Finite-Sample Fragility in Double Machine Learning: Bias Amplification and Diagnostics

Gabriel Saco*

December 2025

Abstract

Double Machine Learning (DML) delivers root-n inference under the assumption that nuisance estimation error is asymptotically negligible ($o(n^{-1/2})$). We show that this assumption depends critically on the conditioning of the orthogonal score. Using an exact finite-sample decomposition, we demonstrate that weak overlap—characterised by a high condition number—multiplies residual nuisance bias. In ill-conditioned regimes, flexible learners such as Random Forests exhibit severe undercoverage not because of variance inflation, but because of *bias amplification*. We propose the standardised diagnostic $\kappa^* = 1/(1 - R^2(D \mid X))$ to accompany all DML estimates, analogous to the first-stage F-statistic in instrumental variables regression. Monte Carlo experiments and re-analysis of LaLonde (1986) confirm that κ^* reliably distinguishes robust inference from fragile estimates.

Keywords: Double Machine Learning, Bias Amplification, Weak Identification, Finite-Sample Inference, Condition Number, Variance Inflation Factor.
JEL Codes: C14, C21, C52.

*Universidad del Pacífico. Email: gsacoalvarado@gmail.com. The Python package `dml_diagnostic` and full replication code are available at https://github.com/gsaco/dml-diagnostic.

1 Introduction

Double Machine Learning (DML; Chernozhukov et al., 2018) represents one of the most significant methodological advances in causal inference of the past decade. The framework promises to separate the complexity of nuisance function estimation from the inferential problem of quantifying uncertainty about causal parameters. By combining Neyman-orthogonal scores—which are locally insensitive to first-order perturbations in nuisance estimates—with cross-fitting, DML delivers asymptotically valid inference even when nuisance functions are estimated with flexible machine learning methods such as Random Forests, neural networks, or gradient-boosted trees. The methodology has been adopted across economics, epidemiology, political science, and the technology industry, with researchers and practitioners routinely reporting point estimates and confidence intervals based on DML. The promise is seductive: let machine learning handle the confounders, and enjoy root-n inference on the causal effect.

This paper demonstrates that the promise has an important caveat. The asymptotic theory of DML rests on a condition that is rarely verified in practice: that nuisance estimation error is "sufficiently negligible"—specifically, $o_P(n^{-1/2})$—in a precise sense. This condition is asymptotic in nature; it says nothing about what happens in finite samples. We show that in finite samples, there exists a structural amplifier that can magnify residual nuisance error into a first-order bias, destroying the coverage guarantees that practitioners take for granted. This amplifier is the *condition number* of the orthogonal score.

The failure mode we identify is particularly insidious because it is *silent*. Unlike variance inflation—where confidence intervals widen to honestly reflect increased uncertainty—bias amplification produces narrow confidence intervals that systematically miss the true parameter. The estimated standard errors, which capture only sampling variance, fail to account for the systematic shift caused by amplified nuisance bias. A practitioner who checks coverage probabilities in well-specified Monte Carlo simulations will find 95% coverage; the same practitioner who applies the method to data with weak overlap may obtain 40% coverage without any warning from the statistical machinery.

The phenomenon is easy to state: when treatment is highly predictable from covariates, there is little residual variation in treatment to identify the causal effect. Identification becomes fragile. What is less obvious—and what this paper demonstrates—is the precise mechanism through which this fragility manifests. The condition number acts as a *lever*, multiplying any residual bias from nuisance estimation into the parameter estimate. If the condition number is 100 and the nuisance learner introduces a regularization bias of 0.01 (in standardized units), the resulting estimation error is 1.0—potentially orders of magnitude larger than the standard error.

The analogy to weak instruments in instrumental variables (IV) regression is both instructive and historically significant. Before the influential work of Staiger and Stock (1997) and the diagnostic framework of Stock and Yogo (2005), applied researchers routinely reported two-stage least squares (2SLS) estimates without diagnosing first-stage strength. The instruments were "valid" in the sense that exclusion restrictions were satisfied; the econometric theory was "correct" in the sense that asymptotic normality held under regularity conditions. Yet in finite samples with weak first-stage relationships, 2SLS estimates could be severely biased, and Wald confidence intervals could have arbitrarily poor coverage.

The introduction of the first-stage F-statistic as a standard diagnostic transformed IV practice. A simple rule of thumb—$F > 10$ for reliable inference—gave practitioners actionable guidance. The F-statistic summarized the strength of identification in a single number, and its routine reporting made the fragility

of weak-instrument designs immediately visible. A generation of applied researchers learned to view IV estimates with $F < 10$ with deep skepticism.

We argue that DML is currently in its "pre-Staiger-Stock" moment with respect to overlap. Researchers report DML estimates without any standardized diagnostic of the score's conditioning. Propensity score plots and overlap summaries diagnose nuisance estimation but not the sensitivity of the final estimator. There is no scalar summary that quantifies, in a single number, how much the estimating equation amplifies residual errors. This paper fills that gap.

The central conceptual contribution of this paper is the distinction between *variance inflation* and *bias amplification*. This distinction is fundamental to understanding why DML can fail in ways that classical regression cannot.

Consider classical OLS regression with collinear regressors. The Variance Inflation Factor (VIF) measures how much collinearity inflates the variance of coefficient estimates. When the VIF is large, confidence intervals widen. Crucially, they widen *honestly*: the standard errors correctly reflect the increased uncertainty, and coverage probabilities remain near nominal levels. The estimator is unbiased; only its precision suffers.

DML with regularized learners operates in a fundamentally different regime. Machine learning methods achieve their predictive power by trading variance for bias—this is the bias-variance tradeoff at the heart of statistical learning theory. Regularization (whether through explicit penalties, early stopping, tree depth limits, or ensemble averaging) shrinks predictions toward smooth or simple functions, introducing systematic bias. This bias is small when the regularization is well-calibrated and the sample size is large. But it is generically *non-zero* in finite samples.

The condition number multiplies this non-zero bias. When the condition number is small, the product $\kappa \cdot B_n$ is negligible, and DML behaves as advertised. When the condition number is large, the product dominates, and coverage collapses. The confidence intervals remain narrow because the standard errors—which measure sampling variance—are unaffected by the bias. The practitioner sees precise estimates that are systematically wrong.

This is the "silent failure" of DML: unlike variance inflation, which announces itself through wide confidence intervals, bias amplification produces confident and incorrect inference.

To make this precise, we derive an exact finite-sample decomposition of the DML estimator error. Working within the canonical Partially Linear Regression model of Chernozhukov et al. (2018), we show that the estimation error admits the representation

$$\widehat{\theta} - \theta_0 = \widehat{\kappa} \cdot S_n' + \widehat{\kappa} \cdot B_n', \tag{1}$$

where $\widehat{\kappa}$ is the empirical condition number, S_n' is the standardized oracle sampling term, and B_n' is the standardized nuisance bias term. This decomposition is an *algebraic identity*, not an approximation—there is no Taylor remainder, no higher-order terms. The exactness follows from the affine structure of the Partially Linear Regression score and implies that the bias amplification mechanism operates at *all* sample sizes, not merely asymptotically.

The decomposition reveals both failure channels. The first term, $\widehat{\kappa} \cdot S_n'$, captures *variance inflation*: sampling uncertainty is scaled by the condition number, widening confidence intervals just as the classical VIF widens intervals for collinear regressors in OLS. The standard error estimator correctly accounts for

this inflation, so coverage probabilities remain near nominal levels. The second term, $\widehat{\kappa} \cdot B'_n$, is the *bias amplification* term. When nuisance learners introduce regularization bias—as virtually all flexible machine learning methods do in finite samples—this bias is multiplied by κ. Unlike variance inflation, bias amplification is invisible to the standard error estimator. The confidence intervals remain narrow while the point estimate is systematically displaced.

The condition number κ arises naturally as the inverse of the Jacobian magnitude in the score equation; equivalently, it equals the ratio $\text{Var}(D)/\mathbb{E}[\text{Var}(D \mid X)]$. For interpretability and cross-study comparability, we work with the **standardized condition number**

$$\kappa^* := \frac{1}{1 - R^2(D \mid X)}, \tag{2}$$

which is scale-invariant and coincides precisely with the classical Variance Inflation Factor (VIF) from regression diagnostics (Belsley et al., 1980). We claim no novelty for this statistic; the VIF has been routinely computed for decades. What is new is its role in governing a failure mode specific to DML: the amplification of nuisance bias into a first-order distortion. We establish a formal connection to the Riesz representer norm from semiparametric efficiency theory (Chernozhukov et al., 2022), showing that κ^* quantifies the "difficulty" of the debiasing problem in a mathematically rigorous sense.

Building on this theoretical foundation, we propose a classification of conditioning regimes that parallels the Stock–Yogo framework for weak instruments. When $\kappa^* < 5$—covariates explain less than 80% of treatment variation—standard DML asymptotics are reliable, and coverage should be near nominal for appropriately specified learners. When $5 \leq \kappa^* \leq 20$, bias amplification becomes a concern; sensitivity analysis across learners with different bias-variance tradeoffs is strongly warranted, and the "effective sample size" n/κ^* may be substantially smaller than the nominal sample size. When $\kappa^* > 20$—covariates explain more than 95% of treatment variation—bias is likely to dominate for any regularized learner, and DML confidence intervals should not be trusted without additional robustness checks. These thresholds, like the $F > 10$ rule for instruments, are guidelines rather than bright lines; they translate abstract asymptotic conditions into actionable diagnostic practice.

We validate the theoretical predictions through Monte Carlo simulation and an empirical application. In simulations, we construct data-generating processes with varying $R^2(D \mid X)$ values and compare the behavior of unbiased learners (OLS) with biased learners (Random Forests). As κ^* increases:

- Unbiased learners exhibit variance inflation: coverage remains near 95%, but confidence intervals widen proportionally to $\sqrt{\kappa}$.
- Biased learners exhibit bias amplification: coverage collapses (to below 40% at $\kappa^* \approx 33$), while confidence intervals remain narrow.

The "smoking gun" is the ratio of bias to standard error: for Random Forests at high κ^*, this ratio exceeds 2, explaining the catastrophic undercoverage.

In the empirical application, we revisit the canonical job-training study of LaLonde (1986). The experimental sample yields $\kappa^* \approx 4$ across learners, and estimates are stable (\$1,455–\$1,793) and consistent with the experimental benchmark. The observational sample yields $\kappa^* > 20$, and estimates range from −\$642 to +\$621—spanning zero and reversing sign across learners. The κ^* diagnostic immediately distinguishes

reliable from fragile inference.

The remainder of the paper proceeds as follows. Section 2 situates our contribution within the literatures on semiparametric efficiency, weak identification, and machine learning for causal inference. Section 3 develops the theoretical framework, deriving the exact decomposition, characterizing the bias amplification mechanism, and connecting the condition number to the Riesz representer. Section 4 defines the conditioning regimes and provides practical guidance for interpreting κ^*. Section 5 presents Monte Carlo evidence. Section 6 applies the diagnostic to the LaLonde data. Section 7 offers practical recommendations. Mathematical proofs are collected in the Appendix.

2 Related Literature

This paper synthesizes four streams of research: semiparametric efficiency theory for treatment effects, finite-sample problems in high-dimensional inference, weak identification diagnostics in instrumental variables and GMM, and machine learning methods for causal inference. We draw on each to illuminate the bias amplification mechanism and its implications for applied practice.

2.1 Semiparametric Efficiency and the Role of Overlap

The semiparametric efficiency bound for treatment effect estimation has been a central object of study since the foundational work of Hahn (1998), who derived the efficiency bound for average treatment effects under unconfoundedness and demonstrated that its magnitude depends critically on the propensity score. For binary treatments, the bound involves terms of the form $1/e(x)$ and $1/(1-e(x))$, where $e(x) = \mathbb{P}(D = 1 \mid X = x)$ is the propensity score. As the propensity score approaches zero or one—that is, as overlap weakens—the efficiency bound diverges, signaling that even optimal estimators must have large variance.

Hirano et al. (2003) extended this analysis by establishing that efficient estimators can achieve the semiparametric bound when propensity scores are estimated nonparametrically. Their efficiency result, however, is predicated on the propensity scores being bounded away from zero and one. Rosenbaum and Rubin (1983) laid foundational work on propensity score methodology, while subsequent work by Imbens (2004) provided a comprehensive framework for nonparametric estimation under unconfoundedness. The efficiency bound provides no operational guidance when overlap conditions fail; it merely indicates that variance must be large.

In the Partially Linear Regression model that we study, the efficiency bound takes the form $V_{\text{eff}} = \mathbb{E}[V^2\varepsilon^2]/(\mathbb{E}[V^2])^2$, where $V = D - m_0(X)$ is the treatment residual and ε is the structural error. As $\mathbb{E}[V^2] \to 0$, the bound diverges. Our condition number $\kappa = 1/\mathbb{E}[V^2]$ (in appropriate units) is thus a finite-sample analogue of the factor governing the efficiency bound. What the semiparametric literature establishes is that *asymptotically*, estimators must have larger variance as overlap weakens. What it does not address—and what we contribute—is the finite-sample mechanism through which weak overlap interacts with nuisance estimation bias.

The comprehensive treatment of Kennedy (2023) provides a unified framework for semiparametric inference, emphasizing influence functions, doubly robust estimators, and the conditions under which root-n inference is achievable. Our analysis complements this literature by providing a computable diagnostic that signals

proximity to the boundary where standard conditions fail.

2.2 Finite-Sample Inference and Post-Selection Bias

A recurring theme in modern econometrics is that asymptotic approximations can fail badly in finite samples, particularly when model selection or regularization is involved. Belloni et al. (2014) developed the post-double-selection estimator for high-dimensional regression, addressing the challenge that LASSO-based variable selection can induce bias in subsequent inference. Their key insight—that inference on a target coefficient requires "protecting" the selection step by also selecting controls for the outcome equation—presages the double-robustness structure of DML.

The finite-sample theory of debiased machine learning has received increasing attention. Chernozhukov et al. (2023) establish Berry-Esseen bounds for the DML t-statistic, providing nonasymptotic guarantees for coverage accuracy. Their results show that the coverage error of DML confidence intervals is of order $n^{-1/2} + \sqrt{n} \cdot r_n$, where r_n is the product of nuisance estimation rates. Importantly, these bounds assume regularity conditions that implicitly bound the Jacobian away from zero—that is, they assume the condition number κ is $O(1)$. Quintas-Martínez (2022) and Jung (2023) provide complementary finite-sample guarantees under similar maintained assumptions.

When the $\kappa = O(1)$ assumption fails, the finite-sample picture changes dramatically. We show that the relevant error bound becomes $O_P(\kappa_n/\sqrt{n} + \kappa_n \cdot r_n)$, with both terms scaled by the condition number. This implies that even small regularization biases, multiplied by a large κ, can produce first-order distortions invisible to standard inference. The finite-sample literature has largely focused on establishing that DML "works" under regularity; our contribution is to characterize precisely when and how it fails.

2.3 Weak Identification in Instrumental Variables and GMM

The weak instruments literature provides our conceptual template. Before the work of Staiger and Stock (1997), the standard treatment of instrumental variables regression focused on consistency and asymptotic normality under the assumption that instruments are "relevant"—correlated with the endogenous regressor. The relevance assumption was treated as binary: either instruments were relevant (and 2SLS was valid) or they were not (and identification failed entirely). What Staiger and Stock demonstrated was that this binary view obscured a *continuous fragility*. When instruments are "weak"—correlated with the endogenous regressor, but only weakly so—2SLS exhibits pathological finite-sample behavior: the estimator is biased toward the OLS coefficient, and confidence intervals can have arbitrarily poor coverage.

Stock and Yogo (2005) translated this theoretical insight into actionable practice. They provided critical values for the first-stage F-statistic, showing that $F > 10$ was a reasonable threshold for inference that is not too severely distorted. This simple rule of thumb transformed empirical practice in applied economics; a generation of researchers learned to diagnose instrument strength before trusting IV estimates. Subsequent work, including the comprehensive survey by Andrews et al. (2019), refined the diagnostics and developed alternative testing procedures.

Stock and Wright (2000) extended the weak identification framework to GMM, classifying designs by identification strength into "strong," "semi-strong," and "weak" regimes, each with distinct asymptotic behavior. Their framework maps directly onto ours: we classify DML designs as well-conditioned ($\kappa^* < 5$,

analogous to strong identification), moderately ill-conditioned ($5 \leq \kappa^* \leq 20$, analogous to semi-strong), or severely ill-conditioned ($\kappa^* > 20$, analogous to weak). The structural parallel extends further: in IV, the Jacobian of the moment condition (the first-stage coefficient π) determines identification strength; in DML, the Jacobian of the orthogonal score ($J_\theta = -\mathbb{E}[V^2]$) plays the same role, and our condition number is its inverse.

Bound et al. (1995) documented the empirical prevalence of weak instruments and their consequences for applied research, while Moreira (2003) developed the conditional likelihood ratio test that provides valid inference regardless of instrument strength. We do not develop conditioning-robust tests for DML—that is an important direction for future work—but establish that κ^* reliably indicates when standard DML confidence intervals become fragile.

2.4 Machine Learning for Causal Inference

The Double Machine Learning framework of Chernozhukov et al. (2018) represents a synthesis of two decades of work on semiparametric inference and modern machine learning. The framework addresses a fundamental challenge: how can we use flexible, potentially slowly-converging machine learning methods to estimate nuisance functions while still obtaining root-n inference for causal parameters? The answer involves two key insights. First, Neyman-orthogonal scores are locally insensitive to first-order perturbations in nuisance estimates—the pathwise derivative with respect to the nuisance parameter vanishes at the truth, so nuisance estimation error enters only through second-order terms. Second, cross-fitting eliminates the "own-observation" bias that would otherwise arise from using the same data for both nuisance estimation and parameter estimation. Together, these insights yield the DML "product rate" condition: if the product of nuisance estimation rates satisfies $r_n = o_P(n^{-1/2})$, standard inference applies.

The product rate condition is stated in terms of nuisance estimation rates alone; it does not mention the condition number. Our exact decomposition shows that the relevant condition is not just $r_n = o_P(n^{-1/2})$ but $\kappa \cdot r_n = o_P(n^{-1/2})$. When κ is large, nuisance estimators must converge *faster* to maintain valid inference.

The Riesz representer approach of Chernozhukov et al. (2022) provides a unifying framework for debiased machine learning that directly connects to our analysis. The Riesz representer $\alpha_0(W) = V/\mathbb{E}[V^2]$ is the function that "reweights" the score to achieve double robustness, and its squared L^2 norm is $\|\alpha_0\|_{L^2}^2 = 1/\mathbb{E}[V^2]$. As we show in Section 3, this norm is precisely κ/σ_D^2—the condition number up to a normalizing constant. The Riesz representer literature establishes that $\|\alpha_0\|_{L^2}$ governs the variance of debiased estimators; we extend this to show that when nuisance estimation introduces bias, the same quantity governs bias amplification.

Recent work addresses weak overlap directly. Crump et al. (2009) propose trimming observations with extreme propensity scores. Ma et al. (2023) develop doubly robust estimators tailored to weak overlap settings. Li et al. (2018) introduce overlap weights to improve balance. These methods complement our diagnostic: κ^* indicates when corrections may be needed and confirms whether conditioning has improved.

2.5 The Gap: No Diagnostic for DML Conditioning

Despite the progress surveyed above, there remains no standardized, interpretable diagnostic that signals when DML inference is at risk of failure due to poor conditioning—analogous to the F-statistic for weak instruments. The semiparametric efficiency literature establishes that variance must increase as overlap

weakens, but provides no operational threshold. The finite-sample DML literature derives bounds under the implicit assumption that $\kappa = O(1)$, without characterizing what happens when this assumption fails. The weak identification literature provides conceptual templates but has not been adapted to the DML setting. And while practitioners routinely examine propensity score plots and balance diagnostics, these tools diagnose nuisance estimation quality rather than the sensitivity of the final estimator.

This paper fills the gap by proposing the standardized condition number κ^* as a routine diagnostic, establishing its connection to both classical VIF and modern Riesz representer theory, and providing threshold-based guidance for interpreting the reliability of DML inference.

3 Theoretical Framework

This section develops the theoretical foundation for the bias amplification mechanism. We work within the Partially Linear Regression (PLR) model, the canonical example in Chernozhukov et al. (2018). We first establish notation and the DML estimator, then define the condition number and derive its properties, and finally present the exact finite-sample decomposition that reveals the bias amplification channel. Complete proofs are provided in Appendix A.

3.1 Model, Notation, and the DML Estimator

We observe independent and identically distributed data $\{W_i\}_{i=1}^n$ drawn from a probability measure P, where each observation $W_i = (Y_i, D_i, X_i)$ consists of a scalar outcome $Y \in \mathbb{R}$, a scalar treatment $D \in \mathbb{R}$, and a p-dimensional covariate vector $X \in \mathbb{R}^p$. The PLR model specifies the data-generating process through two structural equations:

$$Y = \theta_0 D + g_0(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid D, X] = 0, \tag{3}$$

$$D = m_0(X) + V, \qquad \mathbb{E}[V \mid X] = 0. \tag{4}$$

The parameter $\theta_0 \in \mathbb{R}$ is the causal effect of interest: the expected change in outcome per unit change in treatment, holding covariates fixed. The function $g_0 : \mathbb{R}^p \to \mathbb{R}$ is an unknown nuisance function capturing the direct effect of covariates on outcomes. The propensity function $m_0(X) = \mathbb{E}[D \mid X]$ is the conditional expectation of treatment given covariates. The treatment residual $V = D - m_0(X)$ represents the part of treatment variation not explained by covariates; by construction, $\mathbb{E}[V \mid X] = 0$. The structural error ε satisfies the exogeneity condition $\mathbb{E}[\varepsilon \mid D, X] = 0$, which combined with standard unconfoundedness assumptions justifies interpreting θ_0 as a causal effect.

The treatment residual V is central to identification. Under the PLR model, the only variation in D that identifies θ_0 is the residual variation V—the part not explained by covariates. When $\mathrm{Var}(V)$ is small relative to $\mathrm{Var}(D)$, covariates explain most of the treatment variation, and little identifying information remains.

Outcome residual and its decomposition. Define the outcome nuisance function $\ell_0(X) := \mathbb{E}[Y \mid X]$. By taking conditional expectations in (3), we obtain $\ell_0(X) = \theta_0 m_0(X) + g_0(X)$. The outcome residual is

$U := Y - \ell_0(X)$. A key algebraic identity (proven in the Appendix) is:

$$U = \theta_0 V + \varepsilon. \tag{5}$$

This decomposition shows that the outcome residual equals the causal effect times the treatment residual, plus structural noise. The treatment residual V is the identifying variation.

Cross-fitting and nuisance estimation. The DML procedure uses K-fold cross-fitting to estimate nuisance functions. Let $\{I_1, \ldots, I_K\}$ be a random partition of the index set $\{1, \ldots, n\}$ into K approximately equal-sized folds. For each fold k, we train nuisance estimators $\widehat{m}^{(-k)}$ and $\widehat{\ell}^{(-k)}$ on the out-of-fold sample $\{W_i : i \notin I_k\}$. For observations $i \in I_k$, the cross-fitted residuals are:

$$\widehat{V}_i := D_i - \widehat{m}^{(-k)}(X_i), \qquad \widehat{U}_i := Y_i - \widehat{\ell}^{(-k)}(X_i). \tag{6}$$

The key property of cross-fitting is that for $i \in I_k$, the nuisance estimates $\widehat{m}^{(-k)}(X_i)$ and $\widehat{\ell}^{(-k)}(X_i)$ are computed from data independent of observation i, breaking the dependence that would otherwise arise.

The Neyman-orthogonal score. The DML estimator is based on the Neyman-orthogonal score for the PLR model:

$$\psi(W; \theta, \eta) := (D - m(X))\big\{Y - \ell(X) - \theta(D - m(X))\big\}, \tag{7}$$

where $\eta = (\ell, m)$ denotes the nuisance functions. At the true values (θ_0, η_0), the score reduces to $\psi(W; \theta_0, \eta_0) = V\varepsilon$, which has mean zero by exogeneity. The score is "Neyman-orthogonal" because its expectation is locally insensitive to perturbations in the nuisance parameter—the pathwise derivative with respect to η vanishes at the truth (Appendix A).

The DML estimator. The DML estimator $\widehat{\theta}$ solves the empirical moment condition $\Psi_n(\widehat{\theta}, \widehat{\eta}) = 0$, where $\Psi_n(\theta, \eta) := n^{-1} \sum_{i=1}^n \psi(W_i; \theta, \eta_i)$. Because the score is linear in θ, the solution has the closed form:

$$\widehat{\theta} = \frac{\sum_{i=1}^n \widehat{V}_i \widehat{U}_i}{\sum_{i=1}^n \widehat{V}_i^2}. \tag{8}$$

This is a ratio estimator: the numerator is the sample covariance of the cross-fitted residuals, and the denominator is the sample variance of the treatment residuals.

3.2 The Condition Number: Definition and Interpretation

The condition number of the DML estimating equation quantifies the sensitivity of the estimator to perturbations. It arises naturally from the Jacobian of the score with respect to the target parameter.

The empirical Jacobian. The Jacobian of the empirical score with respect to θ is:

$$\widehat{J}_\theta := \frac{\partial}{\partial \theta} \Psi_n(\theta, \widehat{\eta}) = -\frac{1}{n} \sum_{i=1}^n \widehat{V}_i^2 =: -\widehat{\sigma}_V^2. \tag{9}$$

The Jacobian is negative (the score is decreasing in θ), and its magnitude $|\widehat{J}_\theta| = \widehat{\sigma}_V^2$ equals the sample variance of the cross-fitted treatment residuals. This quantity measures the "curvature" of the moment condition at the root—how quickly the score changes as θ moves away from the solution.

When the Jacobian is small in magnitude, the score is nearly flat, and the estimator is sensitive to small perturbations. This is the numerical analysis interpretation of "ill-conditioning": the inverse Jacobian, which maps score perturbations to parameter perturbations, is large.

Definition 3.1 (DML Condition Number)**.** The **DML condition number** is the inverse of the Jacobian magnitude, normalized by sample size:

$$\boxed{\kappa_{\text{DML}} := \frac{1}{|\widehat{J}_\theta|} = \frac{n}{\sum_{i=1}^n \widehat{V}_i^2} = \frac{1}{\widehat{\sigma}_V^2}.} \tag{10}$$

The condition number has units of $1/\text{Var}(D)$, which complicates cross-study comparisons. A treatment measured in dollars will have a different condition number than the same treatment measured in thousands of dollars, even if the underlying design is identical. We therefore define a scale-invariant version.

Definition 3.2 (Standardized Condition Number)**.** The **standardized condition number** is:

$$\boxed{\kappa^* := \kappa_{\text{DML}} \times \widehat{\sigma}_D^2 = \frac{\widehat{\sigma}_D^2}{\widehat{\sigma}_V^2} = \frac{1}{1 - \widehat{R}^2(D \mid X)},} \tag{11}$$

where $\widehat{\sigma}_D^2 := n^{-1}\sum_i (D_i - \bar{D})^2$ is the sample variance of treatment and $\widehat{R}^2(D \mid X) := 1 - \widehat{\sigma}_V^2/\widehat{\sigma}_D^2$ is the out-of-sample R^2 from the cross-fitted propensity model.

The standardized condition number κ^* is dimensionless and depends only on the proportion of treatment variance explained by covariates. It is precisely the classical **Variance Inflation Factor** (VIF) from regression diagnostics (Belsley et al., 1980). This is not a new statistic; the VIF has been known and used for decades. What is new is its role in governing bias amplification in DML.

Remark 3.1 (Effective Sample Size Interpretation). The quantity κ^* admits a concrete interpretation as an "effective sample size" deflator. If $\kappa^* = 10$, then the residual treatment variation available for identifying θ_0 is equivalent to what would be available in a sample of size $n/10$ with orthogonal treatment assignment. Large κ^* signals that covariates explain most of the treatment variation, leaving little residual variation for identification. An effective sample size of n/κ^* may be far smaller than the nominal sample size n.

Remark 3.2 (Population Condition Number). The population analogue of κ^* is:

$$\kappa := \frac{\text{Var}(D)}{\mathbb{E}[\text{Var}(D \mid X)]} = \frac{\sigma_D^2}{\sigma_V^2} = \frac{1}{1 - R^2(D \mid X)}. \tag{12}$$

By the law of total variance, $\sigma_D^2 = \sigma_V^2 + \text{Var}(m_0(X))$, so $\kappa \geq 1$ with equality when $m_0(X)$ is constant (no selection on observables). As $R^2(D \mid X) \to 1$, we have $\kappa \to \infty$.

3.3 Connection to the Riesz Representer

The condition number has a deep connection to modern semiparametric theory through the Riesz representer. This connection grounds our finite-sample diagnostic in the abstract efficiency theory developed by Chernozhukov et al. (2022).

Definition 3.3 (Riesz Representer for PLR)**.** The Riesz representer for the PLR model is the function $\alpha_0(W) = V/\sigma_V^2$ that satisfies:

1. Orthogonality: $\mathbb{E}[\alpha_0(W) \cdot f(X)] = 0$ for all $f \in L^2(P_X)$;

2. Normalization: $\mathbb{E}[\alpha_0(W) \cdot D] = 1$.

The Riesz representer is the "correction weight" needed to debias the naive regression coefficient. Its norm measures how much correction is required. When the Riesz norm is large, substantial reweighting is needed, and inference becomes fragile.

Theorem 3.1 (Condition Number as Riesz Norm)**.** *The standardized condition number equals the squared L^2 norm of the Riesz representer times the treatment variance:*

$$\boxed{\kappa = \sigma_D^2 \cdot \|\alpha_0\|_{L^2}^2, \quad \textit{where} \quad \|\alpha_0\|_{L^2}^2 = \frac{1}{\sigma_V^2}.} \tag{13}$$

The proof is provided in Appendix A. This result shows that κ^* is not merely a heuristic diagnostic but the finite-sample proxy for a fundamental semiparametric quantity. The Riesz representer framework of Chernozhukov et al. (2022) establishes that $\|\alpha_0\|_{L^2}$ governs the variance of debiased estimators. Our contribution extends this: when nuisance estimation introduces bias, the same quantity governs bias amplification.

3.4 The Exact Finite-Sample Decomposition

We now present the central theoretical result: an exact algebraic decomposition of the DML estimator error that reveals both the variance inflation and bias amplification channels.

Oracle and nuisance bias terms. Define the oracle sampling term:

$$S_n := \frac{1}{n} \sum_{i=1}^{n} V_i \varepsilon_i, \tag{14}$$

which is the sample average of the oracle score—what we would compute if we knew the true nuisance functions. By the central limit theorem, $S_n = O_P(n^{-1/2})$.

Define the nuisance bias term as the difference between the estimated and oracle scores at θ_0:

$$B_n := \Psi_n(\theta_0, \widehat{\eta}) - \Psi_n(\theta_0, \eta_0). \tag{15}$$

This term captures the contamination from nuisance estimation. The Appendix provides an explicit expansion:

$$B_n = \underbrace{\frac{1}{n}\sum_i V_i \Delta_i^{\ell}}_{B_n^{(1)}} - \underbrace{\theta_0 \cdot \frac{1}{n}\sum_i V_i \Delta_i^{m}}_{B_n^{(2)}} + \underbrace{\frac{1}{n}\sum_i \Delta_i^{m} \varepsilon_i}_{B_n^{(3)}} + \underbrace{\frac{1}{n}\sum_i \Delta_i^{m} \Delta_i^{\ell}}_{B_n^{(4)}} - \underbrace{\theta_0 \cdot \frac{1}{n}\sum_i (\Delta_i^{m})^2}_{B_n^{(5)}}, \tag{16}$$

where $\Delta_i^m := m_0(X_i) - \widehat{m}^{(-k)}(X_i)$ and $\Delta_i^{\ell} := \ell_0(X_i) - \widehat{\ell}^{(-k)}(X_i)$ are the nuisance estimation errors. Under Neyman orthogonality and cross-fitting, terms $B_n^{(1)}$–$B_n^{(3)}$ have conditional mean zero; the dominant contribution is the "product term" $B_n^{(4)}$, which drives the product-rate requirement.

Theorem 3.2 (Exact Decomposition)**.** *Under the PLR model (3)–(4), the DML estimator satisfies the*

following ***exact algebraic identity****:*

$$\boxed{\widehat{\theta} - \theta_0 = \widehat{\kappa} \cdot S'_n + \widehat{\kappa} \cdot B'_n,} \tag{17}$$

where:

$$S'_n := \frac{S_n}{\widehat{\sigma}_D^2} \quad \textit{(standardized oracle term)}, \tag{18}$$

$$B'_n := \frac{B_n}{\widehat{\sigma}_D^2} \quad \textit{(standardized nuisance bias term)}. \tag{19}$$

This decomposition involves ***no Taylor approximation****. It is exact because the PLR score (7) is affine in θ.*

Proof Sketch. The DML estimator $\widehat{\theta}$ solves $\Psi_n(\widehat{\theta}, \widehat{\eta}) = 0$. Since the score is linear in θ:

$$\Psi_n(\theta, \widehat{\eta}) = \frac{1}{n} \sum_i \widehat{V}_i \widehat{U}_i - \theta \cdot \widehat{\sigma}_V^2. \tag{20}$$

Solving for $\widehat{\theta}$ and subtracting θ_0:

$$\widehat{\theta} - \theta_0 = \frac{\Psi_n(\theta_0, \widehat{\eta})}{\widehat{\sigma}_V^2} = \kappa_{\mathrm{DML}} \cdot \Psi_n(\theta_0, \widehat{\eta}) = \kappa_{\mathrm{DML}} \cdot (S_n + B_n). \tag{21}$$

Multiplying and dividing by $\widehat{\sigma}_D^2$ yields the standardized form (17). The complete proof is in Appendix A. □

3.5 Interpretation: Variance Inflation versus Bias Amplification

The exact decomposition (17) reveals that the condition number $\widehat{\kappa}$ multiplies *both* sources of error. This is the crucial insight of the paper.

Variance inflation: the term $\widehat{\kappa} \cdot S'_n$. The oracle sampling term $S_n = O_P(n^{-1/2})$ represents the irreducible sampling uncertainty from having a finite sample. When the condition number is large, this sampling uncertainty is amplified: the variance of the DML estimator scales as κ/n rather than $1/n$. This is the classical VIF effect from regression theory. Confidence intervals widen, but they widen *honestly*—the standard error estimator correctly captures this inflation, and coverage probabilities remain near nominal levels.

Bias amplification: the term $\widehat{\kappa} \cdot B'_n$. The nuisance bias term B_n represents the contamination from imperfect nuisance estimation. Under the DML product-rate condition, $B_n = O_P(r_n)$ where $r_n = \|\widehat{m} - m_0\|_{L^2} \cdot \|\widehat{\ell} - \ell_0\|_{L^2}$. In asymptotic theory, this rate is assumed to satisfy $r_n = o(n^{-1/2})$, making the bias negligible.

The key insight is that the *effective* bias is:

$$\widehat{\kappa} \cdot B'_n = O_P(\kappa \cdot r_n). \tag{22}$$

When κ is large, even a "small" product rate r_n can produce a first-order bias. The critical condition for

valid inference is not just $r_n = o(n^{-1/2})$ but:

$$\kappa \cdot r_n = o(n^{-1/2}). \tag{23}$$

A concrete numerical example. Suppose $n = 1{,}000$, $\kappa = 50$, and the nuisance learners achieve rates $\|\widehat{m} - m_0\|_{L^2} = \|\widehat{\ell} - \ell_0\|_{L^2} = n^{-1/4} \approx 0.18$. Then:

- The product rate is $r_n = n^{-1/2} = 0.032$.
- The standard DML asymptotic analysis treats $r_n = O(n^{-1/2})$ as negligible.
- But the *amplified* bias is $\kappa \cdot r_n = 50 \times 0.032 = 1.6$.
- The oracle standard error is approximately $1/\sqrt{n} = 0.032$.
- The bias-to-standard-error ratio is $1.6/0.032 = 50$.

A bias of 50 standard errors produces catastrophic undercoverage. The confidence interval is centered on a point 50 standard errors away from the truth.

This example illustrates why regularized learners are more fragile in ill-conditioned designs. By trading variance for bias, they are exposed to the bias amplification channel. A machine learning method that appears "robust" when κ is small can fail catastrophically when κ is large.

Why standard errors fail to detect the problem. The standard DML variance estimator is:

$$\widehat{\sigma}^2_{\widehat{\theta}} = \frac{\sum_i \widehat{V}_i^2 \hat{\varepsilon}_i^2}{(\sum_i \widehat{V}_i^2)^2}, \tag{24}$$

where $\hat{\varepsilon}_i = \widehat{U}_i - \widehat{\theta}\widehat{V}_i$. This estimator correctly captures the sampling variance, including the variance inflation effect. But it does *not* account for the systematic shift caused by the amplified bias term $\kappa \cdot B_n$. The confidence intervals remain narrow because they measure the wrong quantity: they measure sampling uncertainty, not total uncertainty including bias. This is why bias amplification is a "silent" failure—the standard errors give no warning.

3.6 Finite-Sample Probability Bounds

We formalize the preceding discussion with a finite-sample probability bound.

Theorem 3.3 (Finite-Sample Error Bound)**.** *Suppose the condition number satisfies $\kappa_n = O(n^\gamma)$ for some $\gamma \geq 0$, and the nuisance estimators satisfy $\|\widehat{m} - m_0\|_{L^2} = O_P(r_n^m)$ and $\|\widehat{\ell} - \ell_0\|_{L^2} = O_P(r_n^\ell)$ with product rate $r_n = r_n^m \cdot r_n^\ell$. Then:*

$$\boxed{\widehat{\theta} - \theta_0 = O_P\left(\frac{\kappa_n}{\sqrt{n}} + \kappa_n \cdot r_n\right).} \tag{25}$$

*The first term is **variance inflation**; the second is **bias amplification**.*

The proof is in Appendix A. This bound makes precise the claim that both channels scale with the condition number. For root-n inference to hold, we need both terms to be $O_P(n^{-1/2})$. The variance term is automatically $O_P(n^{-1/2})$ when $\kappa_n = O(1)$. The bias term requires $\kappa_n \cdot r_n = o(n^{-1/2})$.

Corollary 3.4 (Critical Rate for Valid Inference)**.** *For $\sqrt{n}$-consistent inference to hold, the following condition is necessary:*

$$\kappa_n \cdot r_n = o(n^{-1/2}). \tag{26}$$

Rearranging: $r_n = o(1/(\kappa_n\sqrt{n}))$. When κ_n is large, nuisance estimators must converge faster to maintain valid inference.

4 Diagnostics and Conditioning Regimes

This section translates the theoretical results into practical guidance. We formalize the conditioning regimes, discuss the effective sample size concept, and provide recommendations for practitioners.

4.1 Conditioning Regimes

Based on the theoretical analysis in Section 3, we define three conditioning regimes using the standardized condition number κ^*. These thresholds are analogous to the rule-of-thumb that $F < 10$ indicates weak instruments (Stock and Yogo, 2005). They are not bright lines but guidelines for interpretation, calibrated against both theoretical considerations and Monte Carlo evidence.

Definition 4.1 (Conditioning Regimes)**.**

- **Well-conditioned** ($\kappa^* < 5$): Standard DML asymptotics apply. The product $\kappa \cdot r_n$ is typically negligible for reasonably accurate nuisance learners. Coverage should be near nominal. The effective sample size n/κ^* is at least 20% of the nominal sample size. Point estimates should be stable across learners with similar accuracy.
- **Moderately ill-conditioned** ($5 \leq \kappa^* \leq 20$): Bias amplification begins to matter. Coverage may deteriorate for learners with substantial regularization bias, even if nuisance accuracy is acceptable by standard criteria. Confidence intervals may be wider than expected, reflecting variance inflation, but may still undercover due to residual bias. Sensitivity analysis across learners with different bias-variance tradeoffs is strongly recommended. The effective sample size is between 5% and 20% of nominal.
- **Severely ill-conditioned** ($\kappa^* > 20$): Bias domination is likely for regularized learners. Standard DML confidence intervals should not be trusted without additional robustness checks. Point estimates may vary substantially across learners, potentially spanning zero or reversing sign. The effective sample size is below 5% of nominal. Consider overlap-aware methods such as trimming (Crump et al., 2009), weak-overlap-robust estimators (Ma et al., 2023), or alternative estimands.

These thresholds are conservative. The threshold $\kappa^* = 5$ corresponds to $R^2(D \mid X) = 0.80$—covariates explain 80% of treatment variation. The threshold $\kappa^* = 20$ corresponds to $R^2(D \mid X) = 0.95$—covariates explain 95% of treatment variation. Both are high by the standards of many empirical applications, yet even at $\kappa^* = 5$, bias amplification is detectable in simulations for learners with moderate regularization bias.

4.2 Effective Sample Size

The condition number has a natural interpretation as an effective sample size deflator. This interpretation provides intuitive guidance for practitioners.

Definition 4.2 (Effective Sample Size)**.** The **effective sample size** for identifying θ_0 is:

$$N_{\text{eff}} := \frac{n}{\kappa^*}. \tag{27}$$

The effective sample size measures how much identifying variation is available after accounting for the predictability of treatment. Consider two designs:

- **Randomized experiment:** Treatment is assigned independently of covariates. Then $R^2(D \mid X) \approx 0$, $\kappa^* \approx 1$, and $N_{\text{eff}} \approx n$. All observations contribute identifying variation.
- **Observational study with $\kappa^* = 25$:** Covariates explain 96% of treatment variation. The effective sample size is $N_{\text{eff}} = n/25$. A study with $n = 2{,}500$ observations has only $N_{\text{eff}} = 100$ effective observations for identifying θ_0.

The effective sample size provides intuition for why estimates can be unstable in severely ill-conditioned designs: the data simply do not contain enough residual variation to pin down the treatment effect precisely.

Remark 4.1 (Relationship to Precision). The effective sample size is related to the precision of the DML estimator. Under homoskedasticity, the variance of $\widehat{\theta}$ is approximately:

$$\text{Var}(\widehat{\theta}) \approx \frac{\sigma_\varepsilon^2}{n \cdot \sigma_V^2} = \frac{\sigma_\varepsilon^2 \kappa}{n \cdot \sigma_D^2} = \frac{\sigma_\varepsilon^2}{\sigma_D^2} \cdot \frac{1}{N_{\text{eff}}}. \tag{28}$$

The precision is governed by the effective sample size, not the nominal sample size.

4.3 Convergence Rates by Regime

The finite-sample bound in Theorem 3.3 implies different convergence rates in each regime.

Corollary 4.1 (Effective Rates by Regime)**.** *Suppose nuisance estimators achieve the product rate $r_n = n^{-\alpha}$ for some $\alpha > 0$. Then under Theorem 3.3:*

(i) ***Well-conditioned*** *($\kappa_n = O(1)$):*

$$\widehat{\theta} - \theta_0 = O_P(n^{-1/2}). \tag{29}$$

Standard root-n asymptotics apply. The bias term $\kappa_n \cdot r_n = O(n^{-\alpha})$ is negligible relative to the variance term $O(n^{-1/2})$ whenever $\alpha > 1/2$.

(ii) ***Moderately ill-conditioned*** *($\kappa_n = O(n^\beta)$ for $0 < \beta < 1/2$):*

$$\widehat{\theta} - \theta_0 = O_P(n^{\beta - 1/2} + n^{\beta - \alpha}). \tag{30}$$

The oracle term degrades to $n^{\beta-1/2} \to 0$ (slower than root-n). The bias term $n^{\beta-\alpha}$ vanishes if $\alpha > \beta$ but can dominate if $\alpha < \beta$.

(iii) ***Severely ill-conditioned*** *($\kappa_n \asymp \sqrt{n}$):*

$$\widehat{\theta} - \theta_0 = O_P(1). \tag{31}$$

The estimator does not converge. Even the oracle term is $O_P(1)$—variance inflation alone prevents convergence. If $\alpha < 1/2$ (typical for nonparametric estimators in moderate dimensions), the bias term $n^{1/2-\alpha} \to \infty$: bias diverges.

This corollary shows that the "critical frontier" where standard inference breaks down is $\kappa_n \asymp \sqrt{n}$. At this frontier, even with oracle nuisance functions ($B_n = 0$), the estimator variance does not shrink. With regularized learners, the situation is worse: bias can diverge.

4.4 Connection to the Semiparametric Efficiency Bound

The condition number is intimately connected to the semiparametric efficiency bound, providing theoretical grounding for the diagnostic.

The semiparametric efficiency bound for estimating θ_0 in the PLR model is:

$$V_{\text{eff}} = \frac{\mathbb{E}[V^2\varepsilon^2]}{(\mathbb{E}[V^2])^2}. \tag{32}$$

Under homoskedasticity ($\mathbb{E}[\varepsilon^2 \mid X] = \sigma_\varepsilon^2$), this simplifies to:

$$V_{\text{eff}} = \frac{\sigma_\varepsilon^2}{\sigma_V^2} = \frac{\sigma_\varepsilon^2 \kappa}{\sigma_D^2}. \tag{33}$$

As $\mathbb{E}[V^2] \to 0$ (i.e., $R^2(D \mid X) \to 1$), the efficiency bound diverges: $V_{\text{eff}} \to \infty$. The treatment effect becomes unidentifiable in the limit, and no estimator—however clever—can achieve finite variance. The condition number κ quantifies distance from this boundary.

Theorem 4.2 (Efficiency Bound and Condition Number)**.** *Under homoskedasticity, the semiparametric efficiency bound satisfies:*

$$V_{\text{eff}} = \frac{\sigma_\varepsilon^2 \kappa}{\sigma_D^2}. \tag{34}$$

The condition number κ is thus the factor by which the efficiency bound exceeds its minimum value (achieved at $\kappa = 1$).

This connection explains why κ^* is the "right" diagnostic: it directly measures the inflation in the efficiency bound relative to an orthogonal-treatment benchmark. Large κ^* signals proximity to the identification boundary where no estimator can work well.

4.5 Practical Guidance for Reporting

We conclude this section with practical recommendations for researchers using DML.

1. **Always compute and report κ^*.** The standardized condition number should be computed as $\kappa^* =$

$1/(1 - \widehat{R}^2(D \mid X))$ using the out-of-sample R^2 from the cross-fitted propensity model. Report it in the main results table alongside point estimates and confidence intervals.

2. **Interpret the effective sample size.** For quick intuition, divide the nominal sample size by κ^*. If $N_{\text{eff}} < 100$, question whether the data contain enough identifying variation for reliable inference.

3. **Compare estimates across learners.** If $\kappa^* > 5$, run DML with at least two learners: one with low bias (e.g., OLS if the nuisance model is approximately linear) and one with high flexibility (e.g., Random Forest). Large discrepancies between estimates are a red flag for bias amplification.

4. **Examine the bias-to-SE ratio.** In Monte Carlo simulations or sensitivity analyses, compute the ratio of estimated bias to standard error. A ratio exceeding 0.5 suggests that bias may be a substantial fraction of the confidence interval width.

5. **Consider robustness to trimming.** In severely ill-conditioned designs, re-estimate after trimming observations with extreme predicted treatment propensities. If estimates change substantially, the original design is fragile.

6. **Be skeptical of narrow confidence intervals.** In ill-conditioned designs, narrow confidence intervals are not reassuring—they may reflect bias amplification masquerading as precision. The standard errors measure sampling variance, not total uncertainty including bias.

5 Monte Carlo Simulations

This section uses Monte Carlo experiments to validate the theoretical predictions and illustrate the distinct failure modes in different conditioning regimes.

5.1 Data Generating Process

We generate data from the PLR model with the following specification. Let $X = (X_1, \ldots, X_{10}) \in \mathbb{R}^{10}$ be drawn from a Gaussian AR(1) process:

$$X_j = \rho X_{j-1} + \sqrt{1 - \rho^2}\, \eta_j, \quad \eta_j \overset{\text{iid}}{\sim} N(0,1), \quad X_0 = 0, \quad \rho = 0.5. \tag{35}$$

The treatment equation is:

$$D = \beta^\top X + \sigma_U \cdot U, \quad U \sim N(0,1), \tag{36}$$

where $\beta = (0.7, 0.7^2, \ldots, 0.7^{10})^\top$ and σ_U is calibrated to achieve the target $R^2(D \mid X)$.

The outcome equation is:

$$Y = \theta_0 D + g_0(X) + \varepsilon, \quad \varepsilon \sim N(0,1), \tag{37}$$

where $\theta_0 = 1$ and $g_0(X) = X_1 + X_2^2 + X_3 X_4$.

We calibrate three overlap levels:

- **High overlap:** $R^2(D \mid X) = 0.75$, yielding $\kappa^* \approx 4$.

- **Moderate overlap:** $R^2(D \mid X) = 0.90$, yielding $\kappa^* \approx 10$.
- **Low overlap:** $R^2(D \mid X) = 0.97$, yielding $\kappa^* \approx 33$.

5.2 Nuisance Learners

We consider three nuisance learners representing different points on the bias-variance spectrum:

- **OLS (LIN):** Linear regression using all covariates. Unbiased when the true nuisance functions are linear. Provides the benchmark for variance-only failure.
- **Lasso (LAS):** ℓ_1-penalised regression with cross-validated penalty. Introduces shrinkage bias but maintains sparsity. In our linear DGP, Lasso bias is minimal.
- **Random Forest (RF):** Ensemble of regression trees with default hyperparameters. Introduces regularisation bias through tree depth and sample constraints. This is the primary example of bias amplification.

All specifications use $K = 5$-fold cross-fitting. For each configuration, we run $B = 500$ Monte Carlo replications with sample sizes $n \in \{500, 2000\}$.

5.3 Results: Two Failure Modes

Table 1 summarises the key findings.

Table 1: Monte Carlo Results by Overlap Level and Learner

Overlap (κ^*)	Learner	Coverage (%)	CI Length	Bias	RMSE	Failure Mode
High (≈ 4)	OLS	95.1	0.145	0.001	0.037	—
	Lasso	94.0	0.146	0.000	0.037	—
	RF	89.0	0.131	−0.023	0.038	Mild bias
Moderate (≈ 10)	OLS	94.8	0.242	0.001	0.062	—
	Lasso	94.2	0.244	0.000	0.063	—
	RF	84.1	0.265	−0.059	0.079	Bias amplification
Low (≈ 33)	OLS	95.0	0.363	−0.000	0.094	Variance inflation
	Lasso	94.4	0.364	−0.000	0.096	Variance inflation
	RF	39.8	0.191	−0.103	0.109	**Severe bias**

Notes: $B = 500$ replications pooled across $n \in \{500, 2000\}$. Coverage is the proportion of 95% CIs containing $\theta_0 = 1$. "Failure Mode" summarises the dominant source of inference degradation.

5.4 Discussion: Understanding the Failure Modes

The results illustrate the two distinct failure modes predicted by theory:

Variance Inflation (OLS/Lasso). For unbiased learners, increasing κ^* causes confidence intervals to widen: from 0.145 at $\kappa^* \approx 4$ to 0.363 at $\kappa^* \approx 33$. Coverage remains near 95% because the intervals honestly reflect the increased uncertainty. This is the classical VIF effect: the term $\kappa_{\text{DML}} S_n$ dominates, but the standard error estimator correctly captures this inflation.

Bias Amplification (Random Forest). For biased learners, the pattern is strikingly different. At $\kappa^* \approx 33$:

- Coverage collapses to 39.8%—catastrophic undercoverage.
- Confidence intervals remain *narrow* (0.191 vs. 0.363 for OLS).
- Point estimates are systematically biased by -0.10.

This is bias amplification in action. The Random Forest introduces regularisation bias B_n that is "small" in absolute terms but non-zero. When multiplied by the large κ_{DML}, this bias becomes first-order:

$$\kappa_{\text{DML}} \cdot B_n \approx 33 \times (-0.003) \approx -0.10.$$

The Bias-to-Standard-Error Ratio: The Smoking Gun. The key diagnostic is the ratio of bias to standard error. For Random Forests at $\kappa^* \approx 33$, the bias is -0.103 and the standard error (implied by CI length) is approximately $0.191/3.92 \approx 0.049$. Thus, the bias-to-SE ratio is:

$$\frac{|\text{Bias}|}{\text{SE}} \approx \frac{0.103}{0.049} \approx 2.1.$$

A bias exceeding two standard errors shifts the confidence interval entirely off the truth, explaining the catastrophic undercoverage. The confidence interval, based on the standard error that captures only sampling variance, fails to account for this systematic shift.

Why Random Forests are more fragile. The key insight is that flexible learners are *more fragile* in ill-conditioned designs, not less. By trading variance for bias through regularisation, they are exposed to the bias amplification channel. A machine learning method that appears "robust" in well-conditioned settings can fail catastrophically when κ^* is large.

6 Empirical Application: LaLonde Job-Training Data

We illustrate the diagnostic using the canonical job-training study of LaLonde (1986), which has become a litmus test for causal inference methods.

6.1 Data and Design

The outcome Y is real earnings in 1978. The treatment D is participation in the National Supported Work (NSW) job-training program. Covariates X include age, education, race, marital status, and earnings in 1974–1975.

We consider two designs:

- **Experimental sample** ($n = 445$): The original randomised experiment. Treatment is assigned independently of covariates, yielding $\kappa^* \approx 4$.

- **Observational sample** ($n = 2{,}675$): NSW treated units combined with a non-experimental comparison group from the Panel Study of Income Dynamics (PSID-1). Treatment is highly predictable from covariates, yielding $\kappa^* > 20$.

For each design, we estimate the PLR-DML model using OLS, Lasso, and Random Forest for nuisance functions with $K = 5$-fold cross-fitting.

6.2 Results

Table 2 reports the main findings.

Table 2: DML Estimates for LaLonde (1986) Job-Training Data

Design	Learner	$\widehat{\theta}$ (\$)	95% CI (\$)	κ^*
Experimental	OLS	1,752	[443, 3,060]	4.0
	Lasso	1,793	[475, 3,111]	4.1
	RF	1,455	[213, 2,698]	3.9
Observational	OLS	621	[−921, 2,163]	21.9
	Lasso	56	[−1,196, 1,307]	15.7
	RF	−642	[−2,438, 1,153]	39.9

Notes: Experimental benchmark ≈ \$1,794 (LaLonde, 1986). Experimental sample: randomised NSW treatment group vs. NSW control group. Observational sample: NSW treated units vs. PSID-1 comparison group.

6.3 Interpretation

Experimental sample. The standardised condition number is $\kappa^* \approx 4$ across all learners—well within the "well-conditioned" regime. Point estimates are stable across learners (\$1,455–\$1,793) and confidence intervals, while wide, consistently exclude zero. The estimates align with the experimental benchmark of approximately \$1,794.

Observational sample. The standardised condition number increases dramatically to $\kappa^* \in [16, 40]$—squarely in the "severely ill-conditioned" regime. The consequences are immediate and stark:

- Point estimates range from −\$642 (RF) to +\$621 (OLS), spanning zero and reversing sign.
- Confidence intervals are very wide and fail to exclude zero for all learners.
- The Random Forest estimate is negative, contradicting the experimental benchmark entirely.

Effective Sample Size. The observational sample contains $n = 2{,}675$ observations, but with $\kappa^* \approx 40$ for Random Forests, the *effective sample size* for identification is merely $n/\kappa^* \approx 2{,}675/40 \approx 67$ observations. This stark reduction explains the instability: the data simply lack sufficient residual treatment variation to identify θ_0 reliably. The observed fluctuations are not merely "noise"—they reflect *bias amplification* driven by the poor condition number.

The κ^* diagnostic immediately distinguishes the two settings:

- Experimental design: $\kappa^* \approx 4$ (well-conditioned). Inference is reliable.

- Observational design: $\kappa^* > 20$ (severely ill-conditioned). Inference is fragile.

7 Conclusion and Practical Recommendations

Double Machine Learning is not magic. It requires overlap. This paper has shown that the conditioning of the orthogonal score—summarised by the standardised condition number κ^*—is a critical determinant of finite-sample reliability.

7.1 Summary of Findings

Standard OLS theory focuses on how poor conditioning inflates variance. We have demonstrated a second channel specific to DML: *bias amplification*. When flexible learners introduce regularisation bias, high κ^* multiplies that bias into a first-order distortion. The result is undercoverage, not variance inflation.

The exact decomposition $\widehat{\theta} - \theta_0 = \kappa_{\text{DML}}(S_n + B_n)$ reveals that the condition number affects both error channels. For unbiased learners, only the variance term matters; for biased learners, the bias term can dominate.

7.2 Practitioner's Checklist

We recommend the following workflow for applied DML analyses:

1. **Compute κ^* before interpreting estimates.** The standardised condition number should be computed as $\kappa^* = 1/(1 - R^2(D \mid X))$ using the cross-fitted treatment residuals.

2. **Report κ^* in main tables.** Just as first-stage F-statistics accompany IV results, κ^* should accompany all DML point estimates and confidence intervals.

3. **Use κ^* to guide sensitivity analysis.** If $\kappa^* > 10$, compare estimates across learners with different bias properties. Large variation is a red flag.

4. **Interpret with caution when $\kappa^* > 20$.** In severely ill-conditioned designs, DML confidence intervals should not be trusted unless the learner is known to be unbiased. Consider overlap trimming (Crump et al., 2009) or alternative estimands (Ma et al., 2023).

5. **Document conditioning in robustness checks.** Report how κ^* changes under alternative specifications (different covariates, trimming rules, learner choices).

The introduction of κ^* as a standard diagnostic aims to bring to DML the same transparency that the F-statistic brought to instrumental variables regression. Asymptotic DML theory remains valid in regular regimes, but its practical reliability hinges on conditioning. Making κ^* a routine part of DML reporting is a simple step toward more transparent and reliable empirical work.

References

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

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.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

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.

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

Chernozhukov, V., Newey, W. K., and Singh, R. (2022). Automatic debiased machine learning via Riesz representers. *Journal of Econometrics*, 226(1):274–302.

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

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.

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

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29.

Jung, Y. (2023). A short note on finite sample analysis on double/debiased machine learning. Manuscript, Purdue University.

Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

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., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Ma, Y., Sant'Anna, P. H., Sasaki, Y., and Ura, T. (2023). Doubly robust estimators with weak overlap. arXiv preprint arXiv:2304.02036.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048.

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

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.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with 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 linear IV regression. In Andrews, D. W. K. and Stock, J. H. (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80–108. Cambridge University Press.

Wüthrich, K. and Zhu, Y. (2024). Omitted variable bias meets machine learning. *Journal of Econometrics*, forthcoming.

A Mathematical Appendix

This appendix provides detailed proofs for the main theoretical results. For complete derivations with full mathematical rigor, including the probability space, function spaces, and stochastic order notation, see the companion Mathematical Appendix document.

A.1 Proof of Theorem 3.2 (Exact Decomposition)

Proof. The DML estimator $\widehat{\theta}$ solves $\Psi_n(\widehat{\theta}, \widehat{\eta}) = 0$. Since the score is linear in θ:

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

Solving for $\widehat{\theta}$ and subtracting θ_0:

$$\begin{aligned}\widehat{\theta} - \theta_0 &= \frac{\sum_i \widehat{V}_i(\widehat{U}_i - \theta_0\widehat{V}_i)}{\sum_i \widehat{V}_i^2} \\ &= \kappa_{\mathrm{DML}} \cdot \Psi_n(\theta_0, \widehat{\eta}).\end{aligned}$$

Now decompose $\Psi_n(\theta_0, \widehat{\eta})$. Define population residuals $V_i := D_i - m_0(X_i)$ and $U_i := Y_i - \ell_0(X_i)$. By the residualization identity, $U_i = \theta_0 V_i + \varepsilon_i$.

The cross-fitted score at θ_0 is:

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

Add and subtract the population quantities:

$$\widehat{V}_i = V_i + (m_0(X_i) - \widehat{m}(X_i)) = V_i - \Delta_i^m,$$
$$\widehat{U}_i = U_i + (\ell_0(X_i) - \widehat{\ell}(X_i)) = U_i - \Delta_i^\ell,$$

where $\Delta_i^m := \widehat{m}(X_i) - m_0(X_i)$ and $\Delta_i^\ell := \widehat{\ell}(X_i) - \ell_0(X_i)$.

Substituting and using $U_i - \theta_0 V_i = \varepsilon_i$:

$$\begin{aligned}\widehat{V}_i(\widehat{U}_i - \theta_0\widehat{V}_i) &= (V_i - \Delta_i^m)[(U_i - \Delta_i^\ell) - \theta_0(V_i - \Delta_i^m)] \\ &= (V_i - \Delta_i^m)[\varepsilon_i - \Delta_i^\ell + \theta_0\Delta_i^m] \\ &= V_i\varepsilon_i - V_i\Delta_i^\ell + \theta_0 V_i\Delta_i^m \\ &\quad - \Delta_i^m\varepsilon_i + \Delta_i^m\Delta_i^\ell - \theta_0(\Delta_i^m)^2.\end{aligned}$$

Averaging and using the definitions of S_n and B_n:

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

where $S_n = n^{-1}\sum_i V_i\varepsilon_i$ is the oracle term and B_n collects all terms involving nuisance estimation error.

Therefore:

$$\widehat{\theta} - \theta_0 = \kappa_{\mathrm{DML}}(S_n + B_n).$$

This decomposition is **exact**—not an approximation—because the score (7) is affine in θ. There is no Taylor remainder. □

A.2 Proof of Theorem 3.3 (Finite-Sample Error Bound)

Proof. From Theorem 3.2:

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

Sampling term. By the CLT, $S_n = n^{-1}\sum_i V_i\varepsilon_i$ satisfies $\sqrt{n}S_n \xrightarrow{d} N(0, \sigma_\psi^2)$ where $\sigma_\psi^2 = \mathbb{E}[V^2\varepsilon^2]$. Thus $S_n = O_P(n^{-1/2})$, and:

$$\widehat{\kappa}S_n' = O_P\left(\frac{\kappa_n}{\sqrt{n}}\right).$$

Bias term. By Neyman orthogonality and the product-rate condition, $B_n = O_P(r_n)$ where $r_n = \|\widehat{m} - m_0\|_{L^2} \cdot \|\widehat{\ell} - \ell_0\|_{L^2}$. Under standard DML conditions, $r_n = o_P(n^{-1/2})$. Thus:

$$\widehat{\kappa}B_n' = O_P(\kappa_n \cdot r_n).$$

Combining:

$$\widehat{\theta} - \theta_0 = O_P\left(\frac{\kappa_n}{\sqrt{n}} + \kappa_n \cdot r_n\right).$$

The nuisance bias term is negligible relative to the sampling term if and only if $\kappa_n \cdot r_n = o_P(n^{-1/2})$. □

A.3 Cross-Fitting Algorithm

The DML estimator is computed as follows:

1. **Partition the data.** Randomly split $\{1, \ldots, n\}$ into K folds $I_1, \ldots, I_K$ of approximately equal size.
2. **Train nuisance estimators.** For each fold $k = 1, \ldots, K$:
   - Train $\widehat{m}^{(-k)}$ on $\{(D_i, X_i) : i \notin I_k\}$.
   - Train $\widehat{\ell}^{(-k)}$ on $\{(Y_i, X_i) : i \notin I_k\}$.
3. **Compute cross-fitted residuals.** For each $i \in I_k$:
   - $\widehat{V}_i = D_i - \widehat{m}^{(-k)}(X_i)$.
   - $\widehat{U}_i = Y_i - \widehat{\ell}^{(-k)}(X_i)$.
4. **Compute the DML estimator.**
$$\widehat{\theta} = \frac{\sum_{i=1}^{n} \widehat{V}_i \widehat{U}_i}{\sum_{i=1}^{n} \widehat{V}_i^2}.$$
5. **Compute the standard error.**
$$\widehat{\mathrm{SE}} = \sqrt{\frac{\sum_{i=1}^{n} \widehat{V}_i^2 \hat{\varepsilon}_i^2}{(\sum_{i=1}^{n} \widehat{V}_i^2)^2}},$$
   where $\hat{\varepsilon}_i = \widehat{U}_i - \widehat{\theta}\widehat{V}_i$.
6. **Compute the condition number.**
$$\kappa^* = \frac{n \cdot \widehat{\mathrm{Var}}(D)}{\sum_{i=1}^{n} \widehat{V}_i^2} = \frac{1}{1 - \widehat{R}^2(D \mid X)}.$$