Finite-Sample Failures and Condition-Number Diagnostics in Double Machine Learning

Gabriel Saco*

Abstract

Standard Double Machine Learning (DML; Chernozhukov et al. (2018)) confidence intervals can exhibit severe finite-sample distortions when the underlying score equations are ill-conditioned, even if nuisance functions are estimated with state-of-the-art methods. This paper shows that a simple, easily computed condition number for the partially linear regression score, $\kappa_{\mathrm{DML}} := 1/|\hat{J}_\theta|$, largely determines when DML inference is reliable. Our first result provides a nonasymptotic coverage error bound for the usual DML t-statistic of order $n^{-1/2} + \sqrt{n}\, r_n$, where r_n summarizes nuisance estimation error. Our second result gives a refined linearization in which both estimation error and confidence interval length scale as $\kappa_{\mathrm{DML}}/\sqrt{n} + \kappa_{\mathrm{DML}} r_n$. These expansions yield three conditioning regimes—well-conditioned, moderately ill-conditioned, and severely ill-conditioned—and imply that valid, shrinking confidence sets require $\kappa_{\mathrm{DML}} = o_{\mathrm{p}}(\sqrt{n})$ and $\kappa_{\mathrm{DML}} r_n \to 0$. Monte Carlo experiments calibrated to typical PLR DML applications via overlap-based $R^2(D \mid X)$ targets show that designs with $\kappa_{\mathrm{DML}} < 1$ deliver near-nominal coverage of 95% intervals, whereas severely ill-conditioned designs with $\kappa_{\mathrm{DML}} \approx 2$–$3$ can suffer substantial undercoverage (coverage around 60% for nominal 95% intervals at $n = 2000$). We therefore propose reporting κ_{DML} alongside DML estimates as a routine diagnostic, in the same spirit as condition-number checks or weak-instrument diagnostics in instrumental variables settings.

Keywords: Double Machine Learning, Orthogonal Scores, Condition-Number Diagnostics, Nonasymptotic Inference, Finite-Sample Coverage, Partially Linear Regression

1 Introduction

Double Machine Learning (DML), introduced by Chernozhukov et al. (2018), provides a principled framework for inference on low-dimensional target parameters in the presence of high-dimensional nuisance functions. The method combines *Neyman-orthogonal scores*—which are locally insensitive to first-order nuisance errors—with *cross-fitting* to accommodate flexible machine learning estimators.[1] Under appropriate regularity conditions, DML estimators are $\sqrt{n}$-consistent and asymptotically normal, extending classical PLR results (Robinson, 1988) and high-dimensional post-selection inference (Belloni et al., 2014; Farrell, 2015).

Asymptotic guarantees, however, provide limited guidance for finite-sample reliability. When the *empirical Jacobian* of the orthogonal score is nearly singular, the usual normal approximation may be misleading even at

*Universidad del Pacífico, Lima, Peru. Email: ga.sacoa@up.edu.pe.

[1]The general construction of orthogonal / locally robust moments in semiparametric GMM is developed by Chernozhukov et al. (2022), building on earlier influence-function work in semiparametric efficiency. See also Newey and Robins (2018) for cross-fit remainder-rate theory.

apparently "large" sample sizes. This short communication isolates and quantifies this conditioning problem in the canonical PLR setting, in the spirit of weak-instrument diagnostics in IV regression (Staiger and Stock, 1997; Stock and Yogo, 2005).[2]

Contributions

We make three contributions.

1. Coverage error bound and κ-amplified linearization. We introduce the DML condition number

$$\kappa_{\mathrm{DML}} := \frac{1}{|\hat{J}_\theta|}$$

and prove that the coverage error of standard DML confidence intervals satisfies

$$\left|\mathbb{P}\big(\theta_0 \in \mathrm{CI}_{\mathrm{std}}\big) - (1-\alpha)\right| \le \frac{C_1}{\sqrt{n}} + C_2\sqrt{n}\, r_n + o(1), \tag{1}$$

for constants $C_1, C_2 > 0$ and a sequence r_n that captures nuisance and remainder terms. At the same time, a refined linearization shows that

$$\hat{\theta} - \theta_0 = \kappa_{\mathrm{DML}}\{S_n + B_n\} + R_n,$$

so that the parameter-scale error and confidence interval length obey

$$|\hat{\theta} - \theta_0| = O_P\Big(\frac{\kappa_{\mathrm{DML}}}{\sqrt{n}} + \kappa_{\mathrm{DML}} r_n\Big).$$

Thus the condition number directly amplifies both variance and bias in θ-space. Our argument is complementary to the general finite-sample DML theory in Chernozhukov et al. (2023) and recent joint-coverage guarantees for high-dimensional DML (Quintas-Martinez, 2022; Jung, 2023).

2. Regime characterization. We show that meaningful DML inference requires controlling the growth of $\kappa_n := \kappa_{\mathrm{DML}}$. In particular, confidence sets shrink only if $\kappa_n = o_P(\sqrt{n})$ and the bias term $\kappa_n r_n$ vanishes. This yields a classification into well-conditioned, moderately ill-conditioned, and severely ill-conditioned regimes, paralleling strong vs. weak identification regimes in IV (Staiger and Stock, 1997; Stock and Yogo, 2005).

3. Simulation evidence. Monte Carlo experiments confirm that κ_{DML} predicts coverage failures: designs with $\kappa_{\mathrm{DML}} \approx 5.6$ exhibit 8.8% coverage of nominal 95% intervals at $n = 2000$. The patterns are consistent with our theoretical decomposition and echo the finite-sample distortions documented in simulation work on DML and doubly robust estimators (e.g., Chernozhukov et al., 2018; Farrell, 2015; Quintas-Martinez, 2022).

We propose κ_{DML} as a simple diagnostic for DML reliability, analogous to first-stage F-statistics in instrumental variables. Developing robust, conditioning-aware inference procedures is left for future work.

[2]For many-instrument settings with sparse first stages, see Belloni et al. (2012), which motivates thinking of first-stage strength and condition numbers jointly.

2 Setup: PLR Model and DML Estimator

We consider the canonical Partially Linear Regression (PLR) model. Observations $W_i = (Y_i, D_i, X_i)$, $i = 1, \dots, n$, are drawn i.i.d. from a distribution P, where $Y_i \in \mathbb{R}$ is the outcome, $D_i \in \mathbb{R}$ is a scalar treatment or policy variable, and $X_i \in \mathbb{R}^p$ is a vector of controls or confounders. The structural model is

$$Y = D\theta_0 + g_0(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid D, X] = 0, \tag{2}$$

where $\theta_0 \in \mathbb{R}$ is the scalar parameter of interest and $g_0 : \mathbb{R}^p \to \mathbb{R}$ is an unknown nuisance function. This model goes back to the semiparametric PLR formulation of Robinson (1988) and has been extensively studied in high-dimensional settings (e.g., Belloni et al., 2014; Farrell, 2015).

Define the nuisance regression functions

$$m_0(X) := \mathbb{E}[D \mid X], \qquad \ell_0(X) := \mathbb{E}[Y \mid X].$$

Using (2), one has

$$\ell_0(X) = \theta_0 m_0(X) + g_0(X),$$

so that ℓ_0 encodes both the structural and reduced-form components.

Orthogonal score. For PLR, the standard Neyman-orthogonal score is

$$\psi(W; \theta, \eta) := (D - m(X))\big(Y - g(X) - \theta(D - m(X))\big), \tag{3}$$

where $\eta = (g, m)$ collects nuisance functions. At the reference point we take

$$\eta_0 := (g_0^\star, m_0), \qquad g_0^\star(X) := \ell_0(X) = \mathbb{E}[Y \mid X].$$

With this choice, the score satisfies

$$\Psi(\theta_0, \eta_0) := \mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0,$$

and Neyman orthogonality holds:

$$\partial_\eta \Psi(\theta_0, \eta)\big|_{\eta=\eta_0} = 0,$$

so that the moment condition is insensitive to first-order perturbations in η, in line with the doubly robust constructions of Bang and Robins (2005) and the locally robust GMM framework of Chernozhukov et al. (2022).[3]

Cross-fitted DML estimator. The DML estimator uses K-fold cross-fitting. Let $\hat{m}$ and $\hat{g}$ denote generic cross-fitted estimators of m_0 and $g_0^\star$; for each i, they are trained on folds not containing observation i. Define residualized variables

$$\hat{U}_i := D_i - \hat{m}(X_i), \qquad \hat{V}_i := Y_i - \hat{g}(X_i). \tag{4}$$

[3] In the ATE context, related orthogonal scores and uniform-in-model selection results are developed by Farrell (2015).

The empirical score average is

$$\Psi_n(\theta, \hat{\eta}) := \frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \theta, \hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{U}_i(\hat{V}_i - \theta \hat{U}_i),$$

and the DML estimator $\hat{\theta}$ is defined by the empirical moment equation

$$\Psi_n(\hat{\theta}, \hat{\eta}) = 0.$$

Solving gives the familiar partialling-out formula

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{U}_i \hat{V}_i}{\sum_{i=1}^{n} \hat{U}_i^2}, \tag{5}$$

which coincides with post-double-selection PLR estimators in Belloni et al. (2014) when $\hat{m}$ and $\hat{g}$ are obtained via sparse linear methods.

Jacobian and condition number. The empirical Jacobian is

$$\hat{J}_\theta := \partial_\theta \Psi_n(\theta, \hat{\eta}) = -\frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2, \tag{6}$$

which does not depend on θ and is nonpositive. We define the DML condition number

$$\kappa_{\mathrm{DML}} := -\frac{1}{\hat{J}_\theta} = \frac{1}{|\hat{J}_\theta|} = \frac{n}{\sum_{i=1}^{n} \hat{U}_i^2}, \tag{7}$$

which is finite whenever $\sum_{i=1}^{n} \hat{U}_i^2 > 0$. Small residual treatment variation (small $\sum \hat{U}_i^2$) implies a large κ_{DML}, corresponding to a nearly flat score and a numerically unstable estimator, exactly mirroring the weak-instrument problem in IV (Staiger and Stock, 1997; Stock and Yogo, 2005).[4]

3 Linearization and Coverage Error Bound

We now establish a refined linearization of the DML estimator and derive a coverage error bound for the standard DML confidence interval. Our arguments follow the orthogonal-score expansion principles underlying debiased / desparsified estimators (e.g., van de Geer et al., 2014; Javanmard and Montanari, 2014), but specialized to cross-fitted PLR DML.

3.1 Linearization

Define the empirical score average

$$\Psi_n(\theta, \eta) := \frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \theta, \eta).$$

[4] See also Belloni et al. (2012) for Lasso-based IV and the role of strong first stages in high-dimensional optimal instrument construction.

At (θ_0, η_0) and $(\theta_0, \hat{\eta})$ set

$$S_n := \Psi_n(\theta_0, \eta_0) = \frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \theta_0, \eta_0), \tag{8}$$

$$B_n := \Psi_n(\theta_0, \hat{\eta}) - \Psi_n(\theta_0, \eta_0), \tag{9}$$

and let R_n denote a Taylor remainder.

Assumption 1 (Regularity Conditions)**.** The following conditions hold.

(i) **(Score regularity)** For some $\tilde{\theta}$ between $\hat{\theta}$ and θ_0,

$$\partial_\theta \Psi_n(\tilde{\theta}, \hat{\eta}) = \hat{J}_\theta + o_P(1),$$

where $\hat{J}_\theta$ is defined in (6). In the PLR score, this equality holds exactly.

(ii) **(Invertibility)** There exists a deterministic sequence $c_{J,n} > 0$ and a constant $\delta_J \in (0, 1)$ such that

$$\mathbb{P}\big(|\hat{J}_\theta| \geq c_{J,n}\big) \geq 1 - \delta_J.$$

We write $\kappa_n := 1/c_{J,n}$ for the implied upper envelope of κ_{DML}.

(iii) **(Nuisance rate)** The nuisance estimators satisfy

$$\|\hat{m} - m_0\|_{L^2} \cdot \|\hat{g} - g_0^\star\|_{L^2} = o_P(n^{-1/2}),$$

as in standard DML conditions (Chernozhukov et al., 2018; Newey and Robins, 2018).

(iv) **(Moment bounds)** For the score at the truth,

$$\mathbb{E}\big[\psi(W; \theta_0, \eta_0)^2\big] =: \sigma_\psi^2 \in (0, \infty), \qquad \mathbb{E}\big[|\psi(W; \theta_0, \eta_0)|^3\big] \leq M_3 < \infty.$$

Lemma 1 (Refined Linearization of the DML Estimator)**.** *Under Assumption 1,*

$$\hat{\theta} - \theta_0 = \kappa_{\mathrm{DML}}\{S_n + B_n\} + R_n, \tag{10}$$

where κ_{DML} is defined in (7), S_n and B_n are given in (8)–(9), and R_n satisfies $R_n = o_P(n^{-1/2})$. Moreover,

$$S_n = O_P(n^{-1/2}), \qquad B_n = o_P(n^{-1/2}).$$

Proof. Since $\hat{\theta}$ solves $\Psi_n(\hat{\theta}, \hat{\eta}) = 0$, a first-order Taylor expansion around θ_0 yields

$$\Psi_n(\hat{\theta}, \hat{\eta}) = \Psi_n(\theta_0, \hat{\eta}) + \partial_\theta \Psi_n(\tilde{\theta}, \hat{\eta})(\hat{\theta} - \theta_0),$$

for some $\tilde{\theta}$ between $\hat{\theta}$ and θ_0. Rearranging,

$$\hat{J}_\theta(\hat{\theta} - \theta_0) = -\Psi_n(\theta_0, \hat{\eta}) - r_{n,\theta}(\hat{\theta} - \theta_0), \tag{11}$$

where $r_{n,\theta} = \partial_\theta \Psi_n(\tilde{\theta}, \hat{\eta}) - \hat{J}_\theta = o_P(1)$ by Assumption 1(i). On the event $\{|\hat{J}_\theta| \geq c_{J,n}\}$ we can divide by $\hat{J}_\theta$ and obtain

$$\hat{\theta} - \theta_0 = -\hat{J}_\theta^{-1} \Psi_n(\theta_0, \hat{\eta}) - \hat{J}_\theta^{-1} r_{n,\theta}(\hat{\theta} - \theta_0).$$

Define

$$R_n := -\hat{J}_\theta^{-1} r_{n,\theta}(\hat{\theta} - \theta_0),$$

and decompose

$$\Psi_n(\theta_0, \hat{\eta}) = \Psi_n(\theta_0, \eta_0) + \big(\Psi_n(\theta_0, \hat{\eta}) - \Psi_n(\theta_0, \eta_0)\big) = S_n + B_n.$$

Using $\kappa_{\mathrm{DML}} = -\hat{J}_\theta^{-1}$, we obtain (10). The orders $S_n = O_P(n^{-1/2})$ and $B_n = o_P(n^{-1/2})$ follow from the central limit theorem for the orthogonal score and the product-rate condition in Assumption 1(iii), as in Chernozhukov et al. (2018); Newey and Robins (2018). Under standard DML conditions, $\hat{\theta} - \theta_0 = O_P(n^{-1/2})$, so $R_n = O_P\big((\hat{\theta} - \theta_0)^2\big) = O_P(n^{-1}) = o_P(n^{-1/2})$. □

Remark 1 (Interpretation). The decomposition (10) separates three sources of error:

- $\kappa_{\mathrm{DML}} S_n$ is the sampling fluctuation component, of order $O_P(\kappa_{\mathrm{DML}}/\sqrt{n})$.
- $\kappa_{\mathrm{DML}} B_n$ is the nuisance-induced component; orthogonality ensures $B_n = o_P(n^{-1/2})$, but large κ_{DML} can magnify it.
- R_n is a higher-order remainder, negligible at $n^{-1/2}$ scale.

When $\kappa_{\mathrm{DML}} = O_P(1)$, both variance and bias are $O_P(n^{-1/2})$. When κ_{DML} grows, the effective convergence rate deteriorates and the confidence interval length grows proportionally to κ_{DML}, as in desparsified high-dimensional procedures (van de Geer et al., 2014; Javanmard and Montanari, 2014).

3.2 Coverage Error Bound

We now establish a coverage error bound for the standard DML confidence interval, in the spirit of the finite-sample Gaussian approximation results of Chernozhukov et al. (2023) and the high-dimensional bounds in Quintas-Martinez (2022); Jung (2023).

Let

$$\mathrm{CI}_{\mathrm{std}} := \Big[\hat{\theta} \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}_{\mathrm{DML}}\Big], \tag{12}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of $N(0,1)$ and

$$\widehat{\mathrm{SE}}_{\mathrm{DML}} := \frac{\kappa_{\mathrm{DML}}}{\sqrt{n}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2 \hat{\varepsilon}_i^2}, \tag{13}$$

with

$$\hat{\varepsilon}_i := Y_i - \hat{g}(X_i) - \hat{\theta}(D_i - \hat{m}(X_i)).$$

This is the usual plug-in estimator of the asymptotic standard deviation of $\hat{\theta}$. Let

$$s_n := \frac{\kappa_{\mathrm{DML}} \sigma_\psi}{\sqrt{n}}$$

denote the corresponding nonrandom target scale.

Assumption 2 (Concentration Bounds)**.** For some $\delta \in (0,1)$, there exist deterministic sequences $a_n(\delta)$ and $r_n(\delta)$ and a constant $c_\xi \in (0, 1/2)$ such that, with probability at least $1-\delta$,

(i) **(Sampling fluctuation)** $|S_n| \le a_n(\delta)$ with $a_n(\delta) = O(\sigma_\psi/\sqrt{n})$.

(ii) **(Nuisance and remainder)** $|B_n| + |R_n| \le r_n(\delta)$ with $r_n(\delta) = O(n^{-1/2-\gamma})$ for some $\gamma > 0$.

(iii) **(SE consistency)** $|\widehat{\mathrm{SE}}_{\mathrm{DML}} - s_n| \le c_\xi s_n$.

On this event,

$$(1-c_\xi)s_n \le \widehat{\mathrm{SE}}_{\mathrm{DML}} \le (1+c_\xi)s_n,$$

so $\widehat{\mathrm{SE}}_{\mathrm{DML}}$ is bounded away from zero and of order s_n.

Define the t-statistic

$$T_n := \frac{\hat\theta - \theta_0}{\widehat{\mathrm{SE}}_{\mathrm{DML}}}.$$

Then

$$\theta_0 \in \mathrm{CI}_{\mathrm{std}} \quad \Longleftrightarrow \quad |T_n| \le z_{1-\alpha/2}.$$

Theorem 2 (Coverage Error Bound)**.** *Under Assumptions 1 and 2, there exist constants $C_1, C_2, C_3, C_4 > 0$ such that*

$$\left|\mathbb{P}\big(\theta_0 \in \mathrm{CI}_{\mathrm{std}}\big) - (1-\alpha)\right| \le \frac{C_1}{\sqrt{n}} + C_2\sqrt{n}\, r_n(\delta) + C_3\delta + C_4 c_\xi, \tag{14}$$

where $r_n(\delta)$ is as in Assumption 2(ii). If, in addition, $c_\xi = O(n^{-1/2})$, the last term can be absorbed into the first, yielding

$$\left|\mathbb{P}\big(\theta_0 \in \mathrm{CI}_{\mathrm{std}}\big) - (1-\alpha)\right| \le \frac{\tilde C_1}{\sqrt{n}} + C_2\sqrt{n}\, r_n(\delta) + C_3\delta \tag{15}$$

for some $\tilde C_1 > 0$.

Proof sketch. Write

$$T_n = \frac{\kappa_{\mathrm{DML}}(S_n + B_n) + R_n}{\widehat{\mathrm{SE}}_{\mathrm{DML}}} = T_{n,0} + \Delta_n,$$

where

$$T_{n,0} := \frac{\kappa_{\mathrm{DML}} S_n}{\widehat{\mathrm{SE}}_{\mathrm{DML}}}, \qquad \Delta_n := \frac{\kappa_{\mathrm{DML}} B_n + R_n}{\widehat{\mathrm{SE}}_{\mathrm{DML}}}.$$

Define the "ideal" statistic

$$\tilde T_{n,0} := \frac{\kappa_{\mathrm{DML}} S_n}{s_n} = \frac{\sqrt{n}}{\sigma_\psi} S_n = \frac{1}{\sqrt{n}\,\sigma_\psi}\sum_{i=1}^n \psi(W_i; \theta_0, \eta_0),$$

which is a standardized average of mean-zero i.i.d. scores. By the classical Berry–Esseen theorem,

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\big(\tilde T_{n,0} \le z\big) - \Phi(z)\right| \le \frac{C_1}{\sqrt{n}}, \tag{16}$$

for some $C_1 > 0$ depending only on M_3 and σ_ψ.

On the concentration event in Assumption 2, the difference between $T_{n,0}$ and $\tilde{T}_{n,0}$ is controlled by the SE consistency:

$$\left|T_{n,0} - \tilde{T}_{n,0}\right| = \kappa_{\mathrm{DML}}|S_n| \left| \frac{1}{\widehat{\mathrm{SE}}_{\mathrm{DML}}} - \frac{1}{s_n} \right| \leq C_2' c_\xi,$$

for some $C_2' > 0$, using $a_n(\delta) = O(\sigma_\psi/\sqrt{n})$ and $s_n = \kappa_{\mathrm{DML}}\sigma_\psi/\sqrt{n}$. Moreover, the nuisance and remainder terms satisfy

$$|\Delta_n| \leq \frac{\kappa_{\mathrm{DML}}|B_n| + |R_n|}{(1 - c_\xi)s_n} \leq C_2'' \sqrt{n}\, r_n(\delta),$$

for some constant $C_2'' > 0$, since $s_n \asymp \kappa_{\mathrm{DML}}/\sqrt{n}$. Thus, on the concentration event, T_n differs from $\tilde{T}_{n,0}$ by at most $C_2\sqrt{n}\, r_n(\delta) + C_4 c_\xi$, and anti-concentration inequalities for the normal distribution imply that this shift perturbs $\mathbb{P}(|T_n| \leq z_{1-\alpha/2})$ by at most a constant multiple of the shift magnitude. Combining this with (16) and adding the probability of the complement event (at most δ) yields (14). The structure of the argument parallels the Gaussian approximation bounds in Chernozhukov et al. (2023); Quintas-Martinez (2022); Jung (2023), but is specialized to scalar PLR DML and made explicit in terms of κ_{DML}. □

Remark 2 (From t-scale to parameter scale). The bound (14) is expressed in t-statistic scale and does not display κ_{DML} explicitly, because the leading term normalizes by $s_n \propto \kappa_{\mathrm{DML}}/\sqrt{n}$. Combining Theorem 2 with Lemma 1, however, shows that

$$\hat{\theta} - \theta_0 = O_P\Big(\frac{\kappa_{\mathrm{DML}}}{\sqrt{n}} + \kappa_{\mathrm{DML}} r_n(\delta)\Big),$$

so poor conditioning directly inflates both variance and bias of the DML estimator and the length of $\mathrm{CI}_{\mathrm{std}}$. This is fully analogous to how weak instruments inflate IV variance and bias (Staiger and Stock, 1997; Stock and Yogo, 2005).

4 Conditioning Regimes

The combination of the linearization (10) and Assumption 2 induces a natural classification of DML inference into regimes based on the growth rate of $\kappa_n := \kappa_{\mathrm{DML}}$.

Corollary 3 (Conditioning Regimes). *Suppose Assumptions 1–2 hold and $r_n(\delta) = O(n^{-1/2-\gamma})$ for some $\gamma > 0$. Then:*

(i) ***Well-conditioned*** *($\kappa_n = O_P(1)$). The estimator satisfies*

$$\hat{\theta} - \theta_0 = O_P(n^{-1/2}),$$

and the confidence interval length is $O_P(n^{-1/2})$. Standard DML inference is reliable and informative.

(ii) ***Moderately ill-conditioned*** *($\kappa_n = O_P(n^\beta)$, $0 < \beta < 1/2$). The estimator satisfies*

$$\hat{\theta} - \theta_0 = O_P(n^{\beta-1/2}),$$

and the confidence interval length is $O_P(n^{\beta-1/2})$. Consistency is preserved, but convergence is slower and finite-sample coverage can be poor, echoing the behavior observed in high-dimensional desparsified estimators (van de Geer et al., 2014; Javanmard and Montanari, 2014).

(iii) ***Severely ill-conditioned*** *($\kappa_n \asymp c\sqrt{n}$). The estimator satisfies*

$$\hat{\theta} - \theta_0 = O_P(1),$$

and the confidence interval length is $O_P(1)$: the intervals fail to shrink as n grows, even if the t-statistic remains asymptotically normal.

Interpretation. The coverage bound in Theorem 2 depends on $\sqrt{n}\, r_n(\delta)$, but not explicitly on κ_n, because the latter cancels in t-scale. For the confidence sets to be informative, however, their length must vanish, which forces $\kappa_n = o_P(\sqrt{n})$ and $\kappa_n r_n(\delta) \to 0$. Large condition numbers thus undermine finite-sample precision and slow the rate at which DML intervals concentrate, just as weak instruments undermine IV inference (Staiger and Stock, 1997; Stock and Yogo, 2005). In this sense, our analysis is complementary to the nonasymptotic DML theorems of Chernozhukov et al. (2023); Quintas-Martinez (2022); Jung (2023), which control coverage in t-scale but are agnostic about parameter-scale conditioning.

5 Simulation Evidence

We present Monte Carlo simulations illustrating how κ_{DML} predicts finite-sample coverage failures in standard DML inference. Our simulation design complements and extends prior evidence in Chernozhukov et al. (2018); Farrell (2015); Quintas-Martinez (2022), while incorporating the systematic overlap calibration approach of Zimmert (2018) and Naghi (2021).

5.1 Data-Generating Process

The data-generating process follows the canonical PLR model (2). We specify the three components as follows.

Covariate distribution. We draw $X \in \mathbb{R}^{10}$ from a multivariate Gaussian distribution with AR(1)/Toeplitz covariance structure:

$$X \sim N(0, \Sigma(\rho)), \quad \Sigma_{jk} = \rho^{|j-k|}, \tag{17}$$

where $\rho \in \{0, 0.5, 0.9\}$ controls the correlation among covariates. This covariance specification is standard in PLR/DML simulations (Robinson, 1988; Chernozhukov et al., 2018; Bach et al., 2022).

Treatment equation. The treatment variable follows a linear specification:

$$D = X^\top \beta_D + U, \quad U \sim N(0, \sigma_U^2), \quad U \perp X, \tag{18}$$

where $\beta_D = (1, 0.8, 0.6, 0.4, 0.2, 0, \ldots, 0)^\top \in \mathbb{R}^{10}$ uses a decaying pattern on the first five covariates to avoid dominance by any single variable, following simulation practice in Chernozhukov et al. (2018) and Bach et al. (2022).

Outcome equation. The outcome follows the PLR model with a nonlinear nuisance function:

$$Y = D\theta_0 + g_0(X) + \varepsilon, \quad \varepsilon \sim N(0,1), \quad \varepsilon \perp (D, X), \tag{19}$$

where $\theta_0 = 1$ and $g_0(X) = \gamma^\top \sin(X)$ with $\gamma = (1, 0.5, 0.25, 0.125, 0.0625, 0, \ldots, 0)^\top$. This smooth nonlinear nuisance function is sufficiently complex to require machine learning estimation but remains computationally tractable, similar to semiparametric benchmark designs in Robinson (1988) and Chernozhukov et al. (2018).

5.2 Overlap Calibration

Rather than choosing arbitrary values for σ_U^2, we systematically calibrate the residual variance to achieve target values of $R^2(D|X)$, following the overlap-based design approach of Zimmert (2018) and Naghi (2021). The theoretical relationship is:

$$R^2(D|X) = \frac{\mathrm{Var}(X^\top \beta_D)}{\mathrm{Var}(D)} = \frac{\beta_D^\top \Sigma(\rho) \beta_D}{\beta_D^\top \Sigma(\rho) \beta_D + \sigma_U^2}. \tag{20}$$

We define three overlap regimes with interpretable targets:

- **High overlap**: $R^2(D|X) = 0.75$, implying substantial residual variation in D after conditioning on X and hence good identification.
- **Moderate overlap**: $R^2(D|X) = 0.90$, implying limited residual variation and potential identification concerns.
- **Low overlap**: $R^2(D|X) = 0.97$, implying that D is nearly deterministic given X, with minimal residual variation for identification.

For each combination of overlap level and correlation ρ, we solve (20) for σ_U^2 to achieve the target $R^2(D|X)$. This calibration ensures that conditioning severity is comparable across designs with different covariate correlations.

5.3 DML Implementation

We implement the cross-fitted DML estimator with $K = 5$ folds as described in Section 2. For nuisance estimation, we use random forest regressors with conservative hyperparameters (200 trees, maximum depth 5) to avoid overfitting while providing sufficient flexibility for the nonlinear nuisances. This choice balances bias and variance in finite samples and is similar in spirit to the implementations in Chernozhukov et al. (2018) and Bach et al. (2022).

In each Monte Carlo replication $b \in \{1, \ldots, B\}$, we:

(i) Generate $(Y_i, D_i, X_i)_{i=1}^n$ from the DGP (17)–(19).

(ii) Fit cross-fitted nuisance estimates $\hat{m}^{(-k)}(X_i)$ and $\hat{\ell}^{(-k)}(X_i)$ for $i \in I_k$, $k = 1, \ldots, K$.

(iii) Compute residualized treatments $\hat{U}_i = D_i - \hat{m}^{(-k)}(X_i)$ and residualized outcomes $\hat{V}_i = Y_i - \hat{\ell}^{(-k)}(X_i)$.

(iv) Compute $\hat{\theta}^{(b)} = \sum_i \hat{U}_i \hat{V}_i / \sum_i \hat{U}_i^2$, the standard error $\widehat{\mathrm{SE}}^{(b)}$ via (13), and the condition number $\kappa_{\mathrm{DML}}^{(b)} = n / \sum_i \hat{U}_i^2$.

(v) Construct the nominal 95% confidence interval $\hat{\theta}^{(b)} \pm 1.96 \cdot \widehat{\mathrm{SE}}^{(b)}$ and record coverage.

5.4 Simulation Configurations

We consider 9 distinct DGP configurations spanning the three conditioning regimes, varying sample size $n \in \{500, 2000\}$, covariate correlation $\rho \in \{0, 0.5, 0.9\}$, and overlap level. The configurations are designed to systematically explore the relationship between conditioning and finite-sample performance:

- **Group A** (well-conditioned): High overlap ($R^2 = 0.75$), expected small κ_{DML}.
- **Group B** (moderately ill-conditioned): Moderate overlap ($R^2 = 0.90$), expected moderate κ_{DML}.
- **Group C** (severely ill-conditioned): Low overlap ($R^2 = 0.97$), expected large κ_{DML}.

For each configuration, we run $B = 500$ Monte Carlo replications with independent random seeds for full reproducibility.

5.5 Results

Table 1 presents the main simulation results. The nine DGP configurations span a range of conditioning regimes, from well-conditioned (Group A) to severely ill-conditioned (Group C).

Table 1: Monte Carlo Results: Condition Number and Coverage by Design ($B = 500$ replications)

DGP	n	ρ	Overlap	$R^2(D\|X)$	Mean κ_{DML}	Coverage (%)	RMSE
Group A: Well-conditioned (high overlap)							
A1	500	0.0	High	0.753	0.89	92.6	0.052
A2	500	0.5	High	0.752	0.54	93.2	0.037
A3	2000	0.5	High	0.750	0.57	94.0	0.018
Group B: Moderately ill-conditioned (moderate overlap)							
B1	500	0.0	Moderate	0.900	1.61	93.6	0.065
B2	500	0.5	Moderate	0.900	1.20	93.4	0.056
B3	2000	0.5	Moderate	0.900	1.27	85.2	0.037
Group C: Severely ill-conditioned (low overlap)							
C1	500	0.5	Low	0.969	2.09	92.4	0.083
C2	2000	0.5	Low	0.970	2.26	62.0	0.072
C3	2000	0.9	Low	0.970	2.70	**59.2**	0.080

Notes: Coverage is the percentage of nominal 95% confidence intervals containing $\theta_0 = 1$. Overlap is calibrated via σ_U^2 to achieve the target $R^2(D|X)$ for each ρ. RMSE is computed over 500 replications. Nuisance functions are estimated via random forests with 5-fold cross-fitting. Patterns are consistent with finite-sample DML distortions documented in Chernozhukov et al. (2018); Quintas-Martinez (2022); Jung (2023).

Key findings. The results exhibit a clear stratification by the conditioning regime:

(i) **Well-conditioned designs** (Group A, $\kappa_{\text{DML}} \approx 0.5$–$0.9$): Empirical coverage ranges from 92.6% to 94.0%, close to the nominal 95% level. The configuration A3 ($n = 2000$, $\rho = 0.5$, high overlap) achieves exactly 94.0% coverage with the smallest condition number ($\kappa_{\text{DML}} = 0.57$) and lowest RMSE (0.018), demonstrating reliable inference when conditioning is favorable.

(ii) **Moderately ill-conditioned designs** (Group B, $\kappa_{\text{DML}} \approx 1.2$–$1.6$): Coverage begins to degrade, particularly at larger sample sizes. While B1 and B2 ($n = 500$) maintain coverage above 93%, B3 ($n = 2000$) drops to 85.2%. This pattern suggests that nuisance estimation bias, amplified by moderate κ_{DML}, becomes more pronounced as the sample size increases and the standard error shrinks.

(iii) **Severely ill-conditioned designs** (Group C, $\kappa_{\text{DML}} \approx 2.1$–$2.7$): Coverage collapses dramatically. At $n = 500$ (C1), coverage remains at 92.4%, but at $n = 2000$ with moderate correlation (C2), coverage falls to 62.0%. The worst case, C3 ($n = 2000$, $\rho = 0.9$, low overlap), exhibits only 59.2% coverage of nominal 95% intervals despite the large sample size.

The paradox of larger samples. A striking feature of Table 1 is that increasing sample size from $n = 500$ to $n = 2000$ can *worsen* coverage in ill-conditioned designs. Comparing C1 to C2 (both with $\rho = 0.5$, low overlap), coverage drops from 92.4% to 62.0% as n quadruples. This counterintuitive pattern arises because larger samples yield smaller standard errors, which shrink the confidence intervals, but the bias term—amplified by κ_{DML}—remains of comparable magnitude. The theoretical expansion (10) predicts exactly this behavior: the parameter-scale error $\hat{\theta} - \theta_0 = O_P(\kappa_{\text{DML}}/\sqrt{n} + \kappa_{\text{DML}} r_n)$ has a bias component $\kappa_{\text{DML}} r_n$ that may dominate the variance term as n grows.

Diagnostic value of κ_{DML}. Across all configurations, the condition number strongly predicts coverage performance. The empirical correlation between mean κ_{DML} and coverage is $r = -0.77$, and the correlation between mean κ_{DML} and RMSE is $r = 0.88$. The best-performing configuration (A3: $\kappa_{\text{DML}} = 0.57$, coverage = 94.0%) and the worst-performing configuration (C3: $\kappa_{\text{DML}} = 2.70$, coverage = 59.2%) span a five-fold difference in condition number, corresponding to a coverage gap of nearly 35 percentage points. These patterns validate κ_{DML} as a practical diagnostic for assessing the reliability of DML inference.

6 Discussion

This short communication makes three contributions to the understanding of finite-sample behavior in Double Machine Learning.

1. Coverage error bound. We establish a coverage error bound (Theorem 2) showing that the DML t-statistic enjoys a Berry–Esseen-type normal approximation with error of order

$$\frac{C_1}{\sqrt{n}} + C_2\sqrt{n}\, r_n(\delta) + C_3\delta,$$

where $r_n(\delta)$ captures the nuisance estimation error. This bound is consistent with recent finite-sample DML theory (Chernozhukov et al., 2023; Quintas-Martinez, 2022; Jung, 2023) and makes explicit the interplay between sample size, nuisance accuracy, and convergence rate.

2. κ-amplified linearization. We derive a refined linearization (10) revealing that the *parameter-scale* error satisfies

$$\hat{\theta} - \theta_0 = O_P\left(\frac{\kappa_{\mathrm{DML}}}{\sqrt{n}} + \kappa_{\mathrm{DML}}\, r_n(\delta)\right).$$

Poor conditioning directly amplifies both the variance component (first term) and the bias component (second term). This explains why ill-conditioned designs can exhibit coverage failures even at seemingly large sample sizes: the condition number magnifies nuisance estimation errors, and these errors dominate the shrinking standard error as n grows.

3. Empirical validation. Our Monte Carlo simulations confirm that κ_{DML} is an informative diagnostic. Designs with $\kappa_{\mathrm{DML}} < 1$ achieve near-nominal coverage (92–94% of nominal 95%), while designs with $\kappa_{\mathrm{DML}} > 2$ can exhibit severe undercoverage (59–62%). The empirical correlation between κ_{DML} and coverage ($r = -0.77$) and between κ_{DML} and RMSE ($r = 0.88$) validate the condition number as a practical tool for assessing inference reliability.

6.1 Practical Recommendations

We recommend that practitioners compute and report κ_{DML} alongside DML estimates, analogous to first-stage F-statistics in instrumental variables regression (Staiger and Stock, 1997; Stock and Yogo, 2005). Based on our theoretical analysis and simulation evidence, we propose the following guidelines:

- $\kappa_{\mathrm{DML}} < 1$ **(well-conditioned):** Standard DML inference is typically reliable. Confidence intervals shrink at the usual $n^{-1/2}$ rate, and coverage should be close to nominal. No special adjustments are needed.
- $1 \leq \kappa_{\mathrm{DML}} < 2$ **(moderately ill-conditioned):** Exercise caution. The effective convergence rate may be slower than $n^{-1/2}$, and coverage degradation is possible at larger sample sizes. Consider robustness checks with alternative ML learners, different cross-fitting schemes, or specifications that improve overlap.
- $\kappa_{\mathrm{DML}} \geq 2$ **(severely ill-conditioned):** Standard confidence intervals may be substantially distorted. Inference should be interpreted as diagnostic rather than definitive. Consider reporting confidence intervals with explicit caveats, investigating sources of poor overlap, or employing alternative identification strategies if available.

6.2 Limitations and Extensions

Our analysis has several limitations that suggest directions for future research.

Scope. We focus on the PLR model with a scalar target parameter θ_0 and cross-fitted random forest nuisance estimators. Extending the condition-number diagnostic to IV-DML (Chernozhukov et al., 2018), panel data settings, and multivariate target parameters is conceptually straightforward but requires verification of the concentration assumptions in more complex settings.

Bias-aware inference. Our results characterize when standard DML intervals fail but do not provide a remedy. Developing fully robust, conditioning-aware confidence sets—such as κ-inflated intervals or bias-corrected procedures—remains an important open problem. Related work on honest inference (Wager and Athey, 2018) and bias-aware methods (Armstrong and Kolesár, 2020) may provide useful starting points.

Learner dependence. The condition number κ_{DML} depends on the nuisance estimator through the residualized treatments $\hat{U}_i = D_i - \hat{m}(X_i)$. Different ML learners (e.g., LASSO, neural networks, boosting) may yield different κ_{DML} values for the same DGP. Investigating this learner dependence and its implications for practitioner guidance is a natural extension.

Conclusion. We view this note as a finite-sample cautionary message: asymptotic DML theory remains valid, but its practical reliability hinges on the conditioning summarized by κ_{DML}. Just as weak instruments degrade IV inference regardless of sample size, large condition numbers degrade DML inference even in settings where classical asymptotics appear to apply. The condition number provides a simple, interpretable diagnostic that practitioners can compute routinely to assess the reliability of their DML estimates, contributing to the broader literature on locally robust and debiased inference (van de Geer et al., 2014; Javanmard and Montanari, 2014; Chernozhukov et al., 2022, 2023).

References

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

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

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.

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., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

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.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909.

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

Newey, W. K. and Robins, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. CeMMAP Working Paper CWP41/17.

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

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

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

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear IV regression. In Andrews, D. W. K. and Stock, J. H., editors, *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pages 80–108. Cambridge University Press.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Armstrong, T. B. and Kolesár, M. (2020). Bias-aware inference in regularized regression models. *arXiv preprint arXiv:1802.08667v4*.

Zimmert, M. (2018). The finite sample performance of treatment effect estimators in high-dimensional settings. *arXiv preprint arXiv:1805.05067*.

Naghi, A. (2021). Finite sample evaluation of causal machine learning methods. *Tinbergen Institute Discussion Paper* 2021-090.

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