail when bias dominates noise.

Panel (b) reveals why coverage fails: the ratio |Bias|/SE crosses 1.0 precisely when coverage drops below 80%. This confirms that undercoverage is due to bias amplification, not variance inflation. Sample-size sensitivity is summarized in Appendix Table 7.

[5]Under sequences with vanishing overlap, $\sqrt{n}$ approximations can fail and standard inference may become nonregular. This is emphasized in limited-overlap theory and high-dimensional overlap analyses, and recent work studies how much weak overlap doubly robust t-statistics can tolerate (Khan and Tamer, 2010; D'Amour et al., 2021; Dorn, 2025).

5.4 Sign-structure mechanism 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× 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 Implications

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]

6 Empirical Application: LaLonde Reanalysis

This section illustrates how the amplification factor $\widehat{\kappa}_{\text{oof}}$ 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 larger $\widehat{\kappa}_{\text{oof}}$ 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 $\widehat{\kappa}_{\text{oof}}$ 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 ($\widehat{\kappa}_{\text{oof}} \approx 1$), 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. For each learner, compute out-of-fold $\widehat{R}^2(D \mid X)$, yielding $\widehat{\kappa}_{\text{oof}} = 1/(1 - \widehat{R}^2_+)$. We run PLR-DML across the same learner set (OLS, Lasso, Ridge, RF, GBM, MLP) and summarize via forest plot.

6.1 Estimable Conditioning Summary

The empirical conditioning summary is:

$$\widehat{\kappa}_{b,\text{oof}} := \frac{1}{1 - \widehat{R}^2_{b,+}(D \mid X)}, \qquad \widehat{R}^2_{b,+}(D \mid X) := \max\{0, \widehat{R}^2_b(D \mid X)\}. \tag{29}$$

where $\widehat{R}^2_b(D \mid X)$ is the out-of-fold R^2 from the first-stage learner b. Out-of-fold $\widehat{R}^2$ can be slightly negative in finite samples; truncation at 0 enforces $\widehat{\kappa}_{b,\text{oof}} \geq 1$, matching the population interpretation.

We use $K = 5$ cross-fitting folds, RF hyperparameters tuned via 5-fold CV with 20 random samples. Out-of-fold $\widehat{R}^2$ is computed as $1 - \sum_i (D_i - \widehat{m}^{(-k)}(X_i))^2 / \sum_i (D_i - \bar{D})^2$. Across learners, mean out-of-fold $\widehat{R}^2(D|X)$ is 0.01 in the experimental sample and 0.25 in the obser-

[6] Applied simulation evidence often finds that causal-ML inference is sensitive to overlap and nuisance choices in finite samples; see practitioner-oriented finite-sample evaluation guidance and large-scale "revisited studies" evidence. (Naghi and Wirths, 2021; Baiardi and Naghi, 2024)

vational sample, yielding mean $\widehat{\kappa}_{\text{oof}} \approx 1$ vs. 1.92 (using the truncated $\widehat{R}^2_+ := \max\{0, \widehat{R}^2\}$). This confirms that conditioning (not just confounding) is systematically stronger in the observational sample, anchoring the instability to the amplification mechanism.

6.2 Results: Cross-learner dispersion and conditioning

This section illustrates the theory's empirical implication: $\widehat{\kappa}_{\text{oof}}$ 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 $\widehat{\kappa}_{\text{oof}}$) predicts cross-learner instability, consistent with the amplification mechanism. Figure 3 summarizes DML estimates across learners for both samples.

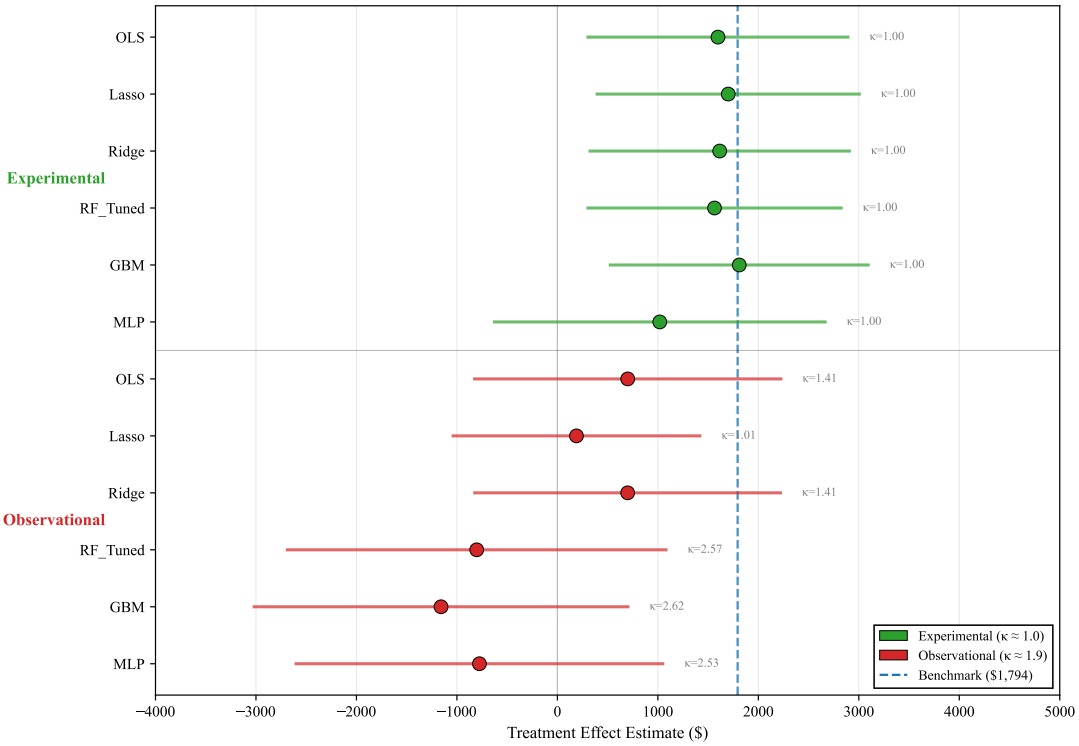

Figure 3: Forest plot of DML estimates by learner and sample. Experimental: (green, $\widehat{\kappa}_{\text{oof}} = 1$): tight clustering. Observational: (red, $\widehat{\kappa}_{\text{oof}} \in [1.0, 2.6]$): larger cross-learner dispersion—amplified sensitivity. Consistent with the theoretical fingerprint.

In the experimental sample, dispersion across learners is \$791 (range \$1,019–\$1,810) with $\widehat{\kappa}_{\text{oof}} = 1.00$ for all learners (after truncation). In the observational sample, dispersion is \$1,858 (range −\$1,158–\$700) with $\widehat{\kappa}_{\text{oof}} \in [1.01, 2.62]$. The observational sample exhibits

2.3× larger dispersion, consistent with the amplification mechanism. In the observational sample, flexible learners (RF, GBM, MLP) achieve higher $\widehat{R}^2(D|X)$, raising $\widehat{\kappa}_{\text{oof}}$ to 2.5–2.6. 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 $\widehat{\kappa}_{\text{oof}}$ is elevated, these differences manifest as large swings in $\hat{\theta}$. The pattern matches the Monte Carlo "fingerprint": stability under low $\widehat{\kappa}_{\text{oof}}$, fragility under high $\widehat{\kappa}_{\text{oof}}$.

When $\widehat{\kappa}_{\text{oof}}$ 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 $\widehat{\kappa}_{\text{oof}}$ as a warning, not a failure.

6.3 Interpretation: Conditioning is not identification

Observational instability can arise from both confounding and conditioning (Smith and Todd, 2005). $\widehat{\kappa}_{\text{oof}}$ quantifies conditioning of the orthogonalized score but does not test unconfoundedness or validate the design. What we show: when $\widehat{\kappa}_{\text{oof}}$ 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 $\widehat{\kappa}_{\text{oof}} = 1$.

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.

6.4 Reporting Implications

Appendix Table 8 reports the full numeric results. The estimated $\widehat{\kappa}_{\text{oof}}$ predicts when learner choice will matter: higher $\widehat{\kappa}_{\text{oof}}$ implies greater learner sensitivity. Stability in the experimental sample versus instability in the observational sample matches the amplification mechanism. In practice, report $\widehat{\kappa}_{\text{oof}}$ plus sensitivity across nuisance specifications, rather

than a single preferred estimate.[7]

In applications with elevated $\widehat{\kappa}_{\text{oof}}$, we recommend reporting: (i) $\widehat{\kappa}_{\text{oof}}$ 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 $\widehat{\kappa}_{\text{oof}}$, suspect model misspecification or violations of identifying assumptions.

7 Conclusion

This paper shows that, in DML, well-conditioning of the orthogonal score is not merely a numerical concern. It is a first-order amplification channel for finite-sample error when residual treatment variation is limited. In the partially linear regression (PLR) model, where the score is affine in θ, we establish an exact finite-sample identity, $\hat{\theta} - \theta_0 = \hat{\kappa}(\hat{S}_n + \hat{B}_n)$, which makes the multiplicative role of the (inverse-Jacobian) condition number explicit at any sample size. We then convert this identity into a stochastic-order bound, $\hat{\theta} - \theta_0 = O_P(\sqrt{\kappa_n/n} + \kappa_n \text{Rem}_n)$, yielding an operational sufficiency requirement for the usual $\sqrt{n}$ approximation beyond oracle noise: $\kappa_n \text{Rem}_n = o(n^{-1/2})$. Consequently, nuisance-rate conditions alone can be misleading in weak-overlap regimes, because κ_n may grow and elevate otherwise second-order remainder terms to first order.

For practice, we recommend reporting an estimable proxy for the amplification scale, $\hat{\kappa}_{\text{oof}} := 1/\{1 - \max(0, \hat{R}^2_{\text{oof}}(D \mid X))\}$, alongside the point estimate. A large $\hat{\kappa}_{\text{oof}}$ indicates a regime in which small, otherwise innocuous differences in nuisance learning choices can translate into large swings in $\hat{\theta}$. In such cases, narrow confidence intervals should be interpreted as potentially fragile. When multiple nuisance learners are plausible, we recommend reporting (i) $\hat{\kappa}_{\text{oof}}$ for each specification and (ii) a concise cross-learner sensitivity summary (e.g. a forest plot), rather than a single preferred estimate.

Our results are exact for PLR-DML, where the score is affine in θ. In more general orthogonal-score problems, the diagnostic framework extends by linearizing the score equation and tracking both (i) the score Jacobian (the stability object) and (ii) the associated linearization remainder. Appendix A.3 provides the generic orthogonal-score identity that

[7] When κ_{oof} is large, it is natural to complement reporting with design-stage overlap checks and remedies (e.g., trimming or overlap weights) rather than relying solely on a point estimate + robust s.e. (Crump et al., 2009; Li et al., 2018)

anchors this template. Developing weak-conditioning-robust inference procedures, and extending the analysis to heterogeneous treatment effects and other orthogonal-score settings, are promising directions.

Acknowledgments

We used Grammarly for language editing. All remaining errors are our own.

Replication and reproducibility.

All tables and figures are generated from raw inputs using a single master script. The replication package includes: (i) code to reproduce every figure/table in the paper, (ii) fixed random seeds, (iii) a manifest listing software versions and computational environment, and (iv) instructions to run each experiment end-to-end. Replication code: https://github.com/gsaco/dml-diagnostic

References

Baiardi, A., & Naghi, A. (2024). The value added of machine learning to causal inference: Evidence from revisited studies. *The Econometrics Journal*, 27(2), 213–234. https://doi.org/10.1093/ectj/utae004

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608–650. https://doi.org/10.1093/restud/rdt044

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press. https://doi.org/10.2307/2533465

Breunig, C., E. Mammen, and A. Simoni (2020). Ill-posed estimation in high-dimensional models with instrumental variables. *Journal of Econometrics* 219(1), 171–200. https://doi.org/10.1016/j.jeconom.2020.04.043

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chernozhukov, V., Newey, W. K., Quintas-Martinez, V., & Syrgkanis, V. (2021). Automatic debiased machine learning via Riesz regression. *arXiv preprint* arXiv:2104.14737.

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3), 967–1027. https://doi.org/10.3982/ECTA18515

Chernozhukov, V., Newey, W. K., & Singh, R. (2023). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika*, 110(1), 257–264. https://doi.org/10.1093/biomet/asac033

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), 187–199. https://doi.org/10.1093/biomet/asn055

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2), 644–654. https://doi.org/10.1016/j.jeconom.2019.10.014

Dorn, J. (2025). How much weak overlap can doubly robust t-statistics tolerate? *arXiv preprint* arXiv:2504.13273.

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181–213. https://doi.org/10.3982/ECTA16901

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1(4), 387–401. https://doi.org/10.2307/1907330

Golub, G. H., & Van Loan, C. F. (2013). *Matrix Computations* (4th ed.). Johns Hopkins University Press.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2), 315–331. https://doi.org/10.2307/2998560

Han, S. and A. McCloskey (2019). Estimation and inference with a (nearly) singular Jacobian. *Quantitative Economics* 10(3), 1019–1068. https://doi.org/10.3982/QE989

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Ichimura, H., & Newey, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics*, 13(1), 29–61. https://doi.org/10.3982/QE826

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Kaji, T. (2021). Theory of weak identification in semiparametric models. *Econometrica* 89(2), 733–763. https://doi.org/10.3982/ECTA16413.

Kang, J. D. Y., & 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. https://doi.org/10.1214/07-STS227

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In H. He, P. Wu, & D.-G. Chen (Eds.), *Statistical Causal Inferences and Their Applications in Public Health Research* (pp. 141–167). Springer. https://doi.org/10.1007/978-3-319-41259-7_8

Kennedy, E. H. (2024). Semiparametric doubly robust targeted double machine learning: A review. In E. B. Laber, M. J. Meyer, B. J. Reich, & R. Wang (Eds.), *Handbook of Statistical Methods for Precision Medicine* (Chap. 10, pp. 207–236). Chapman & Hall/CRC. https://doi.org/10.1201/9781003216223.

Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6), 2021–2042. https://doi.org/10.3982/ECTA7372

Kvalseth, T. O. (1985). Cautionary note about R^2. *The American Statistician*, 39(4), 279–285. https://doi.org/10.1080/00031305.1985.10479448

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4), 604–620.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4), 1027–1048. https://doi.org/10.1111/1468-0262.00438

Naghi, A. A., & Wirths, C. P. (2021). Finite sample evaluation of causal machine learning methods: Guidelines for the applied researcher. Tinbergen Institute Discussion Paper TI 2021-090/III.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2), 99–135. https://doi.org/10.1002/jae.3950050202

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1), 31–54. https://doi.org/10.1177/0962280210386207

Quintas-Martinez, V. M. (2022). Finite-sample guarantees for high-dimensional DML. *arXiv preprint* arXiv:2206.07386.

Riesz, F. (1907). Sur une espèce de géométrie analytique des systèmes de fonctions sommables. *Comptes Rendus de l'Académie des Sciences, Paris*, 144, 1409–1411.

Riesz, F. (1909). Sur les opérations fonctionnelles linéaires. *Comptes Rendus de l'Académie des Sciences, Paris*, 149, 974–977.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56(4), 931–954. https://doi.org/10.2307/1912705

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322–331. https://doi.org/10.1198/016214504000001880

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3), 1139–1151.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1–2), 305–353. https://doi.org/10.1016/j.jeconom.2004.04.011

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press. https://doi.org/10.1017/CBO9780511614491.006

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, $\mathbb{E}[\varepsilon \mid X] = \mathbb{E}[\mathbb{E}[\varepsilon \mid D, X] \mid X] = 0$, hence $\mathbb{E}[(m - m_0)(X) \cdot \varepsilon] = \mathbb{E}[(m - m_0)(X) \cdot \mathbb{E}[\varepsilon \mid X]] = 0$.
- $\mathbb{E}[V \cdot (\ell - \ell_0)(X)] = \mathbb{E}[\mathbb{E}[V \mid X] \cdot (\ell - \ell_0)(X)] = 0$.
- $\mathbb{E}[V \cdot (m - m_0)(X)] = \mathbb{E}[\mathbb{E}[V \mid X] \cdot (m - m_0)(X)] = 0$, since $\mathbb{E}[V \mid X] = 0$. □

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

Proof. Substitute regime-specific κ_n into Theorem 3.11. Assume $\sigma^2_{D,n} \asymp 1$ 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. □

A.3 A Generic Orthogonal-Score Identity

This paper exploits a special feature of PLR: the orthogonal score is affine in θ, so the DML estimating equation can be solved exactly and yields an exact decomposition without any Taylor/von Mises remainder (Theorem 3.8 and Remark 3.16). For completeness, we record the corresponding generic identity for cross-fitted score estimators, which highlights the

same "Jacobian amplification" channel in models where the score need not be affine in θ.

Throughout, probabilities and expectations are under the triangular-array law P_n as in Assumption 3.1 and Definition 3.1.

Setup. Let $\psi(W;\theta,\eta) \in \mathbb{R}^{d_\theta}$ be a (possibly nonlinear) score, where $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and η is a nuisance element. Using the same K-fold cross-fitting structure as in Definition 3.4, let $\hat{\eta}^{(-k)}$ denote the nuisance estimator trained on $\{W_j : j \notin I_k\}$. Define the cross-fitted empirical moment map

$$\Psi_n(\theta,\hat{\eta}) := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\psi\Big(W_i;\theta,\hat{\eta}^{(-k)}\Big). \tag{30}$$

Let $\hat{\theta}$ be any solution to $\Psi_n(\hat{\theta},\hat{\eta}) = 0$.

Mean-value (Jacobian) identity. Assume that $\theta \mapsto \Psi_n(\theta,\hat{\eta})$ is continuously differentiable on the line segment $\{\theta_0 + t(\hat{\theta}-\theta_0) : t \in [0,1]\}$ and define the intermediate Jacobian

$$\widehat{J}_n^e := \int_0^1 \partial_\theta \Psi_n\Big(\theta_0 + t(\hat{\theta}-\theta_0),\hat{\eta}\Big)\, dt. \tag{31}$$

Then the vector mean-value theorem implies

$$0 = \Psi_n(\hat{\theta},\hat{\eta}) = \Psi_n(\theta_0,\hat{\eta}) + \widehat{J}_n^e(\hat{\theta}-\theta_0). \tag{32}$$

If $\widehat{J}_n^e$ is nonsingular, rearranging yields the identity

$$\hat{\theta}-\theta_0 = -(\widehat{J}_n^e)^{-1}\,\Psi_n(\theta_0,\hat{\eta}). \tag{33}$$

Oracle–bias decomposition. Add and subtract the oracle score:

$$\Psi_n(\theta_0,\hat{\eta}) = \underbrace{\Psi_n(\theta_0,\eta_0)}_{=:S_n^{\mathrm{gen}}} + \underbrace{\big\{\Psi_n(\theta_0,\hat{\eta}) - \Psi_n(\theta_0,\eta_0)\big\}}_{=:B_n^{\mathrm{gen}}}. \tag{34}$$

Combining (33) and (34) gives the decomposition

$$\hat{\theta}-\theta_0 = -(\widehat{J}_n^e)^{-1}\Big(S_n^{\mathrm{gen}} + B_n^{\mathrm{gen}}\Big). \tag{35}$$

Connection to PLR. In PLR, the orthogonal score (10) is affine in θ, so $\partial_\theta \Psi_n(\theta, \hat{\eta})$ does not depend on θ, and the intermediate Jacobian $\widehat{J}_n^e$ in (31) collapses to the empirical Jacobian used in the main text. Consequently (35) reduces to the exact identity in Theorem 3.8, with no linearization remainder (Remark 3.16).

For nonlinear scores, (35) remains an exact algebraic identity conditional on differentiability (it is simply the mean-value theorem applied to the sample estimating equation), but controlling B_n^{gen} and replacing $(\widehat{J}_n^e)^{-1}$ by a convenient plug-in typically requires additional smoothness and orthogonality arguments, and yields explicit remainder terms.

B Simulation Appendix

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

B.1 Log–log exponent diagnostic for bias scaling

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

$$\log |\text{Bias}| = a + \alpha \log \kappa + \beta \log \delta + \text{error}. \tag{36}$$

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	

Theorem 3.11 implies that the bias component is of order κRem_n, where Rem_n collects products and squares of nuisance errors. We therefore report $(\hat{\alpha}, \hat{\beta})$ as a descriptive summary of how Monte Carlo bias scales with (κ, δ) over the design grid, not as a test of a sharp theoretical log–log slope in δ.

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	|Bias|	Coverage	|Bias|/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 |Bias|/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 ($R^2 = 0.90$, $\delta = 0.1$, $B = 500$). 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	$\widehat{\kappa}_{\text{oof}}$	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 $\widehat{\kappa}_{\text{oof}} = 1$; (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 $\widehat{\kappa}_{\text{oof}} \in [2.5, 2.6]$.