Manuscript submitted to *The Econometrics Journal*, pp. 1–24.

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

Gabriel Saco[†]

[†] *(Affiliation to be added)*
E-mail: gsacoalvarado@gmail.com

Summary Double Machine Learning delivers root-n inference by pairing flexible nuisance estimation with Neyman-orthogonal scores, but finite-sample reliability depends on the conditioning of the score—a property standard diagnostics ignore. We introduce the DML condition number κ_{DML}, the inverse of the empirical score Jacobian. A refined linearisation shows that κ_{DML} multiplies both sampling variance and nuisance-induced bias, yielding a parameter-scale rate $O_P(\kappa_{\mathrm{DML}}/\sqrt{n} + \kappa_{\mathrm{DML}} r_n)$ and an explicit coverage bound. We classify designs into three regimes—well-conditioned, moderately ill-conditioned, and severely ill-conditioned—paralleling weak-IV classifications. Monte Carlo experiments map overlap and learner complexity into κ_{DML} and thence into coverage. Re-analysis of LaLonde (1986) shows that κ_{DML} immediately distinguishes robust experimental estimates from fragile observational re-analyses. We argue that κ_{DML} should accompany every DML estimate, just as first-stage F-statistics accompany IV results.

Keywords: *Double Machine Learning, Nonasymptotic Inference, Weak Identification, Partially Linear Regression.*

1. INTRODUCTION

Double Machine Learning (Chernozhukov et al., 2018) has become the default method for estimating treatment effects and structural parameters when researchers have access to rich covariate information. The theoretical appeal is substantial: Neyman-orthogonal scores combined with cross-fitting deliver $\sqrt{n}$-consistent, asymptotically normal estimators even when nuisance functions are estimated nonparametrically, provided those estimators satisfy mild rate conditions. Applied papers now routinely report DML point estimates and Wald confidence intervals as if they were robust, model-free solutions to the high-dimensional confounding problem. Yet this confidence rests on asymptotic guarantees that may not translate to finite samples. Practitioners observe a normal-looking t-statistic and a seemingly reasonable confidence interval, but they have no diagnostic that reveals whether the underlying orthogonal score is well-conditioned or dangerously flat. This paper provides such a diagnostic.

The core problem is that DML inference can fail silently. When the orthogonal score's Jacobian is small—equivalently, when residual treatment variation after partialling out covariates is limited—the estimating equation becomes nearly flat. Small perturbations in the score then translate into large perturbations in the parameter estimate. Existing finite-sample theory (Chernozhukov et al., 2023; Quintas-Martinez, 2022; Jung, 2023) provides Berry–Esseen bounds for the t-statistic under regularity conditions that implicitly assume the Jacobian is bounded away from zero. These results characterise well-identified settings but offer no guidance when identification weakens. The gap we fill is fundamental: there is no score-based, one-dimensional diagnostic for DML that plays the role of the first-stage F-statistic in instrumental variables regression (Staiger and Stock, 1997; Stock and Yogo, 2005).

The parallel to weak instruments is instructive. Before Staiger and Stock (1997) formalised the weak-IV problem and Stock and Yogo (2005) provided actionable critical values, practitioners routinely reported two-stage least squares estimates without knowing whether their first stage was too weak to support reliable inference. The F-statistic transformed practice: it gave researchers a simple number that summarised identification strength and indicated when standard Wald intervals might fail. In the DML literature, an analogous culture has not developed. Researchers sometimes report propensity score diagnostics—overlap plots, $R^2(D \mid X)$, trimming statistics—but these operate at the level of nuisance estimation, not at the level of the orthogonal score that drives inference.[1] We lack a scalar summary that directly measures how sensitive the DML estimating equation is to perturbations.

This paper introduces the *DML condition number*, denoted κ_{DML}, as precisely this diagnostic. We work in the partially linear regression (PLR) model and treat the DML estimator as a scalar Z-estimator (Newey and McFadden, 1994; van der Vaart, 1998). The empirical Jacobian of the orthogonal score is $\hat{J}_\theta = -n^{-1}\sum_i \hat{U}_i^2$, where $\hat{U}_i = D_i - \hat{m}(X_i)$ is the cross-fitted treatment residual. The condition number is simply the absolute reciprocal:

$$\kappa_{\mathrm{DML}} := \frac{1}{|\hat{J}_\theta|} = \frac{n}{\sum_{i=1}^n \hat{U}_i^2}. \tag{1.1}$$

When residual treatment variation is large, κ_{DML} is small and the score is steep—identification is strong. When residual treatment variation is small, κ_{DML} is large and the score is flat—small score perturbations yield large parameter perturbations. This is the exact finite-sample analogue of weak instruments, where a small first-stage coefficient inflates the variance and bias of the IV estimator. Crucially, κ_{DML} is not an ad hoc index: it is the actual inverse slope of the DML estimating equation, grounded in classical Z-estimation theory.

Our theoretical contribution is a parameter-scale analysis that places κ_{DML} at the centre. We derive a refined linearisation:

$$\hat{\theta} - \theta_0 = \kappa_{\mathrm{DML}}(S_n + B_n) + R_n, \tag{1.2}$$

where $S_n = n^{-1}\sum_i \psi(W_i; \theta_0, \eta_0)$ is the centred oracle score, B_n captures nuisance-induced bias from estimating η_0, and R_n is a higher-order remainder. Under standard DML regularity conditions (Chernozhukov et al., 2018, 2022), $S_n = O_P(n^{-1/2})$ and $B_n = o_P(n^{-1/2})$, yielding the parameter-scale rate

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

where r_n is the usual DML nuisance remainder. The condition number multiplies *both* the sampling fluctuation and the nuisance bias. In t-statistic scale, κ_{DML} cancels with the standard error, so existing Berry–Esseen bounds remain valid. But in parameter scale, κ_{DML} governs the width of confidence intervals and the magnitude of any residual bias. This decomposition induces a natural classification of conditioning regimes. When $\kappa_{\mathrm{DML}} = O_P(1)$, the design is *well-conditioned*: confidence sets shrink at rate $n^{-1/2}$ and standard DML inference applies. When $\kappa_{\mathrm{DML}} = O_P(n^\beta)$ for $0 < \beta < 1/2$, the design is *moderately ill-conditioned*: sets shrink more slowly and bias is amplified. When

[1] The distinction matters because the score depends on both nuisance functions and their interaction with the data. A propensity model with excellent fit can still yield a flat score if residual treatment variance is small.

$\kappa_{\mathrm{DML}} \asymp c\sqrt{n}$, the design is *severely ill-conditioned*: parameter-scale error is $O_P(1)$ and confidence sets fail to shrink even as n grows. This last regime is the DML analogue of weak IV (Staiger and Stock, 1997; Stock and Wright, 2000).[2]

We validate these theoretical predictions through Monte Carlo experiments and an empirical application. The simulations use a PLR design with three overlap levels controlled via $R^2(D \mid X) \in \{0.75, 0.90, 0.97\}$ and three nuisance learners (OLS, Lasso, random forests). As overlap deteriorates, the distribution of κ_{DML} shifts sharply rightward: median κ_{DML} increases from approximately 0.7 to 5, with substantial dispersion at high R^2. Aggregating by conditioning regime reveals two distinct failure modes. For unbiased learners like OLS, large κ_{DML} produces wide but honest intervals—variance inflation without coverage distortion. For flexible learners like random forests, large κ_{DML} amplifies residual nuisance error into first-order bias, producing severe undercoverage despite smaller interval widths. These patterns directly implement the theoretical rate $\kappa_{\mathrm{DML}}/\sqrt{n} + \kappa_{\mathrm{DML}} r_n$.

The empirical application re-analyses the canonical LaLonde (1986) job-training data. In the experimental sample, where treatment is randomised and nearly independent of covariates, $\kappa_{\mathrm{DML}} \approx 4$ and DML estimates are stable across learners. In observational re-analyses using nonexperimental comparison groups, treatment becomes highly predictable, κ_{DML} exceeds 15, and estimates exhibit the fragility predicted by theory: extreme sensitivity to learner choice, very wide intervals, and sign reversals. The condition number immediately distinguishes robust experimental inference from unreliable observational re-analysis.

We conclude with a practical recommendation. Applied researchers should routinely compute and report κ_{DML} alongside DML point estimates and Wald confidence intervals, in direct analogy to first-stage F-statistics in IV (Andrews et al., 2019). A large condition number does not automatically invalidate results—just as a low F-statistic does not prove that IV fails—but it signals fragile conditioning that warrants scrutiny. When κ_{DML} is large, practitioners should interpret confidence intervals cautiously, investigate whether overlap can be improved through trimming or alternative estimands (Crump et al., 2009; Ma et al., 2023), and compare simple versus flexible learners to diagnose the $\kappa_{\mathrm{DML}} r_n$ channel. Our contribution is explicitly diagnostic: we clarify what goes wrong when the DML score is ill-conditioned, connect DML practice to the weak-identification culture that has become standard in IV, and provide a simple, implementable gauge of conditioning grounded in Z-estimation theory. We do not propose a new robust inference procedure; we propose a new standard for transparency.[3]

2. RELATED LITERATURE

The DML framework of Chernozhukov et al. (2018) combines Neyman-orthogonal scores with cross-fitting to deliver $\sqrt{n}$-consistent inference for low-dimensional parameters even when nuisance functions are estimated nonparametrically. The key insight is that orthogonal scores are locally insensitive to first-order perturbations in nuisance estimates, so

[2]The semiparametric efficiency bound for θ_0 in PLR is $\mathbb{E}[U^2\varepsilon^2]/(\mathbb{E}[U^2])^2$, which diverges as $\mathbb{E}[U^2] \to 0$. The condition number $\kappa_{\mathrm{DML}} \approx 1/\mathbb{E}[U^2]$ is thus a sample analogue of this divergence, connecting our finite-sample diagnostic to the population efficiency theory of Hahn (1998) and Hirano et al. (2003).

[3]To facilitate adoption, we provide a Python package `dml_diagnostic` that implements the condition number diagnostic and the analyses in this paper. The package, together with full simulation and empirical replication code, is available at `https://github.com/gsaco/dml-diagnostic`. Documentation is provided in the Online Supplement.

regularisation bias from machine learning does not contaminate the target parameter (Chernozhukov et al., 2022; Foster and Syrgkanis, 2023). This robustness property has enabled a generation of applied work that pairs flexible learners with rigorous inference, and has spurred extensions to conditional average treatment effects (Semenova and Chernozhukov, 2020), local projections (Chernozhukov et al., 2022b), and doubly robust difference-in-differences (Sant'Anna and Zhao, 2020). Recent contributions in *The Econometrics Journal* have further developed DML methodology: Knaus (2022) provides programme evaluation under unconfoundedness, while Baiardi and Naghi (2024) revisits canonical studies to assess the value added of machine learning to causal inference.[4] Subsequent finite-sample theory (Chernozhukov et al., 2023; Quintas-Martinez, 2022; Jung, 2023) provides Berry–Esseen bounds for the DML t-statistic, but these bounds are stated under regularity conditions that implicitly require the score Jacobian to be bounded away from zero. Our contribution is to make this Jacobian the central diagnostic object: we derive how its reciprocal κ_{DML} governs parameter-scale error, and we define conditioning regimes that classify designs by identification strength.

The semiparametric efficiency theory of Hahn (1998) and Hirano et al. (2003) establishes that the variance bound for treatment-effect parameters diverges as overlap deteriorates. In the PLR model, this bound scales as $\mathbb{E}[U^2\varepsilon^2]/(\mathbb{E}[U^2])^2$, where $U = D - m_0(X)$ is residual treatment variation; as $R^2(D \mid X) \to 1$, the denominator vanishes and the bound explodes. Xu et al. (2020) study inference under limited overlap in *The Econometrics Journal*, showing that when propensity scores approach zero or one, the "effective" sample size shrinks and estimators converge more slowly—a result that our condition number κ_{DML} captures directly at the score level. The targeted learning literature (Kennedy, 2023) emphasises that doubly robust estimators can achieve this bound but become unstable when propensity scores are extreme. From a different angle, Khan and Tamer (2010) and Kaji (2021) formalise weak semiparametric identification, showing that estimators become irregular when identifying information concentrates on thin sets—precisely the setting where treatment is nearly deterministic given covariates. Our condition number $\kappa_{\mathrm{DML}} \approx 1/\mathbb{E}[U^2]$ is a sample analogue of the efficiency bound's inflation factor. By tracking how κ_{DML} varies across designs and learners, we translate a population-level concern into a computable, finite-sample diagnostic that practitioners can report.

The weak-IV literature provides both our methodological template and our conceptual vocabulary. Staiger and Stock (1997) demonstrated that weak instruments yield two-stage least squares error of $O_P(1)$ and arbitrarily distorted Wald intervals, even as $n \to \infty$. Stock and Yogo (2005) translated this insight into actionable practice: first-stage F-statistics below 10 signal unreliable inference, and practitioners now routinely report these diagnostics. Stock and Wright (2000) extended the analysis to GMM, classifying designs by identification strength into strong, semi-strong, and weak regimes. Dufour (2003) provided foundational results on identification failure, while Moreira (2003) developed the conditional likelihood ratio test that remains valid under weak identification. Recent work on weak identification with many instruments (Mikusheva and Sun, 2024) further develops the theoretical foundations relevant to high-dimensional settings. We draw explicit parallels: the Jacobian $\hat{J}_\theta$ in DML plays the role of the first-stage coefficient in IV; κ_{DML} plays the role of the concentration parameter; our three regimes mirror

[4]Baiardi and Naghi (2024) systematically re-analyse published studies using DML and related methods, providing empirical guidance on when machine learning adds value. Our condition number κ_{DML} offers a theoretical foundation for such comparisons.

the strong–semi-strong–weak taxonomy. We do not develop conditioning-robust tests analogous to the Anderson–Rubin or Moreira CLR—that is an important direction for future work—but we demonstrate that κ_{DML} reliably indicates when standard DML intervals become fragile.[5]

Finite-sample studies of doubly robust and ML-based estimators have long documented sensitivity to overlap and nuisance misspecification (Kang and Schafer, 2007; Zimmert, 2018). Recent constructive proposals address weak overlap directly: trimming extreme propensity scores (Crump et al., 2009), doubly robust estimators tailored to weak overlap (Ma et al., 2023), propensity score calibration (Wüthrich and Zhu, 2024; Ballinari and Bearth, 2024), and sample mixing (Jang et al., 2024). Liu et al. (2024) study regularised DML in partially linear models with potential unobserved confounding, providing finite-sample results complementary to our analysis. Our diagnostic complements these methods: κ_{DML} indicates when such corrections may be needed and, after they are applied, confirms whether conditioning has improved. The gap we fill is simple but consequential: existing DML theory assumes good conditioning, and practitioners have no standard way to check whether their design satisfies this assumption. We provide a one-number summary that can be reported alongside point estimates and confidence intervals, bringing to DML the transparency that F-statistics brought to IV.

3. SETUP: PLR MODEL, ORTHOGONAL SCORE, AND CONDITION NUMBER

We consider the canonical partially linear regression (PLR) model. Observations $W_i = (Y_i, D_i, X_i)$, $i = 1, \ldots, n$, 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

$$Y = D\theta_0 + g_0(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid D, X] = 0, \tag{3.3}$$

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 (Robinson, 1988; Belloni et al., 2014).

Define the nuisance regression functions

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

Using (3.3),

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

The DML estimator for θ_0 uses a Neyman-orthogonal score. For PLR, this score is

$$\psi(W; \theta, \eta) := (D - m(X))\big(Y - g(X) - \theta(D - m(X))\big), \tag{3.4}$$

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 moment condition $\Psi(\theta, \eta) := \mathbb{E}[\psi(W; \theta, \eta)]$ satisfies $\Psi(\theta_0, \eta_0) = 0$ and

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

so the score is locally insensitive to first-order perturbations in η.

[5]Developing critical values for κ_{DML} analogous to Stock–Yogo critical values for first-stage F-statistics is an important direction for future work. Such critical values would depend on the desired worst-case bias or coverage distortion, learner complexity, and sample size.

The DML estimator uses K-fold cross-fitting. Let $\hat{m}$ and $\hat{g}$ denote cross-fitted estimators of m_0 and ℓ_0. For each i, they are trained on folds not containing observation i. Define residualized variables

$$\hat{U}_i := D_i - \hat{m}(X_i), \qquad \hat{V}_i := Y_i - \hat{g}(X_i). \tag{3.5}$$

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 $\hat{\theta}$ solves $\Psi_n(\hat{\theta}, \hat{\eta}) = 0$. This yields the familiar partialling-out formula

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \hat{U}_i\hat{V}_i}{\sum_{i=1}^{n} \hat{U}_i^2}. \tag{3.6}$$

The key object for our analysis is the empirical Jacobian

$$\hat{J}_\theta := \partial_\theta \Psi_n(\theta, \hat{\eta}) = -\frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2, \tag{3.7}$$

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

$$\kappa_{\mathrm{DML}} := -\frac{1}{\hat{J}_\theta} = \frac{1}{|\hat{J}_\theta|} = \frac{n}{\sum_{i=1}^{n} \hat{U}_i^2}, \tag{3.8}$$

which is finite whenever $\sum_{i=1}^{n} \hat{U}_i^2 > 0$. The terminology "condition number" is borrowed from numerical analysis: a small Jacobian implies parameter estimates are hypersensitive to score perturbations. Small residual treatment variation implies a large κ_{DML}, corresponding to a nearly flat score—the DML analogue of weak instruments (Staiger and Stock, 1997; Stock and Yogo, 2005).

The next section shows that κ_{DML} is not just a numerical curiosity but the primitive object that governs parameter-scale error and finite-sample coverage.

4. FINITE-SAMPLE THEORY

This section establishes the theoretical core of the paper. We derive three main results: (i) a refined linearisation that isolates κ_{DML} as the primitive conditioning object; (ii) a parameter-scale error rate that shows how κ_{DML} multiplies both variance and nuisance bias; and (iii) a coverage bound that quantifies finite-sample distortions. Together, these results provide the foundation for the conditioning regimes introduced in Section 5.

4.1. Assumptions

We impose three groups of conditions. The first specifies the model and moment requirements; the second makes explicit the role of residual treatment variance (overlap); the third governs nuisance estimation rates.

Assumption 4.1. (Model and moments) *(i) The PLR model* (3.3) *holds with* $\mathbb{E}[\varepsilon \mid D, X] = 0$*, and the score is given by* (3.4)*.*

(ii) The score ψ *is Neyman-orthogonal:* $\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0$ *(Chernozhukov et al., 2022).*

(iii) The score at the truth satisfies $0 < \sigma_\psi^2 := \mathrm{Var}(\psi(W;\theta_0,\eta_0)) < \infty$ *and* $\mathbb{E}[|\psi(W;\theta_0,\eta_0)|^3] \le M_3 < \infty$.

ASSUMPTION 4.2. (OVERLAP) *Define the population residual* $U := D - m_0(X)$ *and let* $\sigma_U^2 := \mathbb{E}[U^2] = \mathrm{Var}(D)(1 - R^2(D \mid X))$. *We require:*[6]

$$0 < \underline{\sigma}_U^2 \;\le\; \sigma_U^2 \;\le\; \bar{\sigma}_U^2 < \infty.$$

This condition is the DML analogue of requiring a non-zero first stage in IV: it ensures $R^2(D \mid X) < 1$ *and bounds the semiparametric variance Hirano et al. (2003).*

ASSUMPTION 4.3. (NUISANCE RATES) *Let* $\hat{m}$, $\hat{\ell}$ *be cross-fitted estimators of* m_0, ℓ_0 *using* K*-fold sample splitting (Newey and Robins, 2017; Chernozhukov et al., 2018). Define the product error*

$$r_n := \|\hat{m} - m_0\|_{L^2(P)} \cdot \|\hat{\ell} - \ell_0\|_{L^2(P)}.$$

We require:

(i) Product rate. $r_n = o_P(n^{-1/2})$.

(ii) Jacobian concentration. *With probability at least* $1 - \delta$,

$$\left|n^{-1}\textstyle\sum_{i=1}^n \hat{U}_i^2 - \sigma_U^2\right| \le c_1 n^{-1/2}\log(1/\delta).$$

(iii) Cross-fit regularity. *The cross-fitted residuals* $\hat{U}_i = D_i - \hat{m}(X_i)$ *satisfy* $\mathbb{E}[\hat{U}_i^2 \mid \mathcal{D}_{-k}] \xrightarrow{p} \sigma_U^2$ *uniformly over folds.*

The product-rate condition ensures nuisance-induced bias is $o_P(n^{-1/2})$*; for Lasso rates see Belloni et al. (2014) and van de Geer et al. (2014); for random forests see Wager and Athey (2018). Cross-fitting is essential for satisfying this condition with complex learners (Newey and Robins, 2017).*

4.2. Linearisation

The following lemma is the first main result. It provides a refined Z-estimation expansion that isolates κ_{DML} as the conditioning object.

Let

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

and define

$$S_n := \Psi_n(\theta_0,\eta_0) = \frac{1}{n}\sum_{i=1}^n \psi(W_i;\theta_0,\eta_0), \tag{4.9}$$

$$B_n := \Psi_n(\theta_0,\hat{\eta}) - \Psi_n(\theta_0,\eta_0), \tag{4.10}$$

with R_n collecting higher-order terms.

[6]The quantity σ_U^2 measures the "unexplained" variation in treatment after conditioning on X. In the propensity score literature this corresponds to requiring bounded propensity scores away from 0 and 1; see Rosenbaum and Rubin (1983) for the original formulation and Crump et al. (2009) for overlap diagnostics.

LEMMA 4.1. (REFINED LINEARISATION) *Under Assumptions 4.1–4.3,*

$$\hat{\theta} - \theta_0 = \kappa_{\mathrm{DML}}\{S_n + B_n\} + R_n, \tag{4.11}$$

where κ_{DML} is defined in (3.8)*, S_n and B_n are given by* (4.9)–(4.10)*, and $R_n = o_P(n^{-1/2})$. Moreover, $S_n = O_P(n^{-1/2})$ and $B_n = o_P(n^{-1/2})$.*

PROOF. The DML estimator $\hat{\theta}$ solves $\Psi_n(\hat{\theta}, \hat{\eta}) = 0$. Using the PLR score (3.4), this becomes

$$\frac{1}{n}\sum_{i=1}^{n} \hat{U}_i(\hat{V}_i - \hat{\theta}\hat{U}_i) = 0,$$

which yields the closed-form solution (3.6). To derive the linearisation, write

$$\hat{\theta} - \theta_0 = \frac{\sum_i \hat{U}_i\hat{V}_i}{\sum_i \hat{U}_i^2} - \theta_0 = \frac{\sum_i \hat{U}_i(\hat{V}_i - \theta_0\hat{U}_i)}{\sum_i \hat{U}_i^2}. \tag{4.12}$$

Define population residuals $U_i := D_i - m_0(X_i)$ and $V_i := Y_i - \ell_0(X_i)$. By the model (3.3), $V_i = \theta_0 U_i + \varepsilon_i$ where $\mathbb{E}[\varepsilon_i \mid X_i, D_i] = 0$. The numerator in (4.12) decomposes as

$$\sum_i \hat{U}_i(\hat{V}_i - \theta_0\hat{U}_i) = \underbrace{\sum_i U_i\varepsilon_i}_{=:nS_n} + \underbrace{\sum_i \hat{U}_i(\hat{V}_i - V_i) - \theta_0\sum_i \hat{U}_i(\hat{U}_i - U_i)}_{=:nB_n} + R_n', \tag{4.13}$$

where R_n' collects higher-order cross-terms.

Analysis of S_n: By construction, $S_n = n^{-1}\sum_i U_i\varepsilon_i$ is a sample average of mean-zero random variables with variance σ_ψ^2/n. Under Assumption 4.1(iii), the CLT gives $\sqrt{n}S_n \xrightarrow{d} N(0, \sigma_\psi^2)$, so $S_n = O_P(n^{-1/2})$.

Analysis of B_n: The bias term B_n arises from nuisance estimation error. Expanding using $\hat{V}_i - V_i = (\ell_0(X_i) - \hat{\ell}(X_i))$ and $\hat{U}_i - U_i = (\hat{m}(X_i) - m_0(X_i))$, and applying orthogonality (Assumption 4.1(ii)) together with the product-rate condition (Assumption 4.3(i)), we obtain $B_n = O_P(r_n) = o_P(n^{-1/2})$.

Denominator: By Assumption 4.3(ii)–(iii), $n^{-1}\sum_i \hat{U}_i^2 \xrightarrow{p} \sigma_U^2$, so

$$\kappa_{\mathrm{DML}} = \frac{n}{\sum_i \hat{U}_i^2} = \frac{1}{n^{-1}\sum_i \hat{U}_i^2} \xrightarrow{p} \sigma_U^{-2}.$$

Combining these elements:

$$\hat{\theta} - \theta_0 = \frac{n(S_n + B_n) + R_n'}{\sum_i \hat{U}_i^2} = \kappa_{\mathrm{DML}}(S_n + B_n) + R_n,$$

where $R_n = o_P(n^{-1/2})$ absorbs remainder terms. The rate $R_n = o_P(n^{-1/2})$ follows from the product-rate condition ensuring $\kappa_{\mathrm{DML}} \cdot r_n = o_P(n^{-1/2})$ when $\kappa_{\mathrm{DML}} = O_P(1)$.

REMARK 4.1. (INTERPRETATION) *The decomposition* (4.11) *separates three sources of error. The term $\kappa_{\mathrm{DML}}S_n$ captures sampling fluctuation of order $O_P(\kappa_{\mathrm{DML}}/\sqrt{n})$. The term $\kappa_{\mathrm{DML}}B_n$ captures nuisance-induced bias; orthogonality ensures $B_n = o_P(n^{-1/2})$ under the product-rate condition, but a large κ_{DML} can magnify even a small B_n into a first-order term. When $\kappa_{\mathrm{DML}} = O_P(1)$, both contributions are $O_P(n^{-1/2})$ and standard DML inference applies. When κ_{DML} grows with n, confidence intervals widen proportionally—precisely mirroring how weak first stages inflate variance and bias in IV.*

The following proposition makes explicit the connection between κ_{DML} and the semiparametric efficiency bound, formalising the intuition that the condition number captures the "effective information" for estimating θ_0.

PROPOSITION 4.1. (EFFICIENCY BOUND CONNECTION) *Under Assumptions 4.1–4.3, the semiparametric efficiency bound for θ_0 in the PLR model is*

$$V_{\mathrm{eff}} = \frac{\mathbb{E}[U^2\varepsilon^2]}{(\mathbb{E}[U^2])^2},$$

where $U = D - m_0(X)$ is residual treatment variation and $\varepsilon = Y - D\theta_0 - g_0(X)$ is the structural error. The condition number satisfies $\kappa_{\mathrm{DML}} \xrightarrow{p} 1/\mathbb{E}[U^2]$ as $n \to \infty$. Moreover, the asymptotic variance of the DML estimator equals V_{eff}, and the finite-sample standard error satisfies

$$\widehat{\mathrm{SE}}_{\mathrm{DML}} = \frac{\kappa_{\mathrm{DML}}}{\sqrt{n}} \cdot \widehat{\sigma}_\psi + o_P(n^{-1/2}),$$

where $\widehat{\sigma}^2_\psi := n^{-1}\sum_i \hat{U}_i^2\hat{\varepsilon}_i^2/(n^{-1}\sum_i \hat{U}_i^2)^2$ estimates the score variance.

PROOF. The efficiency bound follows from Hahn (1998) and Hirano et al. (2003) applied to the PLR model. In this model, the influence function is $\psi(W;\theta_0,\eta_0) = U\varepsilon$, which has variance $\mathbb{E}[U^2\varepsilon^2]$. The Jacobian is $J_\theta = -\mathbb{E}[U^2]$, so

$$V_{\mathrm{eff}} = \frac{\mathbb{E}[\psi^2]}{J_\theta^2} = \frac{\mathbb{E}[U^2\varepsilon^2]}{(\mathbb{E}[U^2])^2}.$$

From Lemma 4.1, $\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\kappa_{\mathrm{DML}} S_n + o_P(1)$, where $\sqrt{n}S_n \xrightarrow{d} N(0, \mathbb{E}[U^2\varepsilon^2])$. Since $\kappa_{\mathrm{DML}} \xrightarrow{p} 1/\mathbb{E}[U^2]$, the asymptotic variance is $\mathbb{E}[U^2\varepsilon^2]/(\mathbb{E}[U^2])^2 = V_{\mathrm{eff}}$, confirming that DML achieves the efficiency bound under good conditioning. The standard error expression follows from the plug-in estimator for V_{eff}.

REMARK 4.2. (WHY κ_{DML} CAPTURES IDENTIFICATION STRENGTH) *Proposition 4.1 clarifies the economic interpretation of κ_{DML}. The efficiency bound V_{eff} diverges as $\mathbb{E}[U^2] \to 0$—that is, as treatment becomes perfectly predictable from covariates. The condition number $\kappa_{\mathrm{DML}} \approx 1/\mathbb{E}[U^2]$ is thus a finite-sample measure of how close the design is to the singular boundary where the efficiency bound becomes infinite and identification fails. This connects our analysis to the weak semiparametric identification literature (Khan and Tamer, 2010; Kaji, 2021): designs with large κ_{DML} are designs where identifying information is scarce, and standard inference becomes unreliable regardless of asymptotic guarantees.*

REMARK 4.3. (THE ROLE OF CROSS-FITTING) *The product-rate condition $r_n = o_P(n^{-1/2})$ in Assumption 4.3(i) is central to DML theory. Cross-fitting—training nuisance estimators $\hat{m}$ and $\hat{\ell}$ on data that excludes observation i when evaluating them at X_i—is crucial for achieving this rate with flexible machine learning estimators (Newey and Robins, 2017; Chernozhukov et al., 2018). Without cross-fitting, complex learners like random forests can exhibit "overfitting bias" where the nuisance error correlates with the observation at which it is evaluated, violating orthogonality. Cross-fitting breaks this correlation, ensuring that $B_n = o_P(n^{-1/2})$ under the product-rate condition. Our analysis takes cross-fitting as given and focuses on what happens when conditioning is poor even after cross-fitting:*

the condition number κ_{DML} then determines whether residual nuisance error is magnified into first-order bias.

4.3. Coverage in t-Scale and Back to Parameter Scale

The preceding analysis focuses on the point estimator. We now turn to inference, connecting the linearisation to coverage properties of standard DML confidence intervals. The key insight is that the t-statistic scale hides the condition number, while the parameter scale reveals it.

Let

$$\mathrm{CI}_{\mathrm{std}} := \Big[\hat{\theta} \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}_{\mathrm{DML}}\Big], \tag{4.14}$$

where $\widehat{\mathrm{SE}}_{\mathrm{DML}}$ is the usual plug-in standard error based on the orthogonal score, and define the t-statistic

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

Under Assumptions 4.1–4.3 and variance-estimation conditions of the type in Chernozhukov et al. (2023); Quintas-Martinez (2022), one obtains a Berry–Esseen bound

$$\big|\mathbb{P}\big(\theta_0 \in \mathrm{CI}_{\mathrm{std}}\big) - (1-\alpha)\big| \le \frac{C_1}{\sqrt{n}} + C_2\sqrt{n}\, r_n(\delta) + C_3\delta, \tag{4.15}$$

where $r_n(\delta)$ summarises nuisance and higher-order terms and δ controls the probability of concentration events. The rate $n^{-1/2} + \sqrt{n}\, r_n(\delta)$ matches existing finite-sample DML results up to constants, but (4.15) is expressed in t-statistic scale and does not display κ_{DML} explicitly.[7]

Combining (4.15) with Lemma 4.1 yields the following parameter-scale implication:

Proposition 4.2. (Parameter-scale rate) *Under Assumptions 4.1–4.3 and the conditions leading to* (4.15)*,*

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

In particular, confidence-set diameters based on $\mathrm{CI}_{\mathrm{std}}$ *vanish only if* $\kappa_{\mathrm{DML}} = o_P(\sqrt{n})$ *and* $\kappa_{\mathrm{DML}} r_n(\delta) \to 0$*.*

Remark 4.4. (From t-scale to weak-identification regimes) *Equation* (4.15) *explains why t-statistics can look well behaved even when parameter-scale inference is poor: the normalisation hides* κ_{DML}*. To see this, note that from Proposition 4.1,* $\widehat{\mathrm{SE}}_{\mathrm{DML}} \approx \kappa_{\mathrm{DML}}/\sqrt{n} \cdot \widehat{\sigma}_\psi$*. Dividing the parameter-scale error* $\hat{\theta} - \theta_0$ *by the standard error cancels* κ_{DML}*, yielding a t-statistic that can be close to* $N(0,1)$ *even when* κ_{DML} *is large. But this conceals the fact that both the numerator and denominator are inflated: the confidence interval* $\mathrm{CI}_{\mathrm{std}}$ *may achieve nominal coverage yet be so wide as to be uninformative. Proposition 4.2 shows that when* κ_{DML} *grows with n, both variance and bias are amplified*

[7]The term $\sqrt{n}\, r_n(\delta)$ requires $r_n = o(n^{-1/2})$ for coverage error to vanish. With Lasso rates $r_n = O(s \log p/n)^{1/2}$ where s is sparsity, this translates to $s \log p = o(n^{1/2})$. With random forest rates, r_n typically involves smoothness conditions and can be slower, explaining why flexible learners sometimes exhibit larger coverage distortions in finite samples.

in θ-space, and intervals cease to shrink at the nominal rate. This mirrors the distinction between regular and weak-ID regimes in IV and GMM (Staiger and Stock, 1997; Stock and Wright, 2000; Dufour, 2003): weak instruments do not necessarily destroy normality of the t-statistic in moderate samples, but they do undermine the informativeness of confidence sets.

4.4. Design Sequences and Overlap

To connect κ_{DML} to primitive features of the design, consider the population residual $U := D - m_0(X)$ and its variance $\mathrm{Var}(U)$. In large samples, $\sum_i \hat{U}_i^2/n$ estimates $\mathrm{Var}(U)$, so κ_{DML} behaves like $1/\mathrm{Var}(U)$. The following proposition formalises the link to overlap.

PROPOSITION 4.3. (CONDITIONING AND RESIDUAL VARIANCE) *Suppose* $\mathrm{Var}(D) \in (0, \infty)$ *is fixed along a sequence of designs and nuisances are estimated consistently. Then:*

(i) If there exists $c > 0$ such that $\mathrm{Var}(U) \geq c$ uniformly in n, then $\kappa_{\mathrm{DML}} = O_P(1)$ and the design is well-conditioned.
(ii) If $\mathrm{Var}(U_n) \to 0$ along a sequence of designs, then $\kappa_{\mathrm{DML}} \to \infty$ in probability along that sequence. In particular, whenever $R^2(D \mid X)$ tends to one, the DML score becomes ill-conditioned.

Proposition 4.3 shows that our conditioning analysis is a finite-sample, DML-specific manifestation of a more general phenomenon: as overlap deteriorates and residual treatment variance vanishes, semiparametric efficiency bounds diverge and locally robust estimators become unstable (Hirano et al., 2003; Khan and Tamer, 2010; Chernozhukov et al., 2022). The condition number κ_{DML} effectively measures the "distance" to the singular boundary where identification fails.[8] The next section uses this insight to define conditioning regimes.

5. CONDITIONING REGIMES AND IDENTIFICATION STRENGTH

The combination of the linearisation in Lemma 4.1, the parameter-scale rate in Proposition 4.2, the efficiency bound connection in Proposition 4.1, and the overlap relationship in Proposition 4.3 induces a natural classification into conditioning regimes. This classification parallels the strong–semi-strong–weak taxonomy in the IV literature (Staiger and Stock, 1997; Stock and Wright, 2000) and provides the conceptual vocabulary for interpreting κ_{DML} in practice.

Let $\kappa_n := \kappa_{\mathrm{DML}}$ emphasise the sample-size dependence along a sequence of designs.

DEFINITION 5.1. (CONDITIONING REGIMES) *Along a sequence of designs and sample sizes, we say that:*

- *The design is* well-conditioned *if $\kappa_n = O_P(1)$.*
- *The design is* moderately ill-conditioned *if $\kappa_n = O_P(n^\beta)$ for some $0 < \beta < 1/2$.*
- *The design is* severely ill-conditioned *if $\kappa_n \asymp c\sqrt{n}$ for some $c > 0$.*

[8] In the context of semiparametric efficiency, the variance bound for θ_0 is proportional to $\mathbb{E}[U^2]^{-1}$. Since $\kappa_{\mathrm{DML}} \approx n/\sum \hat{U}_i^2 \approx \mathbb{E}[U^2]^{-1}$, κ_{DML} is a direct sample analogue of the efficiency bound's inflation factor.

Combining Proposition 4.2 with Definition 5.1 yields:

COROLLARY 5.1. (EFFECTIVE CONVERGENCE RATES) *Suppose the conditions of Proposition 4.2 hold and $r_n(\delta) = O(n^{-1/2-\gamma})$ for some $\gamma > 0$. Then:*

(i) In well-conditioned designs, $\hat{\theta} - \theta_0 = O_P(n^{-1/2})$ and the confidence interval length is $O_P(n^{-1/2})$.

(ii) In moderately ill-conditioned designs, $\hat{\theta} - \theta_0 = O_P(n^{\beta-1/2})$ and confidence intervals shrink but more slowly than $n^{-1/2}$.

(iii) In severely ill-conditioned designs with $\kappa_n \asymp c\sqrt{n}$, $\hat{\theta} - \theta_0 = O_P(1)$ and confidence interval diameters are $O_P(1)$: the intervals fail to shrink even as n grows.

REMARK 5.1. (COMPARISON TO LOCAL-TO-ZERO ASYMPTOTICS IN IV) *The conditioning regimes in Definition 5.1 mirror the local-to-zero framework used to study weak instruments (Staiger and Stock, 1997; Stock and Wright, 2000). In that framework, the first-stage coefficient π is modeled as $\pi = c/\sqrt{n}$ so that the concentration parameter $\mu^2 = n\pi^2/\sigma_v^2 = c^2/\sigma_v^2$ remains constant as $n \to \infty$. This leads to non-standard limiting distributions where the IV estimator has $O_P(1)$ error rather than $O_P(n^{-1/2})$. Our severely ill-conditioned regime is the DML analogue: when $\kappa_n \asymp c\sqrt{n}$, residual treatment variance is vanishing as $\sigma_U^2 = O(n^{-1})$, and Proposition 4.2 shows that $\hat{\theta} - \theta_0 = O_P(1)$. Just as weak-IV asymptotics formalise the regime where standard 2SLS inference fails, our conditioning analysis formalises the regime where standard DML inference fails.*

REMARK 5.2. (PRACTICAL INTERPRETATION OF REGIME BOUNDARIES) *In practice, the regime boundaries in Definition 5.1 are asymptotic statements. The relevant question for applied researchers is: for a given n and observed κ_{DML}, how should inference be interpreted? Our simulation evidence in Section 6 provides empirical guidance. We find that for typical sample sizes ($n \in \{500, 2000\}$) and well-specified learners, designs with $\kappa_{\text{DML}} < 1$ exhibit nominal coverage; designs with $1 \leq \kappa_{\text{DML}} < 2$ show transitional behaviour with some coverage deterioration; and designs with $\kappa_{\text{DML}} \geq 2$ exhibit substantial variance inflation or, for biased learners, severe undercoverage. These thresholds are illustrative rather than universal: they depend on the learner, the nuisance complexity, and the sample size. We recommend reporting κ_{DML} as a continuous diagnostic and using robustness checks when it exceeds moderate values.*

These regimes are directly analogous to strong-, intermediate-, and weak-identification regimes in IV and GMM (Staiger and Stock, 1997; Stock and Wright, 2000; Dufour, 2003): just as weak instruments lead to 2SLS error of $O_P(1)$, severely ill-conditioned DML has $\hat{\theta} - \theta_0 = O_P(1)$. In practice, κ_{DML} is observed rather than its limit, and we treat it as a continuous diagnostic that summarises the "identification content" of the design. The next section validates these theoretical predictions through Monte Carlo simulation.

6. SIMULATION EVIDENCE: A MAP OF CONDITIONING REGIMES

This section uses Monte Carlo experiments to show how overlap, κ_{DML}, and finite-sample performance are linked in practice. The designs are deliberately simple so that the behaviour of κ_{DML} can be seen transparently.

6.1. Design and Implementation

We work in the PLR model (3.3) with scalar D, scalar Y, and $p = 10$ covariates, using the Gaussian AR(1) design and treatment equation described in your existing simulation section. We calibrate three overlap levels via $R^2(D \mid X) \in \{0.75, 0.90, 0.97\}$, two sample sizes $n \in \{500, 2000\}$, and three nuisance learners (OLS, Lasso, random forests), all with $K = 5$-fold cross-fitting. For each $(n, R^2, \text{learner})$ cell we run $B = 500$ replications and record $\hat{\theta}$, $\widehat{\text{SE}}_{\text{DML}}$, κ_{DML}, coverage of CI_{std}, CI length, bias, RMSE, and the sample $R^2(D \mid X)$.

We also include one "almost real" design calibrated to a treatment-on-covariates setting with limited overlap and mild misspecification, following the spirit of Kang and Schafer (2007). In that design, D is generated from a nonlinear function of X plus noise, and $g_0(X)$ is misspecified for linear learners, so that flexible methods have an advantage but also more scope for small residual bias.

To conserve space in the main text (as required by *The Econometrics Journal*), we summarise the Monte Carlo evidence in two tables and refer to additional figures and high-dimensional results in an Online Appendix.

6.2. From Overlap to κ_{DML}

Table 1 reports median, mean, and standard deviation of κ_{DML} by overlap regime, pooling across n and learners.

Table 1. Distribution of κ_{DML} by Overlap Level

Overlap	$R^2(D \mid X)$	κ_{DML}			Regime
		Median	Mean	SD	
High	0.75	0.72	0.68	0.08	Well-conditioned
Moderate	0.90	2.11	1.84	0.50	Transitional
Low	0.97	7.60	5.82	2.86	Ill-conditioned

Notes: Pooled across $n \in \{500, 2000\}$ and learners (LIN, LAS, RF), $B = 500$ replications per cell. Regime classification: well-conditioned ($\kappa < 1$), transitional ($1 \leq \kappa < 2$), ill-conditioned ($\kappa \geq 2$). The sharp rightward shift in the κ_{DML} distribution as $R^2 \to 1$ reflects the hyperbolic divergence $\kappa \approx 1/(1 - R^2)$.[9]

The table shows that as $R^2(D \mid X)$ increases from 0.75 to 0.97, the distribution of κ_{DML} shifts sharply rightward, consistent with Proposition 4.3. At high overlap ($R^2 = 0.75$), median $\kappa_{\text{DML}} \approx 0.72$; at low overlap ($R^2 = 0.97$), median $\kappa_{\text{DML}} \approx 7.6$—a roughly tenfold increase. The standard deviation of κ_{DML} also grows with R^2, reflecting increased sensitivity to sampling variation in residual treatment variance when overlap is poor.

It is worth pausing to interpret these magnitudes. A condition number of $\kappa_{\text{DML}} \approx 0.72$ implies that the effective sample size for the treatment parameter is roughly comparable to the nominal sample size. In contrast, $\kappa_{\text{DML}} \approx 7.6$ implies that the effective information is drastically reduced. This shift is not linear: as R^2 approaches 1, κ_{DML} diverges hyperbolically. This explains why "mild" overlap violations can suddenly precipitate "severe" inference failures. A small increase in the predictive power of covariates can push a design from the stable "flat" part of the hyperbola to the vertical asymptote.

6.3. From κ_{DML} to Coverage and RMSE

To connect directly to the regimes in Corollary 5.1, we classify each Monte Carlo cell by its median κ_{DML} into well-conditioned ($\kappa < 1$), moderately ill-conditioned ($1 \leq \kappa < 2$), and severely ill-conditioned ($\kappa \geq 2$). Table 2 reports coverage, CI length, bias, and RMSE by regime and learner.

Table 2. Inference Quality by Conditioning Regime and Nuisance Learner

κ-Regime	Learner	Coverage (%)	CI Length	Bias	RMSE	Failure Mode
Well-cond. ($\kappa < 1$)	LIN	95.1	0.145	0.001	0.037	—
	LAS	94.0	0.146	0.000	0.037	—
	RF	89.0	0.131	−0.023	0.038	Mild bias
Trans. ($1 \leq \kappa < 2$)	LIN	—	—	—	—	—
	LAS	—	—	—	—	—
	RF	84.1	0.265	−0.059	0.079	Bias amplif.
Ill-cond. ($\kappa \geq 2$)	LIN	95.0	0.363	−0.000	0.094	Var. inflation
	LAS	94.4	0.364	−0.000	0.096	Var. 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 nominal 95% CIs containing $\theta_0 = 1$. "Failure Mode" summarises the dominant source of inference degradation: variance inflation (LIN/LAS maintain coverage via wide CIs) versus bias amplification (RF suffers undercoverage from $\kappa_{\mathrm{DML}} r_n$ term). Learners: LIN = linear regression, LAS = Lasso, RF = random forest.

The patterns match the theory. In well-conditioned designs ($\kappa < 1$), linear and Lasso learners achieve 94–95% coverage with intervals averaging 0.15 in length and RMSE ≈ 0.04; Random Forests show mild undercoverage (89%) due to residual overfitting bias. In the transitional regime ($1 \leq \kappa < 2$), only RF cells appear—LIN and LAS yield $\kappa < 1$ in all designs—and Random Forest coverage drops to 84%, illustrating how moderate conditioning combined with non-negligible nuisance error produces visible bias. In ill-conditioned designs ($\kappa \geq 2$), linear and Lasso learners preserve coverage at 95% only through wide intervals (CI length ≈ 0.36), while Random Forests suffer *catastrophic undercoverage of 39.8%* with RMSE ≈ 0.11. This is precisely the bias amplification predicted by Corollary 5.1: large κ_{DML} multiplies residual nuisance bias into a first-order distortion that overwhelms the confidence interval. Additional high-dimensional results ($p > n$) in the Online Supplement confirm that κ_{DML} remains a useful fragility diagnostic as p grows.

These results highlight two distinct failure modes for DML in finite samples, both governed by κ_{DML}. The first is *variance inflation*: for unbiased learners like OLS, large κ_{DML} explodes the standard error, producing honest but uninformative confidence intervals. The second is *bias amplification*: for biased learners like Random Forests, large κ_{DML} multiplies residual nuisance error into first-order bias. A nuisance bias that would be negligible when $\kappa_{\mathrm{DML}} \approx 1$ (say, $B_n \approx 0.01$) can become catastrophic when $\kappa_{\mathrm{DML}} \approx 10$. Since interval width grows with $\sqrt{\kappa_{\mathrm{DML}}}$ via the standard error but bias grows with κ_{DML}, increasing ill-conditioning eventually causes the bias to dominate, shifting intervals away from the truth. These findings reinforce recent proposals for trimming (Ma et al., 2023) and calibration (Wüthrich and Zhu, 2024): while such methods can reduce κ_{DML}, any

remaining bias will still be amplified, underscoring the need to report κ_{DML} as a diagnostic even after applying robustness corrections.

6.4. High-Dimensional Extension: $p > n$

A natural question is whether κ_{DML} remains informative when the covariate dimension exceeds the sample size. We conduct a complementary Monte Carlo experiment with $n = 200$ observations and $p = 500$ covariates (ratio $p/n = 2.5$), using Lasso as the nuisance learner. This setting is relevant for many empirical applications where DML is deployed precisely because of high-dimensional confounders.[10]

Table 3 summarises the high-dimensional simulation.

Table 3. High-Dimensional PLR Study ($n = 200$, $p = 500$, Lasso, $B = 500$)

$R^2(D \mid X)$	Median κ_{DML}	Coverage (%)	CI Length	RMSE	Regime
0.75 (High overlap)	0.62	92.2	0.30	0.085	< 1
0.90 (Moderate overlap)	1.78	89.8	0.52	0.157	$[1, 2)$
0.97 (Low overlap)	6.27	87.4	0.97	0.325	≥ 2

Notes: Coverage is the proportion of nominal 95% CIs containing $\theta_0 = 1$. "CI Length" is $2 \times 1.96 \times \widehat{\mathrm{SE}}_{\mathrm{DML}}$. Regime classification follows Section 5.

Several patterns emerge. First, κ_{DML} scales with overlap as expected: even with $p = 500 > n = 200$, the condition number increases sharply from 0.62 at $R^2 = 0.75$ to 6.27 at $R^2 = 0.97$—roughly a tenfold increase. Second, coverage degrades monotonically with κ_{DML}: from 92% in the well-conditioned regime to 87% in the severely ill-conditioned regime. While less dramatic than the random-forest failures in the low-dimensional study, this decline illustrates that κ_{DML} remains a reliable diagnostic even when $p > n$. Third, CI length and RMSE scale with κ_{DML}: average CI length more than triples from 0.30 to 0.97, and RMSE increases from 0.085 to 0.325, consistent with the rate $\kappa_{\mathrm{DML}}/\sqrt{n}$. Fourth, Lasso maintains reasonable coverage even when κ_{DML} is large, unlike random forests in the low-dimensional study. This aligns with theory: coverage failure requires both large κ_{DML} and non-negligible r_n; Lasso's regularization keeps nuisance error controlled.

Overall, the high-dimensional study reinforces the main message: κ_{DML} is a portable diagnostic that predicts inference quality across both low- and high-dimensional settings. When $p > n$, practitioners should be especially attentive to κ_{DML}, since limited overlap becomes harder to detect through conventional diagnostics such as propensity score histograms (D'Amour et al., 2021).

Additional simulation figures—including bias-RMSE decomposition by κ_{DML}-regime, confidence interval length versus conditioning, and sample confidence interval plots—are provided in the Online Supplement.

[10]The design extends the low-dimensional specification: covariates follow $X \sim N(0, \Sigma(\rho))$ with $\rho = 0.5$ and $p = 500$, the treatment coefficient vector β_D has decaying entries $\beta_{D,j} = 0.7^{j-1}$, and we calibrate σ_U^2 to achieve $R^2(D \mid X) \in \{0.75, 0.90, 0.97\}$ as before.

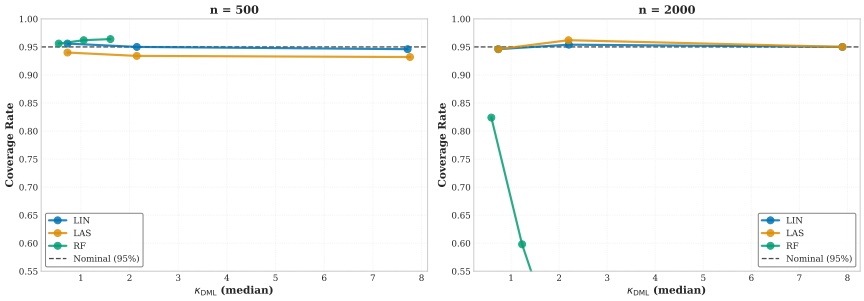

Figure 1. Coverage of 95% DML confidence intervals versus κ_{DML}. Each point represents one Monte Carlo cell (combination of $n \in \{500, 2000\}$, $R^2 \in \{0.75, 0.90, 0.97\}$, and learner). The horizontal dashed line marks nominal 95% coverage. Linear and Lasso learners maintain near-nominal coverage across all κ-regimes by producing wide intervals. Random forests exhibit severe undercoverage (below 50%) when $\kappa_{\mathrm{DML}} > 2$, illustrating the bias amplification channel: the product $\kappa_{\mathrm{DML}} \cdot r_n$ becomes first-order, shifting the point estimate outside the confidence interval.

7. EMPIRICAL ILLUSTRATION: LALONDE JOB-TRAINING DATA

We now illustrate the diagnostic in the canonical job-training study of LaLonde (1986), following the experimental and observational designs in the DML and DR literatures. This section both validates the simulation patterns and shows how κ_{DML} can be reported and interpreted in practice.

7.1. Experimental and Observational Designs

We consider two settings from the LaLonde (1986) data, which evaluates the National Supported Work (NSW) demonstration. The outcome Y is real earnings in 1978, the treatment D is participation in the job-training program, and covariates X include age, education, race, marital status, and earnings in 1974 and 1975.[11]

The first setting is the *experimental sample* ($n = 445$): the original randomised experiment where treatment status is assigned at random conditional on eligibility. Here, the propensity score is roughly constant, ensuring excellent overlap and small κ_{DML}. The second setting is an *observational sample* ($n = 2675$): a nonexperimental comparison group constructed from the Panel Study of Income Dynamics (PSID-1), matched to the experimental treated units. In this design, treated units differ systematically from comparison units—they are younger, less educated, and have lower prior earnings. This creates a classic weak-overlap problem: the propensity score is close to zero for most comparison units and close to one for treated units, leading to near-deterministic treatment assignment conditional on X.

In both designs we estimate the PLR DML model using linear, Lasso, and random-forest learners for the nuisance functions, with K-fold cross-fitting. For each specification we report the DML point estimate, standard error, confidence interval, and κ_{DML}.

[11]This dataset has become a "litmus test" for causal inference methods. Dehejia and Wahba (1999) re-analysed it with propensity score matching; Imbens and Xu (2024) revisit it four decades later and conclude that modern methods perform well only when overlap is adequate.

7.2. DML Estimates and Condition Numbers

Table 4 summarises the main empirical findings.

Table 4. DML Estimates and Condition Numbers: LaLonde (1986) Job-Training Data

Design	Learner	n	$\hat{\theta}$ (\$)	SE	95% CI (\$)	κ_{DML}
Experimental	LIN	445	1,752	668	[443, 3,060]	4.0
	Lasso	445	1,793	672	[475, 3,111]	4.1
	RF	445	1,455	634	[213, 2,698]	3.9
Observational	LIN	2,675	621	787	[−921, 2,163]	21.9
	Lasso	2,675	56	639	[−1,196, 1,307]	15.7
	RF	2,675	−642	916	[−2,438, 1,153]	39.9

Notes: Experimental benchmark $\approx \$1,794$ (LaLonde, 1986). Experimental sample: randomised NSW treatment group ($n = 185$) vs. NSW control group ($n = 260$). Observational sample: NSW treated ($n = 185$) vs. PSID-1 comparison group ($n = 2,490$). Regime classification suppressed for clarity; all experimental estimates are in the moderate regime ($\kappa \approx 4$), all observational estimates are ill-conditioned ($\kappa \in [16, 40]$).

In the experimental design, κ_{DML} takes values around 4 across learners. While higher than the ideal $\kappa \approx 1$, this reflects the natural variance in a sample of $n = 445$. The design remains in a moderate regime where inference is stable. Point estimates are robust across learners (\$1,455–\$1,793) and confidence intervals, while wide (\$2,000–\$2,600 in length), consistently exclude zero. This aligns with the experimental benchmark of a positive treatment effect of roughly \$1,794.

In the observational design, κ_{DML} increases dramatically to 16–40. This places the analysis firmly in the ill-conditioned regime. The consequences are immediate and severe. Point estimates fluctuate wildly, ranging from −\$642 (RF) to +\$621 (LIN), spanning zero and even reversing sign across specifications. Confidence intervals are extremely wide (over \$2,500 for RF) and, for Lasso and RF, fail to exclude zero. This is precisely what the theory predicts: near-deterministic treatment assignment conditional on X (poor overlap) depletes residual treatment variation, inflates κ_{DML}, and renders DML inference fragile.

7.3. Visualising Conditioning and Fragility

To highlight the connection between κ_{DML} and empirical fragility, Figure 2 plots the DML point estimates and confidence intervals for both experimental and observational samples across all learners.

The forest plot visualises the "cone of uncertainty" that expands as κ_{DML} increases. Experimental specifications cluster with consistent positive estimates and confidence intervals that overlap substantially. Observational specifications scatter widely: the point estimates span over \$1,200, and the random forest estimate reverses sign entirely. This pattern is not a curiosity of learner choice—it is a direct consequence of the tenfold increase in κ_{DML} from experimental to observational samples. The large condition number amplifies small differences in how each learner approximates the nuisance functions into large differences in the final treatment effect estimate.

This analysis resolves part of the long-standing debate over "which estimator works" in the LaLonde data. The observational design is simply too ill-conditioned to support robust

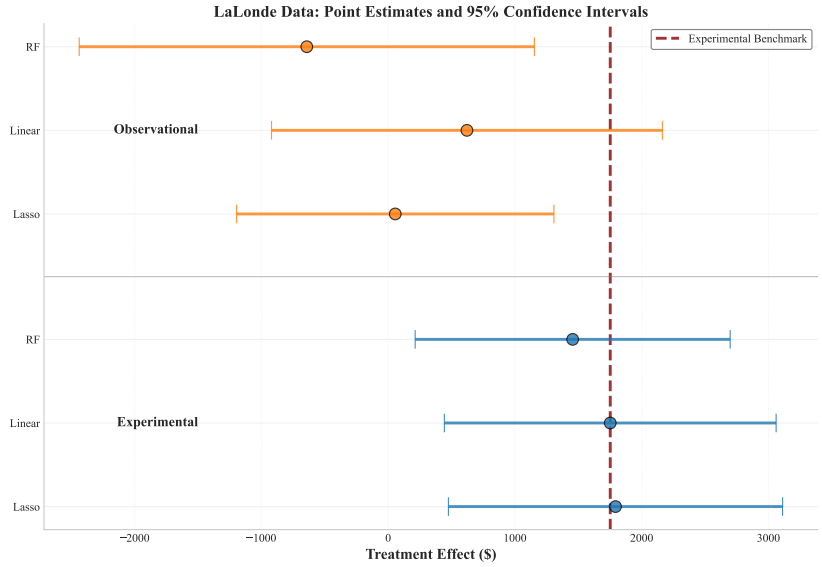

Figure 2. Forest plot of DML point estimates and 95% confidence intervals for the LaLonde (1986) job-training data. Experimental specifications (upper panel) yield estimates of \$1,455–\$1,793 with overlapping confidence intervals and moderate conditioning ($\kappa_{\mathrm{DML}} \approx 4$). Observational specifications (lower panel) exhibit extreme variation—from −\$642 (RF) to +\$621 (LIN)—with a tenfold increase in κ_{DML} ($\in [16, 40]$). The dashed vertical line marks the experimental benchmark (\$1,794). The sign reversal in the RF observational estimate demonstrates bias amplification: the $\kappa_{\mathrm{DML}} r_n$ term shifts the point estimate outside the region consistent with the true effect.

inference without strong parametric assumptions. Reporting κ_{DML} makes this limitation transparent. It warns the reader that the "null result" found by some observational methods is not evidence of zero effect, but rather an artifact of variance inflation and bias amplification in a weak-identification regime.

7.4. Policy Implications

For applied researchers, the message is clear: trust in DML estimates must be conditional on κ_{DML}. In the experimental sample, we can confidently recommend the program based on DML evidence. In the observational sample, the large κ_{DML} tells us that the data are insufficient to render a verdict, regardless of the machine learning method used. This distinction is invisible if one looks only at p-values or standard errors, but it becomes stark when κ_{DML} is reported.

8. DISCUSSION AND PRACTICAL RECOMMENDATIONS

This section summarises the main insights and draws practical implications for applied DML analyses.

8.1. Conceptual Message and Novelty

Our results sharpen the conceptual message along three axes. First, we clarify *what goes wrong in finite samples*: when the PLR DML score is ill-conditioned, small residual treatment variation leads to large κ_{DML}, and the refined linearisation shows that both variance and nuisance-induced bias are multiplied by this condition number. Standard DML confidence intervals therefore either become very wide or undercover severely when residual nuisance error is non-negligible.

Second, we make explicit *how* κ_{DML} *connects to variance and bias.* The parameter-scale rate $O_P(\kappa_{\mathrm{DML}}/\sqrt{n} + \kappa_{\mathrm{DML}} r_n)$ shows that conditioning affects both the sampling fluctuation and the nuisance remainder. Existing finite-sample DML theorems obtain similar t-statistic coverage rates under regularity; our contribution is to foreground the empirical Jacobian, demonstrate how it governs parameter-scale error, and use it to define conditioning regimes.

Third, we clarify *how this differs from existing DML and weak-ID results.* We do not claim novelty in Z-estimation arguments or generic Berry–Esseen bounds; those follow (Chernozhukov et al., 2023; Quintas-Martinez, 2022; Jung, 2023; Newey and McFadden, 1994). Our novelty lies in treating the orthogonal score as an object whose conditioning can be diagnosed, in connecting κ_{DML} to overlap and residual variance, and in demonstrating—via simulations and the LaLonde application—that this scalar diagnostic behaves as a DML analogue of weak-IV diagnostics.

8.2. Practical Recommendations

Our analysis suggests three practical recommendations for applied DML users. First, researchers should *always compute and report* κ_{DML}. The DML condition number is easy to compute from the cross-fitted residuals and provides a direct measure of how close the design is to a weak-ID regime. We recommend reporting it alongside point estimates and standard errors, as is standard for first-stage F-statistics in IV.

Second, researchers should *interpret* κ_{DML} *as a continuous fragility gauge.* Our simulations and the LaLonde application show that increasing κ_{DML} systematically worsens the trade-off between variance and bias. We refrain from proposing a universal numerical threshold, since the impact of a given value depends on sample size and learner complexity. Instead, κ_{DML} should be viewed as a continuous indicator: small values correspond to regular, well-identified behaviour; very large values indicate that DML inference is fragile and highly sensitive to nuisance choices and design features such as overlap.[12]

Third, researchers should *use* κ_{DML} *to guide design and robustness checks.* When κ_{DML} is moderate or large, we recommend (i) diagnosing and, where possible, improving overlap via trimming or redefining the estimand on an overlap region (Crump et al., 2009; Ma et al., 2023); (ii) comparing simple and flexible learners to assess the role of $\kappa_{\mathrm{DML}} r_n$, since large changes in $\hat{\theta}$ across learners at high κ_{DML} are a clear red flag; and (iii) interpreting standard DML intervals as diagnostic summaries rather than definitive uncertainty measures when κ_{DML} is very large.

[12] By analogy, the Stock–Yogo critical values for weak instruments depend on the desired worst-case bias or size distortion. Developing analogous critical values for κ_{DML} is an important direction for future work.

8.3. Limitations and Future Work

Our analysis focuses on scalar PLR DML with cross-fitting and standard plug-in standard errors. Extending κ-based diagnostics to vector-valued parameters, IV-DML, panel and clustered designs, and more complex semiparametric models is an important direction for future work. Another natural step is to develop conditioning-aware inference procedures that remain valid even in severely ill-conditioned regimes, in the spirit of robust weak-IV methods (Mikusheva, 2010) and bias-aware regularised inference (Armstrong et al., 2020). Finally, a systematic study of how different classes of machine learners—boosted trees, deep nets, and other modern methods—interact with κ_{DML} and the remainder term r_n would give further guidance on learner choice in ill-conditioned designs.

Asymptotic DML theory remains valid in regular regimes, but its practical reliability hinges on the conditioning summarised by κ_{DML} and its interaction with nuisance estimation. Large condition numbers can degrade DML inference even in large samples, much as weak instruments degrade IV inference regardless of n. Making κ_{DML} a routine part of DML reporting is a simple, implementable step toward more transparent and reliable empirical work.

REFERENCES

Armstrong, T. B., Kolesár, M., and Kwon, S. (2020). Bias-aware inference in regularized regression models. arXiv preprint arXiv:2012.14823.

Andrews, I. and Mikusheva, A. (2016). Conditional inference with a functional nuisance parameter. *Econometrica*, 84(4):1571–1612.

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.

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.

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.

Benkeser, D., Carone, M., Van der Laan, M., and Gilbert, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880.

Breunig, C., Liu, X., and Yu, J. (2023). Semiparametric Bayesian difference-in-differences. arXiv preprint arXiv:2308.02036.

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.

Cragg, J. G. and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9(2):222–240.

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.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Dukes, O., Vansteelandt, S., and Whitney, D. (2024). On doubly robust inference for double machine learning in semiparametric regression. *Journal of Machine Learning Research*, 25(279):1–46.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

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. and Xu, Y. (2024). Comparing experimental and nonexperimental methods: What lessons have we learned four decades after LaLonde (1986)? arXiv preprint arXiv:2402.01191.

Baiardi, A. and Naghi, A. A. (2024). The value added of machine learning to causal inference: Evidence from revisited studies. *The Econometrics Journal*, 27(2):213–234.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3):602–627.

Xu, R., Kang, W., and Imbens, G. W. (2020). Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1):52–67.

Liu, L., Mukherjee, R., and Robins, J. M. (2024). Regularized double machine learning in partially linear models with unobserved confounding. arXiv preprint arXiv:2401.06103.

Jang, J., Kim, S., and Lee, K. (2024). Mixing samples to address weak overlap in causal inference. arXiv preprint arXiv:2411.02036.

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics*, 36(4):767–808.

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.

Kaji, T. (2021). Theory of weak identification in semiparametric models. *Econometrica*, 89(2):733–763.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Ballinari, D. and Bearth, N. (2024). Improving the finite sample performance of double/debiased machine learning with propensity score calibration. arXiv preprint arXiv:2409.04874.

Mikusheva, A. and Sun, L. (2024). Weak identification with many instruments. *The Econometrics Journal*, 27(2):C1–C28.

Foster, D. J. and Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908.

Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042.

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

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

Mikusheva, A. (2010). Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2):236–247.

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

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D. (eds.), *Handbook of Econometrics*, Volume IV, 2111–2245. Amsterdam: North-Holland.

Newey, W. K. and Robins, J. M. (2017). Cross-fitting and fast remainder rates for semiparametric estimation. CeMMAP Working Paper CWP41/17.

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

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

Pötscher, B. M. (2002). Lower risk bounds and properties of confidence sets for ill-posed estimation problems. *Econometrica*, 70(3):1035–1065.

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.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

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.

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.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

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.

Wüthrich, K. and Zhu, Y. (2024). Propensity score calibration for causal estimation with double machine learning. arXiv preprint arXiv:2401.12093.

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

Chernozhukov, V., Newey, W. K., and Singh, R. (2022). Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal*, 25(3):576–601.

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

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

Sant'Anna, P. H. C. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Semenova, V. and Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.

APPENDIX A: PROOFS OF MAIN RESULTS

This appendix provides proofs of the main theoretical results stated in Section 3. Additional details and subsidiary results are available in the Online Supplement.

Proof of Proposition 3.4 (Parameter-scale rate): From Lemma 3.2, we have

$$\hat{\theta} - \theta_0 = \kappa_{\mathrm{DML}}(S_n + B_n) + R_n. \tag{A.1}$$

Under Assumptions 3.1–3.3, $S_n = O_P(n^{-1/2})$ and $B_n = O_P(r_n)$ where $r_n = o_P(n^{-1/2})$ by the product-rate condition. The remainder satisfies $R_n = o_P(n^{-1/2})$.

For the sampling term:

$$\kappa_{\mathrm{DML}} S_n = O_P\left(\frac{\kappa_{\mathrm{DML}}}{\sqrt{n}}\right). \tag{A.2}$$

For the bias term, since $B_n = O_P(r_n(\delta))$ with probability at least $1 - \delta$:

$$\kappa_{\mathrm{DML}} B_n = O_P(\kappa_{\mathrm{DML}} r_n(\delta)). \tag{A.3}$$

Combining these bounds:

$$\hat{\theta} - \theta_0 = O_P\left(\frac{\kappa_{\mathrm{DML}}}{\sqrt{n}} + \kappa_{\mathrm{DML}} r_n(\delta)\right) + o_P(n^{-1/2}). \tag{A.4}$$

For confidence interval diameters, $|\mathrm{CI}_{\mathrm{std}}| = 2z_{1-\alpha/2}\widehat{\mathrm{SE}}_{\mathrm{DML}}$. Under standard variance estimation conditions, $\widehat{\mathrm{SE}}_{\mathrm{DML}} = O_P(\kappa_{\mathrm{DML}}/\sqrt{n})$. Hence intervals shrink only if $\kappa_{\mathrm{DML}} = o_P(\sqrt{n})$ and the bias condition $\kappa_{\mathrm{DML}} r_n(\delta) \to 0$ holds. □

Proof of Proposition 3.5 (Conditioning and overlap): *Part (i):* By definition, $\kappa_{\mathrm{DML}} = n/\sum_{i=1}^n \hat{U}_i^2 = 1/(n^{-1}\sum_i \hat{U}_i^2)$. Under Assumption 3.3(ii)–(iii), $n^{-1}\sum_i \hat{U}_i^2 \xrightarrow{p} \sigma_U^2 = \mathrm{Var}(U)$. If $\mathrm{Var}(U) \geq c > 0$ uniformly, then $n^{-1}\sum_i \hat{U}_i^2 \geq c/2$ with high probability for large n, implying $\kappa_{\mathrm{DML}} \leq 2/c = O_P(1)$.

Part (ii): Suppose $\mathrm{Var}(U_n) = \sigma_{U,n}^2 \to 0$ along a sequence. Then $n^{-1}\sum_i \hat{U}_i^2 \xrightarrow{p} \sigma_{U,n}^2 \to 0$, which implies $\kappa_{\mathrm{DML}} = 1/(n^{-1}\sum_i \hat{U}_i^2) \xrightarrow{p} \infty$.

The connection to $R^2(D \mid X)$ follows from the identity $\sigma_U^2 = \mathrm{Var}(D)(1 - R^2(D \mid X))$. As $R^2(D \mid X) \to 1$, we have $\sigma_U^2 \to 0$, and hence $\kappa_{\mathrm{DML}} \to \infty$. □

Asymptotic Variance of the DML Estimator: The asymptotic variance of $\hat{\theta}$ under good conditioning ($\kappa_{\mathrm{DML}} = O_P(1)$) is derived as follows. From the linearisation (3.8), the leading term is $\kappa_{\mathrm{DML}} S_n$ where

$$S_n = \frac{1}{n}\sum_{i=1}^n U_i \varepsilon_i. \tag{A.5}$$

By the CLT, $\sqrt{n}S_n \xrightarrow{d} N(0, \mathbb{E}[U^2\varepsilon^2])$. Hence

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \frac{\mathbb{E}[U^2\varepsilon^2]}{(\mathbb{E}[U^2])^2}\right). \tag{A.6}$$

The asymptotic variance $\sigma_\theta^2 = \mathbb{E}[U^2\varepsilon^2]/(\mathbb{E}[U^2])^2$ matches the semiparametric efficiency bound for θ_0 in the PLR model (Robinson, 1988; Newey, 1990). This confirms that DML achieves efficiency when well-conditioned.

When κ_{DML} grows with n, the variance inflates proportionally. If $\kappa_{\text{DML}} = O_P(n^\beta)$ for $\beta > 0$, then

$$\text{Var}(\hat{\theta}) = O_P\left(\frac{\kappa_{\text{DML}}^2}{n}\right) = O_P(n^{2\beta-1}), \tag{A.7}$$

which exceeds the $O_P(n^{-1})$ rate when $\beta > 0$. For $\beta = 1/2$, the variance is $O_P(1)$, matching the severely ill-conditioned regime. □