Debiased Machine Learned Identification for Causal Inference in High-Dimensional Settings with Unobserved Confounders

Tamer Çetin*

October 2025

Abstract

This paper introduces a robust, end-to-end double/debiased Machine-Learned Instrumental Variables (DML-IV) estimator that integrates nonparametric, fully data-driven ML methods with the classical IV framework to deliver transparent and reproducible causal inference in high-dimensional settings with unobserved confounders. Beyond establishing the identification and theoretical guarantees—Neyman orthogonality, double-robustness, $\sqrt{n}$-consistency, asymptotic normality, and attainment of the semiparametric efficiency bound under standard completeness and regularity conditions—the paper provides a complete, full-stack ML engineering pipeline: from cross-fitting, regularization, and hyperparameter tuning, through fully-data-driven, nonparametric feature-importance–guided dimensionality reduction and best-model selection from a rich library of learners, to bias-corrected moment construction and final statistical inference. Monte Carlo experiments and an empirical application estimating the return to education corroborate both the theoretical properties and the practical performance of the proposed estimator.

1 Introduction

Causal inference often relies on instrumental variables (IV) to address endogeneity arising from unobserved confounders. Traditional IV methods exploit instruments correlated with endogenous regressors but uncorrelated with structural errors to isolate exogenous variation and identify causal effects (Angrist and Kruger, 1991; Angrist and Imbens, 1994; Angrist, Imbens and Rubin, 1996; Heckman and Vytlacil , 2005; Horowitz, 2011; Imbens and Rubin,

*University of California, Berkeley. Email: tamercetin@berkeley.edu

2015). However, weak instruments can induce bias and invalidate standard asymptotic results (Staiger and Stock, 1997; Andrews and Stock, 2005). Nonparametric IV techniques offer greater flexibility by relaxing functional form assumptions (Newey and Powell, 2003; Horowitz and Lee, 2007), yet they remain underutilized due to computational and interpretational challenges (Chen and Pouzo, 2012; Darolles et al., 2011).

The rise of ML methods has led to new IV strategies that reframe both the prediction of the endogenous regressor and the construction of instruments and controls as supervised-learning tasks in the first stage. (Belloni et al., 2012; Chen et al., 2022; Hartford et al., 2017; Mullainathan and Spiess, 2017; Sun, Cui, and Tchetgen, 2022). These ML-based methods exploit ML algorithms' ability to handle high dimensionality[1] and nonlinear relationships, potentially improving causal identification beyond traditional parametric frameworks. Within this literature, linear ML models (Lasso, Ridge, Elastic Net) and nonlinear learners (Random Forests, boosting, neural networks) have been employed to stabilize first stages and capture complex patterns (Tibshirani, 1996; Belloni and Chernozhukov, 2013; Hoerl and Kennard, 1970; Hansen and Kozbur, 2014; Zou and Hastie, 2005; Breiman, 2001; Friedman, 2001; Farrell et al., 2021; Chernozhukov et al., 2021; Athey and Imbens, 2019; Wager and Athey, 2018). Finally, DML combines sample splitting and cross-fitting to mitigate overfitting while preserving $\sqrt{n}$-consistency and asymptotic normality (Chernozhukov et al., 2018).

Despite these developments, existing ML–IV approaches often fall short of a fully debiased and fully data-driven solution. Many rely on simple plug-in or split-sample estimators without integrating robust regularization, feature-importance–based dimension reduction, and automated model selection into a Neyman-orthogonal score with explicit bias-correction. In the absence of a fully data–driven, nonparametric double/debiased ML–IV approach, reusing the same sample across both stages can induce first-stage overfitting and bias (Wager and Athey, 2018; Chernozhukov, Newey, and Singh, 2022; Chernozhukov et al., 2022). For instance, Chen et al. (2021) implement a split-sample ML estimator but neither augment the moment condition with a bias-correction term nor verify the Gateaux-derivative condition. Similarly,

[1]In this paper, "high-dimensional" refers solely to the first-stage prediction problem (a supervised ML task); the second-stage IV regression remains low-dimensional. From a modeling perspective, the first stage is treated as fully nonparametric—every conditional expectation (e.g. $E[D_i \mid Z_i, W_i]$, $E[Y_i \mid W_i]$) is estimated by a "black-box" ML learner. In the second stage, however, I impose a low-dimensional linear index for D_i (and for the low-dimensional subset $W_{i,\mathrm{ld}}$), while the remaining variation in W_i is absorbed nonparametrically into $m(W_i)$. Thus the second stage is semiparametric.

Belloni et al. (2012) employ Lasso/post-Lasso to approximate optimal instruments in a linear IV model with many instruments but only low-dimensional controls, yet they fit a single first stage and apply two-stage least squares without orthogonal residualization or cross-fitting, and do not establish double robustness or semiparametric efficiency in high-dimensional or nonlinear contexts. Similarly, Okui et al. (2012) study a semiparametric IV model and propose doubly-robust GMM estimators that require the specification of finite-dimensional working models for both the outcome and treatment equations, showing consistency whenever either of these two working models is correctly specified. However, they assume that both the control and instrument dimensions remain fixed as $n \to \infty$ and do not employ modern ML methods to estimate those nuisance functions, whereas the present approach allows the number of controls (and instruments) to grow with the sample size and leverages fully data-driven ML estimators under Neyman orthogonality. Moreover, in most prior work a single ML algorithm is chosen in an ad hoc manner for the first-stage prediction. In addition to the aforementioned papers, Gold, Lederer, and Tao (2020) employs Lasso in the first-stage prediction.

By contrast, this paper works in the many-controls/instruments regime under an approximate-sparsity assumption and employs all possible linear and nonlinear ML algorithms to estimate the nuisance regressions. It constructs an orthogonal score that remains $\sqrt{n}$-consistent even when using high-dimensional ML estimators and delivers valid, uniformly honest confidence intervals for the IV parameter in this high-dimensional setting. In this way, it extends the doubly-robust IV approach in the existing literature beyond the finite-dimensional, parametric working-model case into the high-dimensional, ML-driven world. Moroever, the proposed estimator adopts a fully data–driven model selection strategy that fits a rich library of both linear and nonlinear ML algorithms and methods to predict the endogenous regressor, evaluates their out-of-sample performance via cross-fitting, and selects the algorithm that delivers the lowest first-stage prediction error. This adaptive procedure flexibly captures the complexity of the data, strengthens the instrument, and underpins improved bias–variance trade-offs.

Thus, the paper fills these gaps by proposing a complete, end-to-end DML-IV estimator that (i) accommodates high-dimensional and potentially nonlinear instruments and controls; (ii) embeds cross-fitting, systematic hyperparameter tuning for regularization, fully data-driven feature-importance–based dimensionality reduction, and model selection from a rich learner library into a Neyman-orthogonal estimating equation; and (iii) rigorously establishes double robustness and semiparametric efficiency. In particular, the estimator implements an adaptive construction of optimal instruments—namely, the conditional expectation of the endogenous regressor given instruments and controls—by leveraging both linear and nonlinear ML models to combine high-dimensional controls and candidate instruments, thereby enhancing predictive accuracy and instrument strength. By integrating cutting-edge ML techniques into a single orthogonal score, the DML-IV approach overcomes limitations of conventional TSLS-IV and plug-in ML-IV methods under unobserved confounding. The resulting estimator remains consistent if either the treatment or outcome nuisance model is correctly specified and achieves the semiparametric efficiency bound—preserving $\sqrt{n}$-consistency and optimal variance even under slow ML convergence rates. Monte Carlo simulations and an empirical application to estimating the return to education (Angrist and Kruger, 1991) corroborate these theoretical findings, demonstrating lower bias, reduced variance, and improved mean squared error relative to conventional TSLS-IV and plug-in ML-IV estimators, with nominal coverage across diverse settings. Under standard regularity and completeness conditions, the proposed DML-IV estimator is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient, thereby delivering reliable inference in complex, high-dimensional applications.

The remainder of the paper is structured as follows. Section 2 presents the structural model, identification strategy, and integration of ML methods. Section 3 delivers the main theoretical results, including proofs of consistency, asymptotic normality, and efficiency. Section 4 reports Monte Carlo simulation outcomes. Section 5 provides the empirical application to returns to education. Section 6 concludes with discussion and directions for future research. Notation, proofs, and additional empirical results are gathered in the online supplementary material.

2 Model and Identification Strategy

2.1 Structural Machine-Learned IV Model and Motivation

The model is given by

$$Y_i = \tau\, D_i + \beta^\top W_{i,\mathrm{ld}} + m(W_i) + \phi_i, \tag{1}$$

$$D_i = f\big(Z_i,\, W_i; \lambda, \theta\big) + \gamma\, A_i + \nu_i, \tag{2}$$

where $W_i = (W_{i,\text{ld}}, W_{i,\text{hd}})$, $\beta^\top W_{i,\text{ld}}$ is the low-dimensional linear component, and

$$m(W_i) \;=\; E\big[Y_i - \tau\, D_i - \beta^\top W_{i,\text{ld}} \;\big|\; W_i\big],$$

which nonparametrically absorbs the remaining high-dimensional controls.[2] Y_i denotes the outcome, D_i the endogenous treatment, Z_i the instruments, and A_i unobserved confounders.[3] The functions $f(\cdot;\lambda,\theta)$ incorporate adaptive ML techniques to predict D_i. The remaining structural shocks satisfy

$$E[\phi_i \mid Z_i, W_i] = 0, \quad E[\nu_i \mid Z_i, W_i] = 0.$$

Although at first glance the decomposition

$$Y_i = \tau\, D_i + \beta^\top W_{i,\text{ld}} + m(W_i) + \phi_i, \quad D_i = f(Z_i, W_i; \lambda, \theta) + \gamma\, A_i + \nu_i$$

may seem nonstandard, it in fact nests virtually every IV- or control-function model in the literature. For example, the classical linear IV model is recovered by taking $W_{i,\text{ld}} = W_i$ and $m \equiv 0$. Nonparametric IV (NPIV) arises by setting $\beta = 0$ and $W_{i,\text{ld}} = \varnothing$, so that all covariates enter through the nonparametric remainder $m(W_i)$ (Newey, 1990; Newey and Powell, 2003; Darolles et al., 2011). Control-function approaches, which posit a finite-dimensional function $h(A_i)$, are also contained here, since once A_i is measurable with respect to W_i, any finite-dimensional $h(A_i)$ can be absorbed into $m(W_i)$. Similarly, latent-factor confounder models—where $A_i \in^r$ and $E[A_i \mid W_i]$ is linear in W_i—fit within this framework by writing

$$\delta^\top A_i = \delta^\top E[A_i \mid W_i] + \big(\delta^\top A_i - \delta^\top E[A_i \mid W_i]\big),$$

absorbing the first term into $m(W_i)$ and treating the second as an orthogonal shock. Thus,

[2] Because all available covariates are used to predict D_i by ML in the first stage, here $W_i \equiv W_{i,\text{hd}}$ is the full (high-dimensional) covariate vector.

[3] Although identification ultimately rests on the single restriction $E\big[\gamma A_i + \phi_i \mid Z_i, W_i\big] = 0$, the model distinguishes A_i (the latent factor driving both treatment and outcome) from ϕ_i (the idiosyncratic shock to Y) and ν_i (the idiosyncratic shock to D) for two purposes: (a) to highlight the source of endogeneity—A_i induces correlation between D_i and the outcome error, whereas ϕ_i affects only the outcome equation; and (b) to clarify the double-robustness argument by separating a "treatment-selection" error, $\gamma A_i + \nu_i$, from a "pure outcome" error, ϕ_i, each of which is orthogonal to (Z_i, W_i) under different maintained assumptions.

by choosing $W_{i,\text{ld}}$ to be the small set of covariates one wishes to interpret linearly (e.g. age or policy dummies) and allowing all remaining controls to enter via $m(W_i)$, one obtains a unified semiparametric specification that preserves interpretability where desired and robustness everywhere else.

I do not assume any finite-dimensional parametric form for

$$q(Z_i, W_i) \;=\; E[D_i \mid Z_i, W_i] \quad \text{or} \quad r(W_i) \;=\; E[D_i \mid W_i],\; \mu(W_i) = E[Y_i \mid W_i].$$

Instead, these conditional expectations are estimated by fully data-driven, nonparametric ML learners (including Lasso, random forests, boosting, neural nets, etc.). In that sense, the entire first stage is nonparametric.

Once I have $\hat{q}(Z_i, W_i)$ and $\hat{m}(W_i) = \hat{\mu}(W_i) - \tau\, \hat{r}(W_i)$, I solve the classical moment equation

$$E\big[(Z_i - \pi(W_i))\big(Y_i - \tau D_i - m(W_i) + \tau\, q(Z_i, W_i)\big)\big] = 0. \tag{3}$$

Here only a finite-dimensional subset of controls $W_{i,\text{ld}}$ enters linearly (through $\beta^\top W_{i,\text{ld}}$), and τ itself is a single scalar. The rest of W_i enters flexibly into $m(W_i)$, but that flexibility is absorbed in a single nonparametric leftover. Consequently, the second stage is semiparametric—it mixes a low-dimensional linear index $(\tau\, D_i + \beta^\top W_{i,\text{ld}})$ with a nonparametric nuisance $m(W_i)$. In short, this paper is nonparametric when building the first-stage predictions of D_i, and semiparametric when forming the final moment condition for τ.

To identify τ, I impose the usual instrumental-variables conditions. First, instrument exogeneity requires that, conditional on the full control vector W_i, the instruments be uncorrelated with both the unobserved confounders and the structural error:

$$E[Z_i A_i \mid W_i] = 0 \quad \text{and} \quad E[Z_i \phi_i \mid W_i] = 0.$$

Second, instrument relevance demands that the instruments retain predictive power for the endogenous treatment after conditioning on the controls,

$$\text{Cov}\big(D_i,\, Z_i \mid W_i\big) \neq 0.$$

Finally, control exogeneity stipulates that the confounders and the outcome error have zero conditional mean given W_i, and that the confounders themselves are mean-zero in the population:

$$E[A_i \mid W_i] = 0, \quad E[\phi_i \mid W_i] = 0, \quad E[A_i] = 0.$$

Because all of W_i enters either linearly (via $W_{i,\mathrm{ld}}$) or nonparametrically (via $m(W_i)$), the condition $E[Z_i A_i \mid W_i] = 0$ is a genuine conditional exogeneity assumption.

A central element of the proposed estimator is the optimal instrument function

$$w(Z_i, W_i) \;=\; E\big[D_i \mid Z_i, W_i\big],$$

estimated by $\hat{w}(Z_i, W_i)$ and used to form the Neyman-orthogonal moment condition via cross-fitting, thereby ensuring robustness to small perturbations in the nuisance estimates.

2.2 Adaptive Neyman Orthogonality and Identification of τ

The paper frames estimation of the causal effect τ via a Neyman–orthogonal moment that accommodates flexible, data-driven ML for the nuisance components. I begin by introducing a hierarchy of parameters. The primary object of interest is the scalar parameter τ, which measures the causal effect. The secondary nuisance functions are collected into

$$\eta = \big(\pi,\, q,\, m\big),$$

where

$$\pi(W_i) = E[Z_i \,|\, W_i], \quad q(Z_i, W_i) = E[D_i \,|\, Z_i, W_i], \quad m(W_i) = E[Y_i - \tau D_i - \beta^\top W_{i,\mathrm{ld}} \mid W_i]. \quad (4)$$

Because $m(W_i) = E[Y_i - \tau D_i \mid W_i]$ depends on τ, I re-express it in terms of two regressions that do not require prior knowledge of τ. Define

$$\mu(W_i) \;=\; E[Y_i \mid W_i], \qquad r(W_i) \;=\; E[D_i \mid W_i]. \quad (5)$$

It follows that

$$m(W_i) = E\big[Y_i - \tau\, D_i \mid W_i\big] = \mu(W_i) \;-\; \tau\, r(W_i),$$

so in practice I cross-fit one ML regression $\hat{\mu}$ of Y_i on W_i and one $\hat{r}$ of D_i on W_i, both independent of τ, and then set

$$\hat{m}(W_i) \;=\; \hat{\mu}(W_i) \;-\; \tau\,\hat{r}(W_i).$$

This construction avoids circularity in estimating m.

Tertiary parameters $\varphi \in \Phi$ govern aspects of the fully data-driven nonparametric ML procedures such as regularization, hyperparameter tuning, feature reduction, and model selection. Here $\varphi = (\lambda, \theta)$ collects *all* of the first-stage adaptive ML techniques , so that the Gateaux derivative

$$\frac{\partial}{\partial \varphi} E[\psi_i(\tau, \eta, \varphi)]\Big|_{(\eta_0, \varphi_0)} = 0 \quad \text{(see Appendix **Appendix C**)},$$

fully accounts for the out-of-sample selection of $\hat{\lambda}, \hat{\theta}$. In other words, the φ–orthogonality condition automatically "knows about" the entire tuning procedure.

Because f may be drawn from any ML class—penalized linear models, trees, boosting, or deep nets—the theory never imposes a parametric form. It suffices to assume the following convergence rates for all four ML nuisances:

$$\|\hat{\pi} - \pi\|_{L^2},\; \|\hat{q} - q\|_{L^2},\; \|\hat{\mu} - \mu\|_{L^2},\; \|\hat{r} - r\|_{L^2} \;=\; o_p(n^{-1/4}).$$

Under cross-fitting, these rates guarantee that first-order perturbations in either $\eta = (\pi, q, m)$ or in the hyperparameter vector $\varphi = (\lambda, \theta)$ enter the estimating equation only at order $o_p(n^{-1/2})$. In particular, the Gateaux derivatives

$$\frac{\partial}{\partial \eta} E[\psi_i(\tau, \eta, \varphi)]\Big|_{(\eta_0, \varphi_0)} \;=\; \frac{\partial}{\partial \varphi} E[\psi_i(\tau, \eta, \varphi)]\Big|_{(\eta_0, \varphi_0)} \;=\; 0 \quad \text{(see Appendix **Appendix C**)},$$

and the mixed second derivative remains bounded (see Appendix **Appendix C**). Consequently the resulting DML–IV estimator is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient even when machine-learning convergence is slow.

In addition, following (Belloni et al., 2012; Hansen, Hausman, and Newey, 2008; Newey, 1990), the optimal instrument function is defined as

$$w(Z_i, W_i) \;=\; E\big[D_i \mid Z_i, W_i\big], \tag{6}$$

which plays a key role in achieving semiparametric efficiency. This function coincides with the first-stage conditional expectation, that is,

$$q(Z_i, W_i) = E\big[D_i \mid Z_i, W_i\big] = w(Z_i, W_i).$$

In practice, this function is estimated from the data—denoted by $\hat{w}(Z_i, W_i)$—and is used to form the plug-in prediction $\hat{D}_i$ in the first stage.

Under the conditional moment restriction $E[\phi_i \mid Z_i, W_i] = 0$ in the structural model, it is a standard result that the efficient instrument in the linear IV setting is

$$w^*(Z_i, W_i) \;=\; E[D_i \mid Z_i, W_i].$$

This follows from Chamberlain's semiparametric efficiency analysis (Chamberlain, 1987, Theorem 1) and its extension to series approximations by Newey (1990). In high-dimensional settings, Belloni et al. (2012) further show that Lasso-based approximations to

$$E[D_i \mid Z_i, W_i]$$

retain $\sqrt{n}$-consistency and asymptotic normality. In this framework I therefore implement the optimal instrument by targeting

$$q(Z_i, W_i) = E[D_i \mid Z_i, W_i]$$

in the orthogonal score (7).

Based on these definitions, the classical orthogonal moment condition is specified as

$$\psi_i\big(\tau, \eta, \varphi\big) \;=\; \big(Z_i - \pi(W_i)\big)\,\Big[Y_i - \tau\, D_i - m(W_i) + \tau\, q(Z_i, W_i)\Big], \tag{7}$$

where the term $(Z_i - \pi(W_i))$ represents the instrument residual, thereby isolating the exogenous variation in Z_i, and the bracketed term isolates τ in expectation.

Here $q(Z, W)$ is chosen to satisfy the conditional moment

$$E\big[(Z - \pi(W))\,(D - q(Z, W)) \mid W\big] = 0.$$

By iterated expectations one then shows

$$E\big[(Z-\pi(W))\,q(Z,W)\big] = E\big[(Z-\pi(W))\,D\big], \quad E\big[(Z-\pi(W))(Y-m(W))\big] = \tau\,E\big[(Z-\pi(W))\,D\big].$$

Hence the two τ–terms cancel exactly and

$$E\Big[(Z-\pi(W))(Y-\tau D-m(W)+\tau q(Z,W))\Big] = 0.$$

This shows that the extra $\tau\,q$–term is precisely the bias–correction needed to ensure the moment is zero.

Under the maintained IV assumptions,

$$E\Big[\psi_i(\tau,\eta,\varphi)\Big] = 0.$$

The optimized ML parameters φ enter the framework indirectly through the nuisance functions, since the functions π, q, and m are estimated using ML models that involve hyperparameter tuning and feature selection to address high-dimensionality and potential overfitting.

The function

$$f(Z_i, W_i; \lambda, \theta)$$

from (2) is nothing but a fully data-driven estimate of the conditional expectation

$$q(Z_i, W_i) = E[D_i \mid Z_i, W_i].$$

I distinguish between two sets of hyperparameters: λ, which includes all regularization penalties (e.g., the Lasso penalty, ridge shrinkage, or penalties on tree depth), and θ, which encompasses all other tuning choices (e.g., learning rate, number of trees, network architecture) as well as any feature-reduction or model-selection decisions made by the fully data-driven ML pipeline.

Within each cross-fitting fold I select $\hat{\lambda}, \hat{\theta}$ by out-of-sample validation and then set

$$\hat{q}(Z_i, W_i) \;\equiv\; f\big(Z_i, W_i; \hat{\lambda}, \hat{\theta}\big) \;\approx\; E[D_i \mid Z_i, W_i].$$

Accordingly, in the orthogonal score

$$\psi_i(\tau, \eta, \varphi) = \big(Z_i - \pi(W_i)\big)\Big[Y_i - \tau D_i - m(W_i) + \tau\, q(Z_i, W_i)\Big],$$

the nuisance function $q(Z_i, W_i)$ is implemented by the first-stage predictor $f(\cdot; \hat{\lambda}, \hat{\theta})$. This makes explicit both the role of f in approximating the optimal instrument and the distinction between tertiary parameters λ and θ.

Neyman orthogonality is an intrinsic property of $\psi_i(\tau, \eta, \varphi)$, namely that

$$\frac{\partial}{\partial \eta} E\big[\psi_i(\tau, \eta, \varphi)\big]\Big|_{(\eta_0, \varphi_0)} = 0, \quad \frac{\partial}{\partial \varphi} E\big[\psi_i(\tau, \eta, \varphi)\big]\Big|_{(\eta_0, \varphi_0)} = 0 \quad \text{(see Appendix } \textbf{Appendix C}\text{)}.$$

Cross-fitting and related sample-splitting schemes are then employed as estimation strategies to ensure that the sample analog of this property holds up to $o_p(n^{-1/2})$, by orthogonalizing the first-stage estimation errors from the second-stage moment evaluation. Moreover, any potential second-order interactions between errors in η and φ are controlled by requiring that the mixed partial derivative

$$\frac{\partial^2}{\partial \varphi\, \partial \eta} E\big[\psi_i(\tau, \eta, \varphi)\big]\Big|_{(\eta_0, \varphi_0)} \quad \text{remains bounded (see Appendix } \textbf{Appendix C}\text{)}.$$

Under these regularity conditions, the estimator for τ is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient when advanced ML techniques are employed. This adaptive setting that integrates both the primary nuisance functions and the advanced ML parameters ensures robustness against overfitting and model misspecification in high-dimensional settings.

Theorem 1 (Identification of τ). Under the adaptive structural model and given assumptions above, the causal effect τ is uniquely identified by the orthogonal moment condition

$$E\Big[\psi_i(\tau, \eta, \varphi) \mid W_i\Big] = 0,$$

where ψ_i is defined in (7).

2.3 Estimation Strategy

2.3.1 Cross-Fitting with Hierarchical Parameters η, φ

To obtain a consistent and robust estimator of τ in the presence of high-dimensional nuisance parameters and advanced ML techniques, I employ a cross-fitting procedure tailored to the hierarchical structure of nuisance parameters η and tertiary parameters φ.

First, partition the full sample $\{1, \ldots, n\}$ into K mutually exclusive and collectively exhaustive subsets (folds) $\mathcal{I}_1, \ldots, \mathcal{I}_K$ of approximately equal size.

For each fold $k \in \{1, \ldots, K\}$, using the data excluding fold k, that is, from the combined set $\mathcal{I}_{-k} = \bigcup_{j \neq k} \mathcal{I}_j$, estimate the secondary nuisance functions $\hat{\eta}^{(-k)} = (\hat{\pi}^{(-k)}, \hat{q}^{(-k)}, \hat{m}^{(-k)})$ along with the corresponding tertiary parameter selections denoted by $\hat{\varphi}^{(-k)}$. Note that these estimations involve advanced ML methods such as regularization, cross-fitting, model selection, and sparsity reduction based on feature importance scores from the selected best prediction algorithms with hyperparameter tuning. For each observation $i \in \mathcal{I}_k$, compute predictions $\hat{\eta}^{(-k)}(W_i)$ and store the associated hyperparameters $\hat{\varphi}^{(-k)}$.

Second, for every observation $i = 1, \ldots, n$, assign the estimates $\hat{\eta}_i$ and $\hat{\varphi}_i$ obtained from the fold where i was held-out, effectively combining the fold-specific predictions into a full-sample set of nuisance estimates.

Lastly, employ a similar cross-fitting approach to estimate the optimal instrument function $w(Z_i, W_i) = E\big[D_i \mid Z_i, W_i\big]$. For each fold k, compute $\hat{w}^{(-k)}(Z_i, W_i)$ for $i \in \mathcal{I}_k$ using models trained on $\mathcal{I}_{-k}$, and aggregate these to obtain $\hat{w}(Z_i, W_i)$ for the whole sample.

This cross-fitting procedure ensures first-order orthogonality because each fold's nuisance estimates $\hat{\eta}_i$ and tertiary parameter choices $\hat{\varphi}_i$ are obtained from data not used in evaluating the moment conditions at observation i; small estimation errors in these nuisance components do not induce first-order bias in the estimator of τ. Moreover, performing hyperparameter selection out-of-sample in each fold mitigates overfitting and supports approximate second-order orthogonality with respect to φ.

For each fold $k = 1, \ldots, K$, on the training set $\mathcal{I}_{-k}$ estimate

$$\hat{\mu}^{(-k)}(W_i) \approx E[Y_i \mid W_i], \qquad \hat{r}^{(-k)}(W_i) \approx E[D_i \mid W_i].$$

Then for each held-out $i \in \mathcal{I}_k$ set

$$\hat{m}(W_i) \;=\; \hat{\mu}^{(-k)}(W_i) \;-\; \tau\, \hat{r}^{(-k)}(W_i).$$

By construction no initial value of τ is needed: these $\hat{m}$ simply enter the orthogonal score

$$\psi_i(\tau) = \big(Z_i - \hat{\pi}(W_i)\big)\Big[Y_i - \tau D_i - \hat{m}(W_i) + \tau\, \hat{q}(Z_i, W_i)\Big],$$

and one shows algebraically that the solution is

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \hat{q}(Z_i, W_i)\, Y_i}{\sum_{i=1}^{n} \hat{q}(Z_i, W_i)\, D_i},$$

so there is never any circularity in estimating m before τ.

2.3.2 Incorporating Advanced ML Techniques

Following (Chernozhukov, Newey, and Singh, 2022; Sun, Cui, and Tchetgen, 2022), the estimator integrates modern ML methods into the estimation of both secondary nuisance functions $\eta = (\pi, q, m)$ and tertiary parameters φ. Concretely, within each fold k I train a suite of candidate pipelines (indexed by $\varphi^{(-k)}$) on the training set $\mathcal{I}_{-k}$. These pipelines may include regularized linear methods (e.g. Lasso, Ridge) parameterized by penalty weights λ, and nonlinear learners (e.g. Random Forests, boosting, neural nets) parameterized by hyperparameters θ (e.g. tree depth, learning rate, network architecture). Feature reduction and model selection are driven by those regularization and hyperparameter-tuning processes, denoted λ and θ.

After fitting each pipeline on $\mathcal{I}_{-k}$, its out-of-sample prediction error is recorded on the held-out fold $\mathcal{I}_k$. The pipeline with the lowest validation error (and its associated $\hat{\varphi}^{(-k)}$) is selected as the "best" for that fold.

Once a pipeline is chosen, I refit it on $\mathcal{I}_{-k}$ under the selected $\hat{\varphi}^{(-k)}$ to produce

$$\hat{\pi}^{(-k)}, \quad \hat{q}^{(-k)}, \quad \hat{\mu}^{(-k)}, \quad \hat{r}^{(-k)},$$

and hence $\hat{m}^{(-k)} = \hat{\mu}^{(-k)} - \tau\, \hat{r}^{(-k)}$. For nonlinear learners, I then compute feature-importance scores (e.g. variable importance in Random Forests or SHAP values in boosting) and prune low-importance covariates from $\{\pi, q, \mu, r\}$. In linear models, dimensionality is already reduced by the ℓ_1 (Lasso) or ℓ_2 (Ridge) penalty, so no further pruning is required. The result is a refined set of predictors for each nuisance function.

For each held-out index $i \in \mathcal{I}_k$, I set

$$\hat{\eta}_i = \big(\hat{\pi}^{(-k)}(W_i),\, \hat{q}^{(-k)}(Z_i, W_i),\, \hat{m}^{(-k)}(W_i)\big), \qquad \hat{\varphi}_i = \hat{\varphi}^{(-k)}.$$

Because every $\hat{\eta}_i$ and $\hat{\varphi}_i$ is computed on data that excludes the ith observation, the first-order effect of any regularization bias or hyperparameter choice vanishes from the second-stage estimating equation. In particular, first-order Neyman orthogonality implies that small perturbations in $\hat{\eta}$ or $\hat{\varphi}$ do not bias $\hat{\tau}$ at order $n^{-1/2}$ (Chernozhukov et al., 2018). Any higher-order interaction between η and φ estimation errors is $o_p(n^{-1/2})$ under the bounded-Hessian and rate conditions detailed in Appendix **Appendix C**.

In summary, each fold k proceeds as follows: (i) fit multiple candidate pipelines—regularized linear (varying λ) and nonlinear (varying θ)—on $\mathcal{I}_{-k}$, including any feature-importance–driven pruning or dimensionality-reduction steps as part of model selection; (ii) select the pipeline with lowest validation error on $\mathcal{I}_k$, recording its hyperparameters, feature-pruning rules, and any model-selection choices as $\hat{\varphi}^{(-k)}$; (iii) refit the chosen pipeline on $\mathcal{I}_{-k}$ under $\hat{\varphi}^{(-k)}$ to obtain $\hat{\pi}^{(-k)}, \hat{q}^{(-k)}, \hat{\mu}^{(-k)}, \hat{r}^{(-k)}$, applying the recorded feature-reduction rules, and compute $\hat{m}^{(-k)} = \hat{\mu}^{(-k)} - \tau\, \hat{r}^{(-k)}$; (iv) for each $i \in \mathcal{I}_k$, set $\hat{\eta}_i = \big(\hat{\pi}^{(-k)}(W_i),\, \hat{q}^{(-k)}(Z_i, W_i),\, \hat{m}^{(-k)}(W_i)\big)$ and $\hat{\varphi}_i = \hat{\varphi}^{(-k)}$. Because cross-fitting isolates each $\hat{\eta}_i$ and $\hat{\varphi}_i$ from (Y_i, D_i, Z_i, W_i), any bias introduced by regularization or hyperparameter selection is orthogonal to the moment condition for τ. As a result, the DML-IV estimator remains $\sqrt{n}$-consistent and asymptotically normal despite using complex, high-dimensional ML pipelines for nuisance estimation.

2.4 Optimal Instruments and Regularity Conditions

Optimal instruments are defined as the conditional expectation of the endogenous variable given both the instruments and the high-dimensional controls. Building on the foundational work of Amemiya (1974), Chamberlain (1987), and Newey (1990), Belloni et al. (2012) demonstrated that optimal instruments can be effectively approximated via sparse methods such as Lasso and post-Lasso. Other studies have contributed alternative strategies. For instance, Bai and Ng (2009) employed boosting-based instrument selection, Caner (2009) introduced a Lasso-type GMM estimator, Carrasco (2012) proposed a ridge regression approach to handle many instruments, and Gautier and Tsybakov (2011) developed a square-root Lasso method for high-dimensional IV estimation. In contrast to these approaches, which typically rely on a single regularization strategy, I extend this framework by integrating a broader array of ML techniques, including cross-fitting, advanced regularization, hyperparameter tuning, data-driven feature selection, and the selection of the best prediction model in the first stage from among all flexible ML algorithms based on a fully data-driven, nonparametric approach, to approximate $w(Z_i, W_i)$ adaptively. This adaptive construction not only preserves the flexibility of the nonparametric projection but also ensures robustness and semiparametric efficiency in high-dimensional settings, even when some instruments are weak.

As defined in (6), the optimal instrument function is

$$w(Z_i, W_i) \;=\; E\big[D_i \mid Z_i, W_i\big].$$

This definition is nonparametric, imposing no fixed linear or nonlinear functional form on the relationship between the instruments Z_i and the controls W_i. Instead, $w(Z_i, W_i)$ flexibly characterizes how the endogenous variable D_i can be optimally predicted from the combined information in Z_i and W_i using data-driven methods. In effect, $w(Z_i, W_i)$ serves as the projection of D_i onto the sigma-algebra generated by (Z_i, W_i), thereby minimizing the mean squared error $E[\|D_i - w(Z_i, W_i)\|^2]$. This property ensures that $w(Z_i, W_i)$ captures the most informative statistical patterns that predict D_i, accommodating both linear and nonlinear relationships as revealed by the data.

Within the adaptive DML-IV estimator in this paper, the function $w(Z_i, W_i)$ underpins the estimation of nuisance parameters and the construction of efficient instruments. By accurately summarizing the relationship between Z_i, W_i, and D_i, $w(Z_i, W_i)$ aids in reducing residual variation when predicting the endogenous variable. In practice, $\hat{w}(Z_i, W_i)$ is estimated using advanced ML algorithms that adapt to the underlying data structure—capturing linearities, nonlinearities, and interactions—through regularization, hyperparameter tuning, and cross-fitting. This adaptive estimation ensures that the final instrument remains robust, efficient, and orthogonal to the second-stage estimation errors.

In this setting, the optimal instrument $w(Z_i, W_i)$ indirectly informs the estimation of nuisance parameters $\pi(W_i)$, $q(Z_i, W_i)$, and $m(W_i)$ through the moment function:

$$\psi_i\big(\tau, \eta, \varphi\big) \;=\; \big(Z_i - \pi(W_i)\big)\,\Big[Y_i - \tau\, D_i - m(W_i) + \tau\, q(Z_i, W_i)\Big].$$

Stacking this primary moment condition with additional valid restrictions yields a vector of moment conditions:

$$\Psi_n(\tau) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} (Z_i - \pi(W_i))(Y_i - \tau D_i - m(W_i) + \tau q(Z_i, W_i)) \\ \psi_i^{(2)}(\tau, \eta, \varphi) \\ \vdots \\ \psi_i^{(J)}(\tau, \eta, \varphi) \end{pmatrix} = 0. \tag{8}$$

where subsequent components $\psi_i^{(j)}(\tau, \eta, \varphi)$ may incorporate overidentifying restrictions or additional moments informed by $w(Z_i, W_i)$. The accurate estimation of $w(Z_i, W_i)$ enhances the precision and efficiency of these nuisance estimators and, consequently, the IV estimator $\hat{\tau}$. Regularity conditions are as follows:

RC1: The covariates W_i, instruments Z_i, and outcome Y_i are bounded or sub-Gaussian:

$$\|W_i\|, \|Z_i\|, \|Y_i\| < C < \infty,$$

and the error terms are exogenous:

$$E[A_i \mid W_i] = 0, \quad E[\phi_i \mid W_i] = 0.$$

This condition controls the variability of the data and ensures that W_i does not introduce bias via correlation with unobservables, which is critical for the reliability of the projection $w(Z_i, W_i)$ and the stability of the nuisance parameter estimates.

RC2: The instruments Z_i possess sufficient predictive power for D_i given W_i:

$$\mathrm{Var}\Big(E[D_i \mid Z_i, W_i]\Big) \;\geq\; c > 0.$$

Strong instruments guarantee that the relationship between Z_i and D_i is informative, enabling the optimal instrument $w(Z_i, W_i)$ to effectively summarize the variation in W_i that is relevant for predicting D_i, which is key to achieving efficient estimation of τ.

RC3: The instruments are valid:

$$E[Z_i \, \phi_i] \;=\; 0, \quad E[Z_i \, A_i] \;=\; 0.$$

Exogeneity of instruments ensures that the projections and moment conditions are not contaminated by omitted variable bias or correlation with error terms, preserving the integrity of $w(Z_i, W_i)$ and the identification strategy for τ.

RC4: The nuisance functions $\pi(W_i)$, $q(Z_i, W_i)$, and $m(W_i)$ are Lipschitz continuous:

$$|\pi(W_i + \Delta) - \pi(W_i)| \leq L\|\Delta\|,$$

and similarly for q and m. The optimal instrument function satisfies

$$E\Big[w(Z_i, W_i)^2\Big] < \infty.$$

Square-integrability ensures that $w(Z_i, W_i)$ has finite variance, which is essential for valid asymptotic inference and for guaranteeing that the estimation process yields stable and reliable instruments.

RC5: The nuisance estimators $\hat{\pi}$, $\hat{q}$, $\hat{m}$, and $\hat{w}$ converge at rates:

$$\|\hat{\pi} - \pi\|_{L^2}, \; \|\hat{q} - q\|_{L^2}, \; \|\hat{m} - m\|_{L^2}, \; \|\hat{w} - w\|_{L^2} = o_p(n^{-1/4}).$$

These convergence rates ensure that the estimation errors for the nuisance functions and the optimal instrument shrink quickly enough so that their impact on the estimation of τ diminishes, preserving the efficiency and consistency of the IV estimator. In other words, this condition ensures that the estimation errors for the nuisance functions and the optimal instrument diminish sufficiently fast.

RC6: The moment function satisfies first-order orthogonality:

$$\frac{\partial}{\partial \eta} E\big[\psi_i(\tau, \eta, \varphi)\big]\Big|_{\eta=\eta_0,\, \varphi=\varphi_0} \;=\; 0.$$

Additionally, for advanced ML parameters φ, I require

$$\frac{\partial}{\partial \varphi} E\big[\psi_i(\tau, \eta, \varphi)\big]\Big|_{\eta=\eta_0,\, \varphi=\varphi_0} \approx 0,$$

and that the mixed second-order derivative

$$\frac{\partial^2}{\partial \varphi\, \partial \eta} E\big[\psi_i(\tau, \eta, \varphi)\big]\Big|_{\eta=\eta_0,\, \varphi=\varphi_0} \text{ is bounded,}$$

thereby controlling the influence of advanced ML parameters on the moment conditions.

First-order orthogonality ensures that small deviations in nuisance parameters do not introduce bias into the estimation of τ.

In (7), the moment function is defined as:

$$\psi_i(\tau, \eta, \varphi) = (Z_i - \pi(W_i))\big[Y_i - \tau D_i - m(W_i) + \tau q(Z_i, W_i)\big].$$

Let $\eta = (\pi, q, m)$ and consider a small perturbation $\Delta\eta = (\Delta\pi, \Delta q, \Delta m)$. The perturbed moment function becomes:

$$\begin{aligned} \psi_i(\tau, \eta + \Delta\eta) = \psi_i(\tau, \eta) & \\ - \Delta\pi(W_i)\, \omega_i & \\ - \big(Z_i - \pi(W_i)\big) \big(\Delta m(W_i) - \tau\, \Delta q(Z_i, W_i)\big). & \end{aligned} \tag{9}$$

where $\omega_i = Y_i - \tau D_i - m(W_i) + \tau q(Z_i, W_i)$.

By (RC1), W_i, Z_i, and Y_i are bounded, so $\|\omega_i\|$ is finite. Using (RC5), the perturbations in nuisance functions satisfy:

$$\|\Delta\pi\|, \|\Delta q\|, \|\Delta m\| = o_p(n^{-1/4}).$$

These bounds ensure that the impact of estimation errors in η on ψ_i diminishes as $n \to \infty$.

Taking expectations conditional on W_i:

$$\begin{aligned} E\big[\psi_i(\tau, \eta + \Delta\eta) \mid W_i\big] = E\big[\psi_i(\tau, \eta) \mid W_i\big] & \\ - E\big[\Delta\pi(W_i)\, \omega_i \mid W_i\big] & \\ - E\Big[(Z_i - \pi(W_i)) \big(\Delta m(W_i) - \tau\, \Delta q(Z_i, W_i)\big) \,\Big|\, W_i\Big]. & \end{aligned} \tag{10}$$

By (RC6):

$$E[\psi_i(\tau, \eta) \mid W_i] = 0.$$

By the cross-fitting procedure, $\Delta\pi$, Δq, and Δm are orthogonal to the residual terms ω_i and $Z_i - \pi(W_i)$:

$$E[\Delta\pi(W_i) \cdot \omega_i \mid W_i] = 0, \quad E[(Z_i - \pi(W_i)) \cdot (\Delta m - \tau \Delta q) \mid W_i] = 0.$$

Thus:

$$E[\psi_i(\tau, \eta + \Delta\eta) \mid W_i] = 0.$$

Second-order orthogonality ensures that interactions between errors in nuisance parameter estimation (η) and hyperparameter selection (φ) are asymptotically negligible.

Expanding $E[\psi_i(\tau, \eta, \varphi)]$ around (η_0, φ_0):

$$\begin{aligned} E\big[\psi_i(\tau, \eta_0 + \Delta\eta, \varphi_0 + \Delta\varphi)\big] = E\big[\psi_i(\tau, \eta_0, \varphi_0)\big] \\ + (\Delta\eta)^\top \frac{\partial E[\psi_i]}{\partial \eta} \ + \ (\Delta\varphi)^\top \frac{\partial E[\psi_i]}{\partial \varphi} \\ + \text{higher-order terms.} \end{aligned} \tag{11}$$

By (RC6), first-order terms vanish:

$$\frac{\partial E[\psi_i]}{\partial \eta} = 0, \quad \frac{\partial E[\psi_i]}{\partial \varphi} = 0.$$

By (RC4) and (SO1):

$$\left\| \frac{\partial^2 E[\psi_i]}{\partial (\eta, \varphi)^2} \right\| \leq C < \infty.$$

By (RC5), the errors in η and φ satisfy:

$$\|\Delta\eta\| = O_p(n^{-1/4}), \quad \|\Delta\varphi\| = O_p(n^{-1/4}),$$

ensuring:

$$\|\Delta\eta\| \cdot \|\Delta\varphi\| = o_p(n^{-1/2}).$$

Under (RC1)–(RC6), first-order orthogonality holds, and second-order mixed terms are bounded by $o_p(n^{-1/2})$, ensuring Neyman orthogonality and robustness to advanced ML methods. These conditions ensure that the moment function is robust to small perturbations in both the nuisance parameters and advanced ML hyperparameters. The nullified first-order derivatives imply that minor errors in estimating η and setting φ do not bias $\hat{\tau}$, while controlling the mixed second-order derivatives prevents interactions between errors in η and φ from inducing substantial bias, thereby safeguarding the estimator's asymptotic properties.

3 Main Results

3.1 Neyman Orthogonality

Neyman Orthogonality ensures that the estimator is insensitive to small perturbations in the nuisance parameters, thereby reducing bias and enhancing robustness. This property removes first-order bias in regression-based parameters, making estimators robust to model misspecification (Neyman, 1959, 1979; Belloni et al., 2017; Chang, 2020; Chernozhukov et al., 2018, 2022). To analyze the robustness of the estimator in the presence of advanced ML parameters, both first-order and second-order orthogonality results are established with respect to the nuisance parameters η and φ.

3.1.1 First-Order Orthogonality

Lemma 1 (Gateaux Orthogonality). Let $\psi_i(\tau, \eta, \varphi)$ be as in (7). Define for any direction $h = (h_\pi, h_q, h_m)$ the Gateaux derivative

$$D_\eta \, E[\psi_i(\tau, \eta, \varphi)]\big[h\big] \;=\; \frac{d}{d\varepsilon} \, E[\psi_i(\tau, \eta + \varepsilon \, h, \varphi)]\Big|_{\varepsilon=0}.$$

Under regularity conditions (RC1–RC6) and the definition of q and m in (2), I have

$$D_\eta \, E[\psi_i(\tau, \eta_0, \varphi_0)]\big[h\big] \;=\; 0 \quad \text{for every nuisance-direction } h.$$

This result implies that small perturbations in the nuisance parameters η do not affect the expected value of the moment function to first order in Lemma 1. Consequently, the moment condition $\psi_i(\tau, \eta)$ satisfies Neyman Orthogonality, ensuring robustness of the estimator $\hat{\tau}$

against first-order estimation errors in the nuisance parameters. This is particularly crucial in high-dimensional settings, where cross-fitting and regularization techniques mitigate overfitting and exploit sparsity. Cross-fitting ensures that the data used to estimate nuisance parameters is separate from that used in the moment condition evaluation, which orthogonalizes the estimation errors.

For the second-order analysis, the following additional assumptions are introduced.

SO1: There exists a constant $C > 0$ such that for all (η, φ) in a neighborhood of the true values (η_0, φ_0),

$$\left\| \frac{\partial^2 E[\psi_i(\tau, \eta, \varphi)]}{\partial(\eta, \varphi)^2} \right\| \leq C.$$

This assumption, common in semiparametric analysis (e.g., Newey 1990), ensures that the curvature of the moment function is controlled.

SO2: The estimation errors for the nuisance parameters and the advanced ML parameters satisfy

$$\|\hat{\eta} - \eta_0\| = O_p(n^{-1/4}) \quad \text{and} \quad \|\hat{\varphi} - \varphi_0\| = O_p(n^{-1/4}),$$

ensuring that the errors decay sufficiently fast.

SO3: Any higher-order remainder terms in the Taylor expansion of $E[\psi_i(\tau, \eta, \varphi)]$ around (η_0, φ_0) are $o_p(n^{-1/2})$, so that the second-order expansion captures the dominant behavior of the remainder.

3.1.2 Second-Order Orthogonality

Lemma 2 (Second-Order Gateaux Orthogonality)*.* Define for any pair of directions $(h, \tilde{h})$, with $h = (h_\eta, h_\varphi)$ and similarly for $\tilde{h}$, the mixed Gateaux derivative

$$D^2_{(\eta,\varphi)} E[\psi_i(\tau, \eta, \varphi)]\big[(h, \tilde{h})\big] := \frac{\partial^2}{\partial \varepsilon\, \partial \delta}\, E\big[\psi_i(\tau, \eta_0 + \varepsilon h_\eta, \varphi_0 + \delta h_\varphi)\big]\bigg|_{\varepsilon=\delta=0}.$$

Under RC1–RC6 and the bounded-Hessian condition SO1,

$$\big|D^2_{(\eta,\varphi)} E[\psi_i(\tau, \eta_0, \varphi_0)]\big[(h, \tilde{h})\big]\big| \;\leq\; C\, \|h\|\, \|\tilde{h}\|.$$

If moreover $\|\hat{\eta} - \eta_0\| = O_p(n^{-1/4})$ and $\|\hat{\varphi} - \varphi_0\| = O_p(n^{-1/4})$ (SO2), then

$$D^2_{(\eta,\varphi)} E[\psi_i(\tau, \hat{\eta}, \hat{\varphi})]\big[(\hat{\eta} - \eta_0, \hat{\varphi} - \varphi_0)\big] = o_p(n^{-1/2}).$$

By Lemmas 1 and 2, the moment function is exactly orthogonal to first-order perturbations in (η, φ) and has only negligible second-order sensitivity $o_p(n^{-1/2})$. This fully establishes Neyman orthogonality in Gateaux form and underpins the $\sqrt{n}$-consistency, asymptotic normality, and semiparametric efficiency of the Adaptive DML-IV estimator.

Under the maintained assumptions and regularity conditions, first-order derivatives vanish and second-order mixed derivatives are bounded, such that any bias introduced by simultaneous small errors in η and φ is $o_p(n^{-1/2})$. This preservation of $\sqrt{n}$-consistency, asymptotic normality, and semiparametric efficiency of the Adaptive DML-IV estimator holds even when employing advanced ML methods.

Theorem 2 (Neyman Orthogonality of the ML-IV Estimator). The ML-IV estimator $\hat{\tau}_{\text{ML-IV}}$ is defined by

$$\hat{\tau}_{\text{ML-IV}} \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{w}(Z_i, W_i)\, Y_i}{\frac{1}{n}\sum_{i=1}^{n} \hat{w}(Z_i, W_i)\, D_i}. \tag{12}$$

satisfies Neyman Orthogonality with respect to the nuisance parameters $\eta = (\pi, q, m)$ when the optimal instrument $\hat{w}(Z_i, W_i) = E\big[D_i \mid Z_i, W_i\big]$ is used and cross-fitting along with regularization techniques are employed in the estimation of η. The construction of moment function here is based on (Chernozhukov et al., 2018, 2022).

Remark 1 (Equivalence to the classical IV equation). Although I have derived the orthogonal DML-IV moment in (7), in fact it collapses algebraically to the usual IV condition. This shows that the more elaborate-looking moment in this paper is exactly equivalent to solving $\frac{1}{n}\sum_i \hat{w}_i\,(Y_i - \tau D_i) = 0$ with $\hat{w}_i = [D_i \mid Z_i, W_i]$.

Substitute

$$\pi(W_i) = E[Z_i \mid W_i], \quad q(Z_i, W_i) = E[D_i \mid Z_i, W_i], \quad m(W_i) = E\big[Y_i - \tau D_i - \beta^\top W_{i,\text{ld}} \;\big|\; W_i\big],$$

into

$$0 = E\Big[\big(Z_i - \pi(W_i)\big)\big(Y_i - \tau D_i - m(W_i) + \tau\, q(Z_i, W_i)\big)\Big].$$

By iterated expectations,

$$E\big[(Z_i - \pi(W_i))(Y_i - \tau D_i - m(W_i))\big] = E\big[(Z_i - \pi(W_i))\,\tau D_i\big],$$

and

$$E\big[(Z_i - \pi(W_i))\,q(Z_i, W_i)\big] = E\big[(Z_i - \pi(W_i))\,D_i\big].$$

Hence the two τ–terms cancel, leaving

$$E\big[(Z_i - \pi(W_i))(Y_i - \tau D_i)\big] = 0,$$

which, using $\pi(W_i) = E[Z_i \mid W_i]$, is equivalent to

$$E\big[w(Z_i, W_i)\,(Y_i - \tau D_i)\big] = 0, \quad w(Z_i, W_i) = E[D_i \mid Z_i, W_i].$$

Thus the DML-IV estimator—defined by solving $\frac{1}{n}\sum_i \hat{w}_i\,(Y_i - \tau D_i) = 0$—is *identical* to the orthogonal moment solution.

Lemma 3 (Adaptive Second-Order Orthogonality)*.* Under assumptions (SO1)–(SO3), the mixed partial derivative

$$\frac{\partial^2}{\partial\varphi\,\partial\eta}\,E\big[\psi_i(\tau, \eta, \varphi)\big]\Big|_{(\eta_0, \varphi_0)}$$

is bounded, and the product $\|\Delta\eta\|\|\Delta\varphi\| = o_p(n^{-1/2})$. Consequently, the second-order terms in the Taylor expansion of $E[\psi_i(\tau, \eta, \varphi)]$ around (η_0, φ_0) are $o_p(n^{-1/2})$.

The expansion demonstrates that, by RC6 and assumptions (SO1)–(SO3), both first-order and second-order contributions of estimation errors in nuisance parameters η and advanced ML parameters φ to the moment function are negligible. This robust control through cross-fitting, regularization, and sparsity ensures Neyman orthogonality, effectively addressing overfitting and sparsity issues in high-dimensional adaptive ML-IV models.

3.2 Double Robustness

Double Robustness ensures that the estimator for τ remains consistent if at least one of the two nuisance models—either the treatment model $q(Z_i, W_i)$ or the outcome model $m(W_i)$—is correctly specified, even if the other is misspecified (Athey, Imbens and Wager, 2018; Belloni et al., 2017; Chernozhukov et al., 2018).

I now show that the ML–IV estimator remains consistent if *either* the first-stage nuisance function

$$q(Z_i, W_i) = E[D_i \mid Z_i, W_i]$$

or the second-stage nuisance function

$$m(W_i) = E\big[Y_i - \tau D_i - \beta^\top W_{i,\text{ld}} \bigm| W_i\big]$$

is correctly specified, even if the other is misspecified. Note that the structural outcome equation

$$Y_i \; = \; \tau \, D_i \; + \; \beta^\top W_{i,\text{ld}} \; + \; \delta \, A_i \; + \; \phi_i$$

includes only the low-dimensional subset $W_{i,\text{ld}} \subseteq W_i$; however, since $(\mathrm{W}_i) = E\big[Y_i - \tau D_i - \beta^\top W_{i,\text{ld}} \bigm| W_i\big]$ conditions on the full W_i, the presence of $W_{i,\text{ld}}$ in the structural form is fully absorbed into $m(W_i)$ and does not alter the double-robustness proof below.

Theorem 3 (Double Robustness). Under Assumptions IV1–IV5 and Regularity Conditions RC1–RC6, let

$$\psi_i(\tau, \eta) \; = \; (Z_i - \pi(W_i))\big[Y_i - \tau D_i - m(W_i) + \tau \, q(Z_i, W_i)\big].$$

Suppose that either

$$\|\hat{q}(Z_i, W_i) - q(Z_i, W_i)\|_{L^2} = o_p(1) \quad \text{or} \quad \|\hat{m}(W_i) - m(W_i)\|_{L^2} = o_p(1).$$

Then

$$\frac{1}{n} \sum_{i=1}^{n} \psi_i(\tau, \hat{\eta}) \; = \; o_p(n^{-1/2}),$$

and consequently any root-n-consistent solution $\hat{\tau}$ to $\frac{1}{n} \sum_i \psi_i(\hat{\tau}, \hat{\eta}) = 0$ satisfies $\hat{\tau} \xrightarrow{p} \tau$.

Consider two cases:

Case 1: Correct Endogenous Variable (Treatment) Model $q(Z_i, W_i)$ Assume $q(Z_i, W_i) = E[D_i \mid Z_i, W_i]$ is correctly specified, while $m(W_i)$ may be misspecified.

Starting from the moment function,

$$\psi_i(\tau, \eta, \varphi) \;=\; \Big(Z_i - \pi(W_i)\Big)\Big(Y_i - \tau D_i - m(W_i) + \tau q(Z_i, W_i)\Big),$$

take expectations conditional on W_i:

$$\begin{aligned} E[\psi_i(\tau, \eta, \varphi) \mid W_i] = \;& E\Big[\big(Z_i - \pi(W_i)\big)\Big(Y_i - \tau D_i - m(W_i) + \tau q(Z_i, W_i)\Big) \;\Big|\; W_i\Big] \\ = \;& E\Big[\big(Z_i - \pi(W_i)\big)\Big(\delta A_i + \phi_i - m(W_i) + \tau(D_i - q(Z_i, W_i))\Big) \;\Big|\; W_i\Big], \end{aligned}$$

where I substituted $Y_i = \tau D_i + \delta A_i + \phi_i$.

Since $q(Z_i, W_i) = E[D_i \mid Z_i, W_i]$, Ihave $E[D_i - q(Z_i, W_i) \mid Z_i, W_i] = 0$. Under Assumptions (A1) and (A3), $E[(Z_i - \pi(W_i))A_i \mid W_i] = 0$ and $E[(Z_i - \pi(W_i))\phi_i \mid W_i] = 0$. Thus,

$$E[\psi_i(\tau, \eta, \varphi) \mid W_i] = \; E\Big[(Z_i - \pi(W_i))(-m(W_i)) \;\Big|\; W_i\Big] \;=\; 0,$$

because $E[Z_i - \pi(W_i) \mid W_i] = 0$.

Hence, even with a misspecified $m(W_i)$, the moment condition holds when the treatment model is correct. Cross-fitting and regularization ensure that errors in estimating $m(W_i)$ do not affect the first-order properties of $\hat{\tau}_{\text{ML-IV}}$. Therefore, $\hat{\tau}_{\text{ML-IV}}$ converges in probability to τ.

Case 2: Correct Outcome Model $m(W_{i,\text{ld}})$ Assume

$$m(W_{i,\text{ld}}) \;=\; E\big[Y_i - \tau D_i \mid W_{i,\text{ld}}\big]$$

is correctly specified, while the first-stage model $q(Z_i, W_i)$ may be misspecified. Starting from the orthogonal moment

$$\psi_i(\tau, \eta, \varphi) = \; (Z_i - \pi(W_i))\big(Y_i - \tau D_i - m(W_{i,\text{ld}}) + \tau\, q(Z_i, W_i)\big),$$

take expectations conditional on $W_{i,\text{ld}}$:

$$\begin{aligned} E\big[\psi_i(\tau, \eta, \varphi) \mid W_{i,\text{ld}}\big] &= E\Big[(Z_i - \pi(W_i))\big(Y_i - \tau D_i - m(W_{i,\text{ld}}) + \tau\, q(Z_i, W_i)\big) \;\Big|\; W_{i,\text{ld}}\Big] \\ &= E\Big[(Z_i - \pi(W_i))\big(\delta\, A_i + \phi_i + \tau\,(D_i - q(Z_i, W_i))\big) \;\Big|\; W_{i,\text{ld}}\Big], \end{aligned}$$

where I used $Y_i = \tau D_i + \beta^\top W_{i,\text{ld}} + \delta A_i + \phi_i$ and the definition of $m(\cdot)$. Since

$$E\big[(Z_i - \pi(W_i))\, A_i \mid W_{i,\text{ld}}\big] = 0, \quad E\big[(Z_i - \pi(W_i))\, \phi_i \mid W_{i,\text{ld}}\big] = 0, \quad E\big[D_i - q(Z_i, W_i) \mid Z_i, W_i\big] = 0,$$

it follows that

$$E\big[\psi_i(\tau, \eta, \varphi) \mid W_{i,\text{ld}}\big] = 0.$$

Thus $\hat{\tau}_{\text{ML-IV}}$ remains consistent even if $q(Z_i, W_i)$ is misspecified, provided $m(W_{i,\text{ld}})$ is correctly specified.

In both cases, the estimation of nuisance parameters η uses cross-fitting with advanced ML techniques governed by φ. While the proof here assumes $\varphi = \varphi_0$ for first-order consistency, the second-order analysis (discussed earlier) ensures that the inclusion of φ does not affect consistency. Specifically, boundedness and convergence rate conditions on $\Delta\eta$ and $\Delta\varphi$ guarantee that any bias introduced by advanced ML methods is asymptotically negligible.

3.3 Asymptotic Normality and Semiparametric Efficiency

Theorem 4 (Consistency and Asymptotic Normality of $\hat{\tau}_{\text{ML-IV}}$)*.* Under Assumptions IV1–IV5, Regularity Conditions RC1–RC4, and Rate Conditions RC5–RC6, the estimator $\hat{\tau}_{\text{ML-IV}}$ is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\tau}_{\text{ML-IV}} - \tau) \xrightarrow{d} \mathcal{N}(0, V),$$

where the asymptotic variance V is given by

$$V \;=\; \frac{\text{Var}\Big(w(Z_i, W_i)\,\big(\delta A_i + \phi_i\big)\Big)}{\Big(E\big[w(Z_i, W_i)\, D_i\big]\Big)^2}. \tag{13}$$

3.4 Efficiency Bound and Semiparametric Optimality

Theorem 4 established that

$$\sqrt{n}\big(\hat{\tau} - \tau\big) \;\xrightarrow{d}\; \mathcal{N}\big(0, V\big), \qquad V = \frac{\text{Var}\big(w(Z_i, W_i)\,(\delta A_i + \phi_i)\big)}{\big[E\{w(Z_i, W_i) D_i\}\big]^2}.$$

To verify that V is the semiparametric efficiency bound, an explicit asymptotic-linear representation is required together with a demonstration that the corresponding influence function lies in the tangent space of the model.

Lemma 4 (Tangent–Space Completeness). Assume the completeness condition $E\big[f(Z,W) \mid W\big] = 0 \Rightarrow f = 0$ a.s. Then the *closed* linear span of all score directions generated by perturbations of (π, q, m) is

$$\overline{\text{span}}\{\text{scores}\} = L_{2,0}(P), \quad L_{2,0}(P) := \{g \in L_2(P) : E[g] = 0\}.$$

Consequently every influence function in that space—and in particular $\Psi_i = w(Z_i, W_i)\,(Y_i - \tau D_i)/E[wD]$ —attains the minimum asymptotic variance among regular estimators (Chamberlain, 1987; Newey, 1990).

Lemma 5 (Remainder–Term Bound). Suppose $\|\hat{\eta} - \eta_0\|_{L^2} = o_p(n^{-1/4})$ and that the map $\eta \mapsto \psi_i(\tau, \eta)$ has a bounded second Gateaux derivative. Then the plug-in error

$$R_n \;=\; \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\psi}_i(\tau) - \psi_i(\tau)\big] \;=\; o_p(n^{-1/2}).$$

Lemma 6 (Semiparametric Efficiency of $\hat{\tau}$). Under IV1–IV5 and RC1–RC6, the completeness condition in Lemma 4, and the remainder control in Lemma 5,

$$\hat{\tau} - \tau \;=\; \frac{1}{n}\sum_{i=1}^{n}\Psi_i \;+\; o_p\big(n^{-1/2}\big), \qquad \Psi_i = \frac{w(Z_i, W_i)\,\big(Y_i - \tau D_i\big)}{S}, \quad S = E\{w(Z_i, W_i)D_i\}.$$

Hence $\mathrm{Var}(\Psi_i) = V$, and no regular estimator can attain an asymptotic variance strictly below V.

4 Monte Carlo Simulations

4.1 Data Generation Process

The data generation process adheres to the structural model specified in (1) and (2). For each simulation run, n observations are generated for the instrumental variables Z_i, covariates W_i, unobserved confounder A_i, endogenous variable D_i, and outcome variable Y_i.

Z_i consist of 5 independent variables drawn from a uniform distribution:

$$Z_{ij} \sim \mathcal{U}(0,1), \quad j = 1, \ldots, 5.$$

These instruments exhibit strong relevance for D_i and satisfy exogeneity assumptions, making them suitable for IV estimation.

The covariates W_i include $p = 200$ high-dimensional control variables:

$$W_{ik} \sim \mathcal{N}(0,1), \quad k = 1, \ldots, 200.$$

This setup reflects real-world scenarios where high-dimensional data necessitates effective regularization and feature selection to mitigate overfitting.

The unobserved confounder A_i is introduced as:

$$A_i \sim \mathcal{N}(0,1),$$

capturing unmeasured influences on both D_i and Y_i that generate endogeneity.

The endogenous variable D_i is generated using a combination of nonlinear and linear components:

$$D_i = g(Z_i, W_i) + \gamma A_i + \nu_i,$$

where $\gamma = 1.0$, $\nu_i \sim \mathcal{N}(0,1)$, and

$$g(Z_i, W_i) = \beta_1 Z_{i1} + \beta_2 Z_{i2}^2 + 0.5 \sin(Z_{i3}) + \sum_{j=4}^{5} Z_{ij} W_{ij}.$$

The linear component includes contributions from W_i and A_i:

$$\text{Linear Component} = \sum_{k=1}^{200} W_{ik} \left(\frac{\gamma}{200} \right) + \gamma A_i.$$

To ensure positivity of D_i, I impose:

$$D_i = \begin{cases} D_i, & \text{if } D_i > 0, \\ 0.1, & \text{otherwise.} \end{cases}$$

The outcome variable Y_i is generated as:

$$Y_i = \tau D_i + \sum_{k=1}^{200} W_{ik} \left(\frac{\gamma}{200}\right) + \delta A_i + \varepsilon_i,$$

where $\tau = 1.0$, $\delta = 0.5$, and $\varepsilon_i \sim \mathcal{N}(0, 1)$.

The true value of $\tau = 1.0$ serves as a benchmark for evaluating the estimated $\hat{\tau}$, with deviations providing insights into the estimator's bias, variance, and consistency. The accuracy of $\hat{\tau}$ across simulation runs directly supports the theoretical claim of Neyman orthogonality for the moment conditions.

The Monte Carlo simulations assess the estimator's performance across four strategies: Benchmark, Cross-Fitted, Regularized, and Cross-Fitted & Regularized. Each strategy involves the use of ML models to estimate the nuisance functions $\pi(W)$, $q(Z, W)$, and $m(W)$, which are then used to compute $\hat{\tau}$ by solving the orthogonal moment condition, as defined in (7). Note that the first-stage predictor $q(Z_i, W_i) = E[D_i \mid Z_i, W_i]$ is equivalent to the optimal instrument function $w(Z_i, W_i)$.

The Benchmark strategy does not employ regularization or cross-fitting, using the same data for both stages. This approach serves as a baseline to highlight potential overfitting and bias. The Cross-Fitted strategy partitions the data into $k = 5$ folds, with nuisance functions estimated on $k - 1$ folds and evaluated on the held-out fold. Cross-fitting mitigates overfitting and ensures orthogonality between the first and second stages. The Regularized strategy applies regularization and hyperparameter tuning techniques for linear and nonlinear algorithms to stabilize parameter estimates, optimize performance, and improve efficiency. The Cross-Fitted & Regularized strategy combines the benefits of cross-fitting and regularization. This approach reduces bias, variance, and overfitting while ensuring robustness to finite-sample challenges. In each strategy, the estimated $\hat{\tau}$ is compared against the true value.

4.2 Simulation Results

Figures 1 and 2 summarize the simulation results for the second-stage coefficient estimates (with the true effect equal to one) and their corresponding standard errors, respectively. Under the Benchmark strategy—using the same data to estimate both the nuisance functions and the causal effect—the estimated coefficient is consistently below one, particularly in smaller samples. This downward bias results from overfitting due to data reuse, which disrupts the orthogonality conditions necessary for rapid convergence of the first-stage estimators. In contrast, when cross-fitting is employed, the bias is substantially reduced. Partitioning the data so that the nuisance functions are estimated on folds independent of those used for causal inference preserves the orthogonality conditions central to the theoretical framework. For example, the cross-fitted and regularized version of XGBoost yields an estimated coefficient of approximately 1.107 in small samples; although this implies a slight upward bias (0.107) relative to the true value, the bias decreases with increasing sample size in line with the fast convergence rates predicted by the model. This indicates that cross-fitting not only counteracts the bias from data reuse but also accelerates the convergence of the first-stage estimates, enhancing second-stage inference.

The Benchmark models display substantially larger standard errors across all sample sizes. In contrast, cross-fitting markedly reduces the variability of the second-stage estimates. When combined with regularization, the resulting standard errors are consistently lower. The full DML estimator, which applies cross-fitting and residualization (as specified in (7)), achieves the smallest standard errors, confirming its $\sqrt{n}$-consistency and asymptotic normality. In the Reduced DML framework, where feature selection further refines the set of instruments and controls, even models that are typically less stable (e.g., non-linear SVR and MLP) exhibit improved precision with more consistent and lower standard errors.

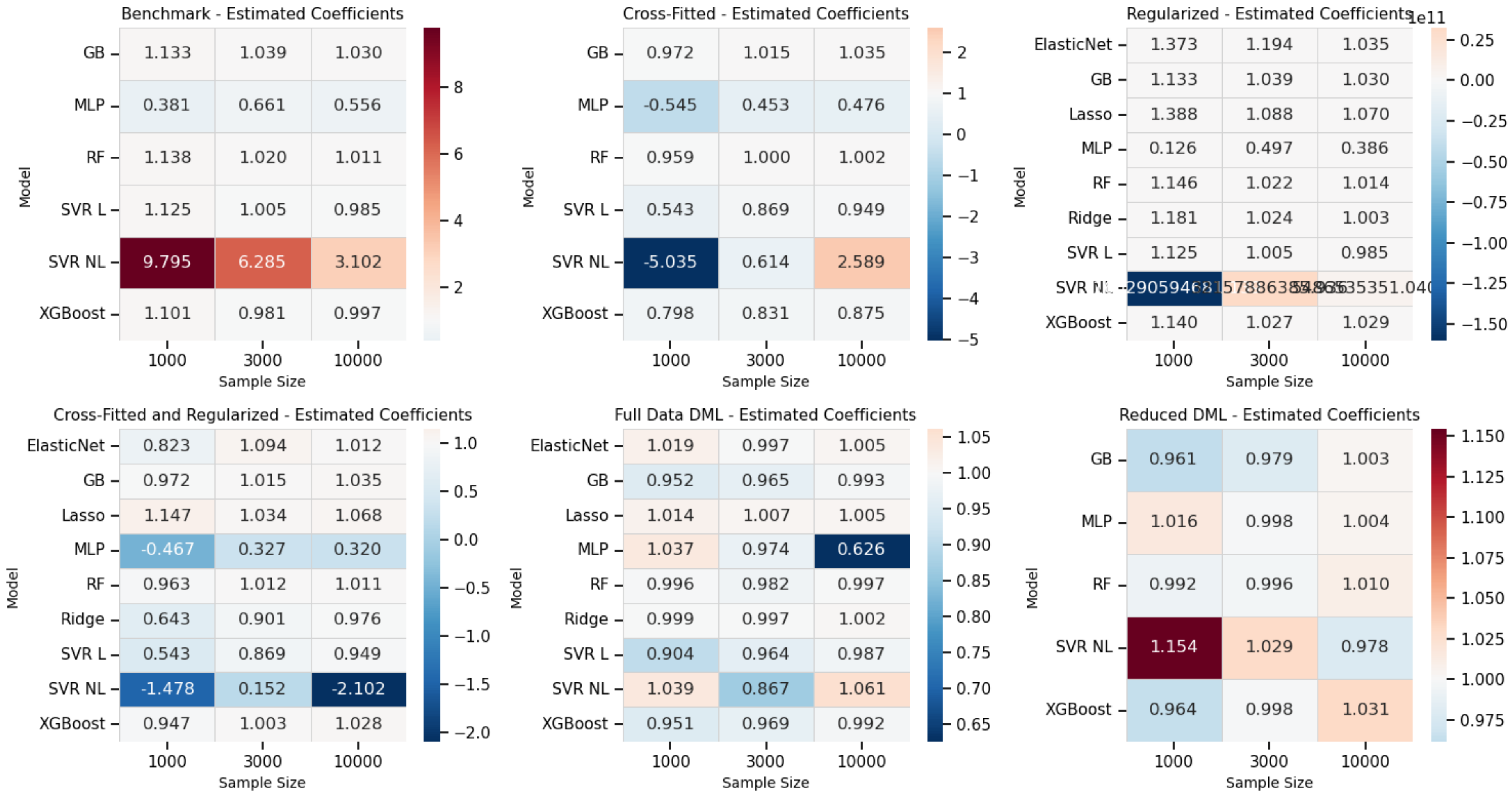

Figure 1: Estimated Coefficients for the Structural Parameter τ

The combination of cross-fitting with regularization yields a significant reduction in variance for the DML estimates relative to the cross-fitted and regularized plug-in ML estimates. Although regularization may introduce additional bias through shrinkage, its integration with cross-fitting in the DML framework mitigates this effect, thereby improving the overall bias–variance trade-off. In non-DML estimates, even when cross-fitting and regularization are applied, the estimated coefficients tend to deviate further from the true value and have higher standard errors than their DML counterparts. This difference underscores the importance of applying orthogonality with residualization.

Among the flexible ML algorithms considered, the regularized and cross-fitted linear methods (Ridge, Lasso, EN) yield $\hat{\tau}$ estimates closest to the true value in the full-data DML framework at $n = 10,000$, with the corresponding standard errors being the smallest (as shown in Figure 2). For nonlinear models, the Reduced DML estimates provide more significant improvements for MLP and SVR, with slight improvements for GB, while full-data DML estimates for GB and RF yield the best coefficients overall. These findings suggest that, among all strategies, Ridge from full-data DML offers the most accurate and precise estimation of τ. This corroborates the importance of fully nonparametric, data-driven model selection when implementing ML methods in IV estimation for high-dimensional settings. As discussed in Section 2.2, I introduce the nuisance parameter vector $\eta = (\pi, q, m)$ and impose the associated convergence rate and Neyman orthogonality conditions (see Equation (7)). These theoretical conditions are fundamental to the proposed estimator, ensuring that the estimation errors in the nuisance functions do not contaminate the second-stage causal inference. This framework—together with the optimal instrument function $w(Z_i) = E[W_i \mid Z_i]$ introduced in the same section—justifies the use of fully nonparametric, data-driven model selection. The improved performance of Ridge in the full-data DML estimates, as evidenced in the simulation results, thus corroborates the importance of these theoretical properties.

In sum, the estimated coefficients and standard errors validate the theoretical properties of the estimator. The evidence demonstrates that cross-fitting is essential for eliminating bias from data reuse, and that combining cross-fitting with regularization—and further refining the nuisance functions through feature selection in Reduced DML—ensures that the causal effects converge to the true parameter and that the standard errors decrease in accordance with $\sqrt{n}$-asymptotic theory. The findings support the claims of optimal instrument approximation, semiparametric efficiency, and double robustness in the proposed estimator.

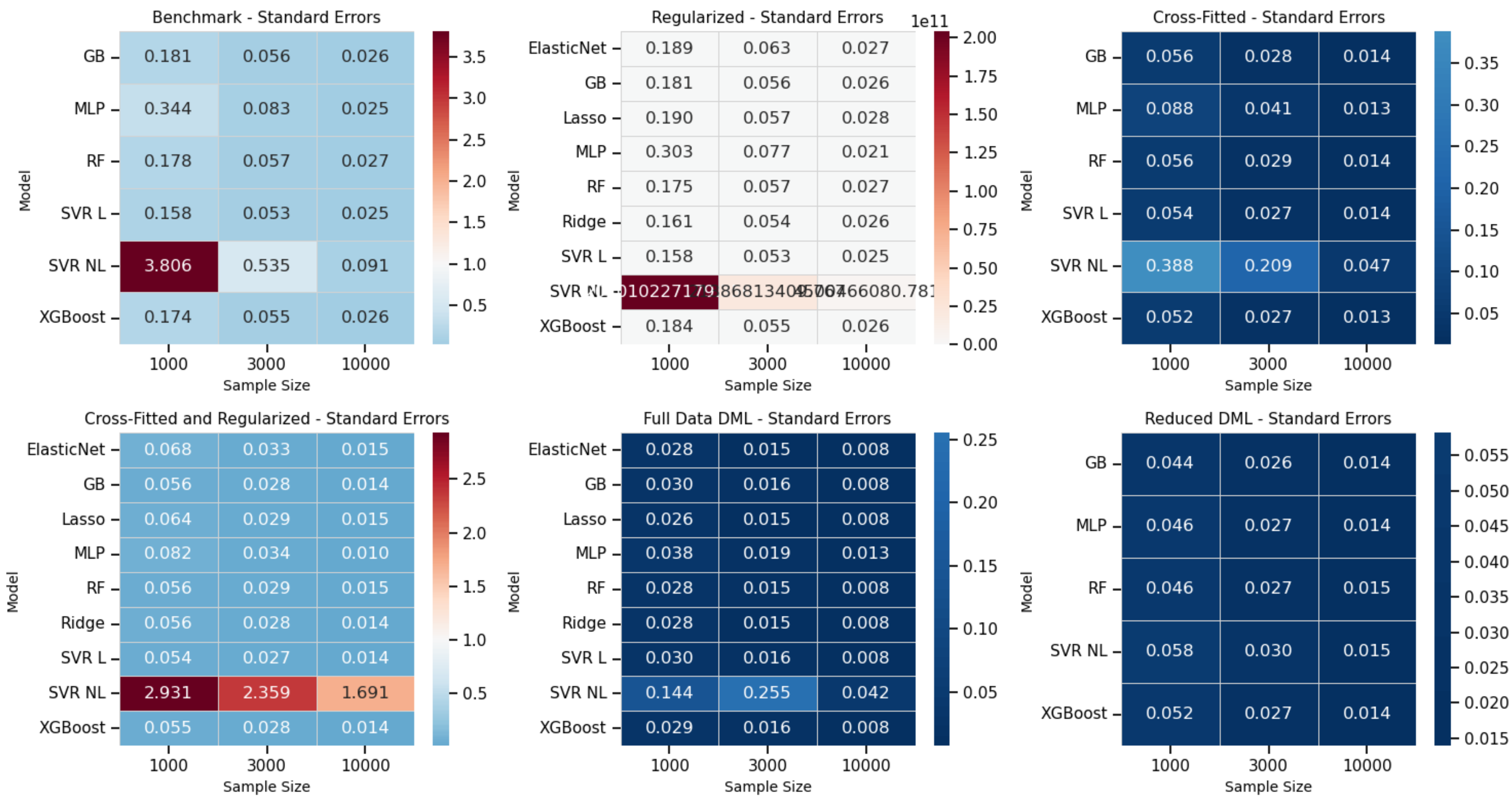

Figure 2: Standard Errors for the Structural Parameter τ

Figure 3 in Appendix **Appendix E** reports the first-stage prediction metrics—bias, variance, mean squared error (MSE), and coverage—for the various estimation strategies. In the first stage, the aim is to predict the endogenous variable D_i using instruments Z_i and controls W_i. Note that in the cross-fitted and regularized strategies, there is no difference between plug-in ML and DML estimators, as residualization is performed only in the second stage.

Across all strategies, as the sample size increases from $n = 1000$ to $n = 10\,000$, both the bias and MSE generally decrease for almost all models, with the notable exceptions of MLP and SVR in some cases. Importantly, in the cross-fitted and regularized estimates, the prediction variance decreases consistently only for Ridge regression; in contrast, the variance for most other ML models tends to increase with larger sample sizes. This finding is in line with the second-stage results, which consistently indicate that Ridge regression is the best prediction model.

Moreover, among the ML models, the highest coverage rates in the regularized and cross-fitted estimates are observed for Ridge and the nonparametric linear prediction (NLP) approach, although NLP is less stable than Ridge. For the regularized and cross-fitted strategies, the bias for Ridge, Lasso, and Elastic Net lies between the bias observed in the solely cross-fitted and solely regularized models, yet remains significantly lower than that of the benchmark models. For nonlinear models, such as XGBoost and Gradient Boosting, the regularized and cross-fitted estimates yield consistently lower bias compared to the Benchmark, Regularized, and Cross-fitted approaches alone, suggesting that these models capture the nonlinearity in the data effectively. Nonetheless, Ridge regression remains the best overall model, providing the most accurate and precise first-stage predictions.

Furthermore, when comparing the full-data DML with the Reduced DML approaches (where Reduced DML employs feature selection to retain only the most important instruments and controls) in Figure 4 in Appendix **Appendix E**, the results from a sample size of $n = 10\,000$ indicate that the Reduced DML estimates achieve smaller bias and MSE for nonlinear models such as XGBoost, MLP, and SVR—with a particularly significant decrease in variance for MLP. These findings from the Reduced DML approach highlight that effective feature selection and dimensionality reduction further stabilize the estimates for models that are otherwise highly variable. Nonetheless, Ridge regression consistently yields the closest predictions to the true value with the smallest standard errors across all strategies and sample sizes.

The empirical findings from Monte Carlo simulations closely mirror the inferences drawn in the literature [e.g., Chernozhukov et al., 2018, 2022], demonstrating that cross-fitting plays a critical role in mitigating the bias arising from data reuse. Models that do not employ cross-fitting—such as the Benchmark and solely Regularized strategies—consistently exhibit substantially higher bias and variance, particularly in smaller samples or for complex model classes. In contrast, cross-fitted estimators, especially when combined with carefully tuned regularization, yield estimated coefficients converging toward the true value with reduced variability and improved confidence interval coverage. While cross-fitting or regularization alone partly mitigates overfitting, their combined use fully leverages asymptotic guarantees required for reliable second-stage inference, substantially reducing bias, variance, and coverage deficiencies. Penalized linear methods (e.g., Lasso, Elastic Net) and well-tuned boosting algorithms (e.g., GB, XGBoost) tend to achieve a desirable balance of low bias and stable variance at moderate sample sizes, whereas kernel-based or neural network models may require larger data to reach comparable stability. Across all sample sizes, the synergy of cross-fitting and shrinkage ensures that nuisance estimates satisfy the requirements for a doubly robust, Neyman-orthogonal IV estimator, thereby enabling the proposed ML-IV method to accurately recover the causal effect τ even under high dimensionality, unobserved confounding, and the pursuit of an optimal instrument function.

5 Estimating the Return to Education

Estimating the causal effect of education on earnings remains a central challenge in applied econometrics due to the potential endogeneity of schooling. Angrist and Krueger (1991) illustrated that OLS estimates may be biased because unobserved factors are likely correlated with educational attainment. Following Angrist and Krueger (1991), I reproduce the core empirical estimation strategy (Models I and II in Table V) using U.S. Census data for men born between 1930 and 1939 and constructs an instrument set that includes interactions between quarter-of-birth and year-of-birth dummies. The replication is validated by closely matching the traditional OLS and TSLS estimates of the return to education.

Table 1 reports the traditional OLS and TSLS-IV estimates along with the benchmark plug-in ML models. Only in the last benchmark strategy, train/test split is applied. Accordingly, in my replication of Angrist and Krueger (1991), the traditional TSLS-IV estimator yields a causal effect of roughly 0.089 with a robust standard error of about 0.016, while OLS produces a lower estimate of 0.071 with negligible standard errors due to endogeneity bias. Extending the analysis, naive full-sample plug-in ML models as benchmark ML strategy generate point estimates that closely replicate the TSLS-IV result. However, these full-sample ML models exhibit larger standard errors, likely due to overfitting in the first-stage estimation when the same data are reused in both stages, which in turn inflates the variance in the second-stage inference.

Table 1: OLS, Traditional TSLS, and Benchmark ML Models for the Return to Education

OLS and Traditional TSLS IV		Benchmark ML (without Train/Test Split)				Benchmark ML (with Train/Test Split)			
OLS	TSLS	LR	RF	GB	XGBoost	LR	RF	GB	XGBoost
0.0711	0.0891	0.0891	0.0886	0.0923	0.0830	0.0846	0.0841	0.0944	0.0800
(0.0000)	(0.0162)	(0.0172)	(0.0169)	(0.0228)	(0.0180)	(0.0167)	(0.0165)	(0.0227)	(0.0175)
[0.0700-0.0720]	[0.0573-0.1209]	[0.0555-0.1228]	[0.0554-0.1218]	[0.0476-0.1370]	[0.0478-0.1183]	[0.0519-0.1172]	[0.0517-0.1166]	[0.0498-0.1389]	[0.0457-0.1143]

Notes: This table reports traditional OLS and TSLS-IV estimates along with benchmark plug-in ML models that do not employ cross-fitting. The estimated coefficient τ represents the return to education; standard errors are presented in parentheses and the corresponding 95% confidence intervals are provided. The results are based on 329,509 observations and an instrument set of 30 variables.

When a simple train–test split is introduced the ML estimates maintain similar point estimates but with notably reduced standard errors. This reduction results from partitioning the data so that the same observations are not used for both the first-stage nuisance estimation and the second-stage regression, thereby mitigating overfitting and reducing the upward bias in variance estimates. Overall, while both traditional TSLS-IV and the full-sample plug-in ML approaches deliver comparable point estimates, the additional variance introduced by data reuse in the full-sample approach can be alleviated by a simple train–test split, leading to more efficient and reliable inference.

Table 2 presents plug-in ML estimators that incorporate a fivefold cross-fitting strategy, thereby differing from the full-sample and train–test split approaches reported in Table 1. In Panel A, I report estimates from models that are cross-fitted without further regularization/tuning-up, whereas Panel B displays the corresponding estimates when regularization/hyperparameter tuning is additionally applied. The plug-in ML models estimated with cross-fitting in Panel A yield systematically lower point estimates relative to those obtained from the simple train–test split approach reported in Table 1. In addition, the standard errors are smaller and, when combined with the lower point estimates, lead to narrower confidence intervals relative to those obtained under the train–test split framework. This improvement is consistent with the theoretical properties of the model that cross-fitting reduces overfitting in the first-stage estimation and consequently alleviates the upward bias in variance estimates that typically arises from reusing the same data in both stages.

Turning to the regularized models in Panel B, several important differences emerge relative to both the train–test split results in Table 1 and the non-regularized cross-fitted models in Panel A in Table 1. First, the point estimates for RF, GB, and XGBoost often shift, sometimes notably, from their Panel A values. RF decreases from about 0.0606 to 0.0338, while GB remains at 0.0524 but with an altered variance structure. Second, the standard errors for the regularized RF and XGBoost in Panel B exceed those in Panel A (and in some instances, those in the train–test split results as well), leading to wider confidence intervals. This increase in variance can arise when hyperparameter tuning or penalization shrinks the first-stage predictions in ways that amplify second-stage sampling variability—particularly if the regularization inadvertently reduces the effective strength of the instruments or imposes constraints that introduce more sensitivity in finite samples.

A similar pattern is evident among the linear estimators in Panel B. Compared to (unpenalized) LR, the Ridge, Lasso, and EN estimates differ in both magnitude and precision. Lasso (0.0758) and EN (0.0723) yield higher point estimates than LR (0.0629 or 0.0643 for Ridge), but their standard errors—especially Lasso's (0.0267)—are larger, reflecting the extra variability introduced by penalization. I include both LR and these penalized linear methods in Panel B to illustrate how regularization within a linear-model framework can reshape estimates of the causal parameter. In principle, regularization should guard against overfitting in high-dimensional or flexible specifications, but it can also increase variance in the second stage if the penalized fit is highly sensitive to tuning parameters or if the first-stage predictions become too aggressively shrunk. Thus, while combining cross-fitting with penalization and tuning often mitigates overfitting, it might lead to higher standard errors in finite samples; the ultimate effect depends on how the penalty and hyperparameters interact with the data-generating process

Table 2: Plug-in ML Estimates with Cross-Fitting for the Return to Education ($K = 5$)

Panel A			
Plug-in ML Models (Cross-Fitted but not Regularized)			
LR	RF	GB	XGBoost
0.0629	0.0606	0.0524	0.0557
(0.0165)	(0.0165)	(0.0220)	(0.0174)
[0.0304-0.0953]	[0.0284-0.0929]	[0.0092-0.0956]	[0.0217-0.0897]

Panel B						
Plug-in ML Models (Cross-Fitted and Regularized)						
LR	Ridge	Lasso	EN	RF	GB	XGBoost
0.0629	0.0643	0.0758	0.0723	0.0338	0.0524	0.0524
(0.0165)	(0.0170)	(0.0267)	(0.0219)	(0.0237)	(0.0220)	(0.0220)
[0.0304-0.0953]	[0.0310-0.09576]	[0.0235-0.1282]	[0.0294-0.1153]	[-0.0127-0.0803]	[0.0092-0.0956]	[0.0092-0.0956]

Notes: Table reports plug-in ML estimates obtained via a fivefold cross-fitting procedure. Panel A displays estimates without additional regularization (corresponding to standard linear regression models and default settings for nonlinear methods), whereas Panel B shows estimates when regularization and hyperparameter tuning are applied. In Panel B, Ridge, Lasso, and Elastic Net estimators are included to assess the effect of penalization, and the nonlinear models are tuned (e.g., Random Forest with a maximum depth of 5, Gradient Boosting with a maximum depth of 3, and XGBoost with corresponding parameter choices) to further mitigate overfitting and improve estimation precision. Standard errors are reported in parentheses and the 95% confidence intervals in square brackets; the estimated coefficient τ denotes the return to education.

Table 3: DML Estimates with Cross-Fitting for the Return to Education ($K = 5$)

Panel A

DML Models (Cross-Fitted but not Regularized)

LR	RF	GB	XGBoost
0.0908	0.0899	0.0670	0.0894
(0.0205)	(0.0204)	(0.0227)	(0.0230)
[0.0507-0.1309]	[0.0449-0.1299]	[0.0225-0.1114]	[0.0443-0.1345]

Panel B

DML Models (Cross-Fitted and Regularized)

LR	Ridge	Lasso	EN	RF	GB	XGBoost
0.0908	0.0909	0.0904	0.0880	0.0627	0.0670	0.0672
(0.0205)	(0.0209)	(0.0249)	(0.0221)	(0.0262)	(0.0227)	(0.0227)
[0.0507-0.1309]	[0.0500-0.1318]	[0.0417-0.1392]	[0.0447-0.1314]	[0.0114-0.1141]	[0.0225-0.1114]	[0.0227-0.1117]

Notes: Table reports DML estimates obtained via a fivefold cross-fitting procedure and Neyman orthogonality. Panel A presents models estimated without additional regularization or hyperparameter tuning, whereas Panel B displays models that incorporate regularization/tuning. In Panel B, Ridge, Lasso, and Elastic Net (EN) estimators are included to assess the impact of different penalty structures relative to unpenalized linear regression (LR). For the nonlinear methods, tuning parameters (e.g., maximum tree depth, learning rate) are adjusted to control model complexity. Standard errors are reported in parentheses and the 95% confidence intervals in square brackets; the estimated coefficient τ denotes the return to education.

and the strength of the instruments.

Table 3 reports DML estimates of the return to education, τ, obtained under a cross-fitting framework with $K = 5$ folds. Panel A presents cross-fitted models without additional regularization or hyperparameter tuning, whereas Panel B displays models in which regularization/tuning is applied. For consistency, the same regularization and hyperparameter tuning methods are employed for the same linear and non-linear ML algorithms in the plug-in ML estimation strategy.

Both cross-fitting without and with regularization yield point estimates that are broadly consistent with traditional TSLS-IV and the benchmark plug-in ML models. The combination of cross-fitting with regularization/tuning provides a nuanced trade-off between bias and variance. In Panel B, the inclusion of Ridge, Lasso, and EN models alongside standard LR facilitates a direct comparison of different linear regularization approaches. The observed differences underscore the importance of careful hyperparameter selection to balance the benefits of mitigating overfitting against the potential for increased variability.

Linear DML models such as Linear Regression, Ridge, Lasso, and EN closely match the TSLS-IV benchmark. In contrast, nonlinear DML models yield lower estimates, suggesting that controlling for nonlinearity via cross-fitting and residualization effectively corrects the upward bias observed in naïve plug-in models. Although the standard errors for the nonlinear DML models are slightly larger, the confidence intervals remain distinct from those of the benchmark models. Overall, the DML-IV strategy achieves near-zero bias in the first stage, consistent variance, and high coverage rates (around 93.35%), thereby ensuring valid inference with $\sqrt{n}$-consistency.

6 Conclusions

The paper has introduced a novel DML-IV estimator for causal inference in high-dimensional settings with unobserved confounders, effectively addressing endogeneity. Although the existing literature has established that such estimators can attain $\sqrt{n}$–consistency and asymptotic normality despite nuisance estimation errors, the contribution of this paper lies in the full integration of cutting-edge, fully data–driven ML methods into the IV framework. Specifically, the proposed estimator employs cross-fitting, advanced regularization/hyperparameter tuning, data-driven feature reduction, and adaptive model selection to stabilize first-stage predictions.

In addition, it implements an adaptive construction of the optimal instrument—namely, the conditional expectation $E[D_i \mid Z_i, W_i]$—by leveraging both linear and nonlinear ML models to combine high-dimensional controls and candidate instruments, thereby enhancing predictive accuracy and instrument strength. This approach not only improves identification strength and mitigates bias, but it also ensures that small estimation errors in the nuisance functions affect the estimator only at a second-order level (i.e., $o_p(n^{-1/2})$).

The theoretical analysis demonstrates that the estimator satisfies Neyman orthogonality—its moment function's first derivative with respect to the nuisance parameters vanishes at the true values—thus ensuring that the first-order asymptotic properties (consistency, asymptotic normality, and semiparametric efficiency) remain unchanged even when these functions are estimated via complex, flexible ML methods. Moreover, because the optimal instrument is estimated in a fully data–driven, nonparametric fashion, the estimator exhibits double robustness and attains the semiparametric efficiency bound for τ, establishing it as optimally efficient among regular, asymptotically linear estimators.

Empirical evidence from extensive Monte Carlo simulations and an application to estimating the return to education corroborate these theoretical findings. The DML-IV estimator consistently outperforms conventional TSLS-IV and plug-in ML estimators by achieving lower bias, reduced variance, and improved mean squared error, while maintaining appropriate coverage probabilities across varied sample sizes and model specifications. Thus, the paper makes a significant contribution to the literature by (i) providing a fully data–driven, adaptive estimation framework that leverages cutting-edge ML methods—including the construction of optimal instruments—for robust causal inference; (ii) rigorously establishing that the estimator's robust properties safeguard its first-order asymptotic behavior against second-order estimation errors; and (iii) demonstrating, through both theory and empirical evidence, that the proposed approach enhances the precision and reliability of causal effect estimation in complex, high-dimensional settings. These advances extend the existing theoretical foundations and offer practical improvements for applications where traditional IV methods and plug-in ML estimators fall short.

References

Amemiya, T. (1974). The non-linear two-stage least squares estimator. *Journal of Econometrics*, 2:105–110.

Andrews, D. W.K. and Stock, J. H. (2005). Inference with weak instruments. *Annals of Economics and Finance*, 6(1):1–60.

Angrist, J. D., and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4), 979-1014.

Angrist, J. D. and Imbens, G. W. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Athey, S. and Imbens, G. W. (2019). Machine learning methods economists should know about. *Annual Review of Economics*, 11:685–725.

Athey, S., Imbens, G. W. and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B*, 80(4):597–623.

Bai, J. and Ng, S. (2009). Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, 1(1):Article 4.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

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., Fernandez-Val, I. and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Caner, M. (2009). LASSO-type GMM estimator. *Econometric Theory*, 25:270–290.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334.

Chang, N. C. (2020). Double/debiased machine learning for difference-in-differences models. *Econometrics Journal*, 23(2):177–191.

Chen, J., Chen, D. L., and Lewis, G. (2021). Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models. *Working Paper*, Harvard Business School; Toulouse School of Economics; Microsoft Research.

Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321.

Chen, X., Shi, Z. and Xu, Z. (2022). Deep learning and alternative modeling strategies in econometrics. *Journal of Econometrics*, 228(2):129–150.

Chernozhukov, V., Chetverikov, D., Demirer, M., et al. (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. (2021). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

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

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

Darolles, S., Fan, Y., Florens, J.-P. and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

Farrell, M. H., Liang, T. and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Gautier, E. and Tsybakov, A. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.

Gold, D., Lederer, J., and Tao, J. (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1):79–111.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press.

Hansen, B. E. and Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, 182(2):290–308.

Hansen, B. E. Hausman, J, and Newey, D. (2008). Estimation with many instrumental variables. *Journal of Business Economic Statistics*, 26(3):398–422.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1414–1423.

Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669-738.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394.

Horowitz, J. L. and Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75(4):1191–1207.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(5):809–837.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Newey, W. K. and Robins, J. M. (2018). Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation. Working paper.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics*, pages 416–444. Wiley.

Neyman, J. (1979). c() tests and their use. In *Sankhya*, pages 1–21.

Okui, R., Small, D. S., Tan, Z., and Robins, J. M. (2012). Doubly robust instrumental variable regression. *Statistica Sinica*, 22(1):173–205.

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

Sun, B., Cui, Y. and Tchetgen, E. (2022). Selective machine learning of the average treatment effect with an invalid instrumental variable. *Journal of Machine Learning Research*, 23:1–40.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

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.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Appendix A Notation

In the paper, data are organized as a triangular array $\{z_{i,n}\}_{i=1}^{n}$ on a common probability space $(\Omega, \mathcal{A}, P)$. For each sample size n, each observation $z_{i,n}$ is a vector comprising the outcome Y_i, the endogenous treatment D_i, and the covariate vector X_i, where X_i includes both the instrumental variables Z_i and the controls W_i[4]. Write $W_i = (W_{i,\text{ld}}, W_{i,\text{hd}})$, where only $W_{i,\text{ld}}$ enters linearly in the structural equation and the remaining controls $W_{i,\text{hd}}$ are absorbed nonparametrically into the outcome-nuisance function $m(W_i)$. In the empirical implementation, I begin by including every available covariate in the first-stage ML predictor of D_i (i.e. $W_i \equiv W_{i,\text{hd}}$). After fitting a flexible library of learners, I use feature-importance scores to prune $W_{i,\text{hd}}$ down to a smaller subset for the reduced-DML runs. By contrast, for the second-stage structural equation I select a low-dimensional subset $W_{i,\text{ld}}$ a priori on economic grounds (e.g. age, gender, baseline covariates known to affect Y_i). All remaining high-dimensional variation is absorbed nonparametrically in $m(W_i)$.

Although observations are independent across i, they are allowed to be nonidentically distributed; parameters characterizing the distribution of $z_{i,n}$ are implicitly indexed by n, though this index is suppressed for notational simplicity. The empirical expectation of a function $f(z)$ is denoted by $\mathbb{E}_n[f(z)] = \frac{1}{n}\sum_{i=1}^{n} f(z_i)$, and the population expectation is represented by $\mathbb{E}[f(z)]$.

For any vector $v \in \mathbb{R}^p$, the ℓ_2-norm is denoted $\|v\|_2 = \sqrt{\sum_{j=1}^{p} v_j^2}$ and the ℓ_1-norm by $\|v\|_1 = \sum_{j=1}^{p} |v_j|$. $\|v\|_0$ represents the number of nonzero components of v. The L_2 norm of a function f is defined by $\|f\|_{2,n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f(z_i)^2}$. For a vector $v \in \mathbb{R}^p$ and an index set $T \subset \{1, \dots, p\}$, v_T refers to the vector that agrees with v on T and is zero elsewhere, while T^c denotes the complement of T. For any two positive sequences a_n and b_n, $a_n \asymp b_n$ indicates that there exist constants $c, C > 0$ independent of n such that $c\,b_n \le a_n \le C\,b_n$. The standard notations $O_p(\cdot)$ and $o_p(\cdot)$ are employed for convergence in probability.

In the proposed adaptive ML-IV model, the endogenous regressor D_i is generated by $f(Z_i, W_{i,\text{hd}}; \lambda, \theta)$, which is estimated via ML methods incorporating regularization, hyperparameter tuning, feature reduction, and model selection. Here λ and θ parameterize these adaptive ML techniques. Thus the entire control vector W_i enters nonparametrically via $m(W_i)$, while only $W_{i,\text{ld}}$ appears in the linear second stage.

[4] In the first-stage ML predictor I use the full covariate vector W_i ("high-dimensional," so $W_i \equiv W_{i,\text{hd}}$), while in the second-stage regression I include only a selected subset $W_{i,\text{ld}}$ (low-dimensional). Thus $W_{i,\text{ld}} \subset W_i \equiv W_{i,\text{hd}}$.

In the model, the optimal instrument function is defined as $w(Z_i, W_i) = E[D_i \mid Z_i, W_i]$, which plays a key role in achieving semiparametric efficiency. In practice, this function is estimated from the data—denoted by $\hat{w}(Z_i, W_i)$—and is used to form the plug-in prediction $\hat{D}_i$ in the first stage and subsequently in the construction of the moment condition for estimating τ. The main moment condition is given by $\psi_i(\tau, \eta, \varphi)$, which, under the maintained IV assumptions and mild smoothness conditions, identifies the causal parameter τ. The nuisance estimators are assumed to converge at rates satisfying $\|\hat{\eta} - \eta\|_{L^2} = o_p(n^{-1/4})$, ensuring that the estimator for τ is $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient.

Appendix B Proofs

Appendix B.1 Proof of Theorem 1

Start from the structural equation with the low-dimensional controls included:

$$Y_i \;=\; \tau\, D_i \;+\; \beta^\top W_{i,\mathrm{ld}} \;+\; \delta\, A_i \;+\; \phi_i.$$

Rearrange to isolate the part orthogonal to W_i:

$$Y_i - \beta^\top W_{i,\mathrm{ld}} \;=\; \tau\, D_i + \delta\, A_i + \phi_i.$$

Multiply both sides by Z_i and take conditional expectation given W_i:

$$E\big[Z_i\,(Y_i - \beta^\top W_{i,\mathrm{ld}}) \mid W_i\big] = \tau\, E\big[Z_i\, D_i \mid W_i\big] \;+\; \delta\, E[Z_i\, A_i \mid W_i] \;+\; E[Z_i\, \phi_i \mid W_i].$$

By the IV exogeneity assumptions $E[Z_i A_i \mid W_i] = 0$ and $E[Z_i \phi_i \mid W_i] = 0$, this simplifies to

$$E\big[Z_i\,(Y_i - \beta^\top W_{i,\mathrm{ld}}) \mid W_i\big] = \tau\, E\big[Z_i\, D_i \mid W_i\big].$$

Next, subtract $\pi(W_i)\,(Y_i - \beta^\top W_{i,\mathrm{ld}})$ from the left-hand side and $\tau\,\pi(W_i)\,D_i$ from the right-hand side, where $\pi(W_i) = E[Z_i \mid W_i]$. I obtain

$$E\big[(Z_i - \pi(W_i))\,(Y_i - \beta^\top W_{i,\mathrm{ld}}) \mid W_i\big] = \tau\, E\big[(Z_i - \pi(W_i))\, D_i \mid W_i\big].$$

Rewriting the left side as a single conditional expectation gives

$$E\big[(Z_i - \pi(W_i))\,(Y_i - \beta^\top W_{i,\mathrm{ld}} - \tau\, D_i) \mid W_i\big] = 0.$$

Finally, insert the definitions of the nuisance functions

$$m(W_i) = E[Y_i - \beta^\top W_{i,\mathrm{ld}} - \tau D_i \mid W_i], \quad q(Z_i, W_i) = E[D_i \mid Z_i, W_i],$$

and add and subtract $m(W_i)$ and $\tau\, q(Z_i, W_i)$ inside the expectation:

$$0 = E\Big[(Z_i - \pi(W_i))\big(Y_i - \beta^\top W_{i,\text{ld}} - \tau\, D_i - m(W_i) + \tau\, q(Z_i, W_i)\big) \;\Big|\; W_i\Big].$$

This is exactly the conditional form of the orthogonal moment $\psi_i(\tau, \eta, \varphi)$ in (7), hence τ is identified by $E[\psi_i(\tau, \eta, \varphi) \mid W_i] = 0$, completing the proof.

B.2 Proof of Lemma 1

Write

$$\psi_i(\tau, \eta{+}\varepsilon h, \varphi) \;=\; (Z_i{-}\pi(W_i){-}\varepsilon\, h_\pi(W_i))\Big[Y_i{-}\tau D_i{-}m(W_i){-}\varepsilon\, h_m(W_i){+}\tau\{q(Z_i, W_i){+}\varepsilon\, h_q(Z_i, W_i)\}\Big].$$

Differentiating under the expectation and evaluating at $\varepsilon = 0$ gives

$$D_\eta E[\psi_i(\tau, \eta_0, \varphi_0)]\big[h\big] = -E\big[h_\pi(W_i)\, \omega_i\big] - E\big[(Z_i - \pi(W_i))\,(h_m(W_i) - \tau\, h_q(Z_i, W_i))\big],$$

where $\omega_i = Y_i - \tau D_i - m(W_i) + \tau\, q(Z_i, W_i)$.

By the definitions

$$E\big[(Z_i - \pi(W_i))\,(D_i - q(Z_i, W_i)) \mid W_i\big] = 0, \quad E\big[Y_i - \tau D_i - m(W_i) \mid W_i\big] = 0,$$

and the law of iterated expectations, each of the two terms above vanishes.

Hence $D_\eta E[\psi_i(\tau, \eta_0, \varphi_0)]\big[h\big] = 0$. □

B.3 Proof of Lemma 2

Define the map

$$\mathcal{T} : (\eta, \varphi) \;\mapsto\; E\big[\psi(\theta_0, \eta, \varphi)\big].$$

Let

$$h \;=\; (\hat\eta - \eta_0,\, \hat\varphi - \varphi_0),$$

and perform a second–order Gateaux expansion of $\mathcal{T}$ around (η_0, φ_0). For some $t \in (0, 1)$,

$$\mathcal{T}(\eta_0 + h_\eta,\ \varphi_0 + h_\varphi) = \mathcal{T}(\eta_0, \varphi_0) + D\mathcal{T}_{(\eta_0,\varphi_0)}[h] + \tfrac{1}{2}\, D^2\mathcal{T}_{(\eta_0,\varphi_0)}[h, h] + R(h),$$

where

$$D\mathcal{T}_{(\eta_0,\varphi_0)}[h] \ = \ E\Big[\psi_\eta(\theta_0, \eta_0, \varphi_0)[h_\eta] \ + \ \psi_\varphi(\theta_0, \eta_0, \varphi_0)[h_\varphi]\Big],$$

and the remainder satisfies

$$R(h) \ = \ \tfrac{1}{6}\, D^3\mathcal{T}_{(\eta_0+t\,h)}[h, h, h] \ = \ o\big(\|h\|^2\big) \quad \text{by SO3.}$$

By the Neyman-orthogonality condition (SO0), the first-order term vanishes:

$$D\mathcal{T}_{(\eta_0,\varphi_0)}[h] = 0.$$

Hence

$$\mathcal{T}(\hat{\eta}, \hat{\varphi}) - \mathcal{T}(\eta_0, \varphi_0) = \tfrac{1}{2}\, D^2\mathcal{T}_{(\eta_0,\varphi_0)}[h, h] \ + \ R(h).$$

By SO1 the bilinear form $D^2\mathcal{T}$ is uniformly bounded:

$$\big|D^2\mathcal{T}_{(\eta_0,\varphi_0)}[h, h]\big| \ \le \ C\, \|h\|^2.$$

Finally, SO2 ensures $\|h\|^2 = O_p(n^{-1/2})$.

Combining these bounds gives

$$\big|\mathcal{T}(\hat{\eta}, \hat{\varphi}) - \mathcal{T}(\eta_0, \varphi_0)\big| \ \le \ \tfrac{1}{2}\, C\, \|h\|^2 \ + \ o\big(\|h\|^2\big) \ = \ O_p(n^{-1/2}),$$

as claimed. □

B.4 Proof of Theorem 2

Consider the moment function in (7). Introduce a small perturbation $\epsilon h = (\epsilon h_\pi, \epsilon h_q, \epsilon h_m)$ to the nuisance parameters η, while treating φ as fixed (i.e., $\Delta\varphi = 0$). That is,

$$\eta + \epsilon h = (\pi + \epsilon h_\pi,\ q + \epsilon h_q,\ m + \epsilon h_m).$$

Substitute these perturbed parameters into ψ_i, expand, and retain only first-order terms:

$$\psi_i(\tau, \eta + \epsilon h, \varphi) \approx \psi_i(\tau, \eta, \varphi) \\ - \epsilon\, h_\pi(W_i)\, \omega_i \;-\; \epsilon\, (Z_i - \pi(W_i))\big(h_m(W_i) - \tau\, h_q(Z_i, W_i)\big),$$

where

$$\omega_i \;=\; Y_i \;-\; \tau\, D_i \;-\; m(W_i) \;+\; \tau\, q(Z_i, W_i).$$

Taking expectations conditional on W_i gives:

$$E\big[\psi_i(\tau, \eta + \epsilon h, \varphi) \mid W_i\big] \approx E\big[\psi_i(\tau, \eta, \varphi) \mid W_i\big] \\ - \epsilon\, E\Big[h_\pi(W_i)\, \omega_i \;W_i\Big] \\ - \epsilon\, E\Big[(Z_i - \pi(W_i))\big(h_m(W_i) - \tau\, h_q(Z_i, W_i)\big) \;W_i\Big].$$

By Neyman orthogonality (from Lemma 1), the derivative with respect to η vanishes at $\eta = \eta_0$ (with φ fixed at φ_0), thus

$$\frac{\partial}{\partial \epsilon} E\big[\psi_i(\tau, \eta + \epsilon h, \varphi) \mid W_i\big]\Big|_{\epsilon=0} \;=\; 0.$$

This implies that first-order perturbations in η do not bias the moment condition, confirming Neyman orthogonality for the Adaptive DML-IV estimator. □

B.5 Proof of Lemma 3

The proof follows directly from Lemmas 1 and 2, extended to the setting of advanced ML-IV models.

I expand the expectation of the moment function around the true parameter values (η_0, φ_0)

with small perturbations $\Delta\eta = \hat{\eta} - \eta_0$ and $\Delta\varphi = \hat{\varphi} - \varphi_0$. Specifically, consider

$$\begin{aligned}
E\big[\psi_i(\tau, \eta_0 + \Delta\eta, \varphi_0 + \Delta\varphi)\big] \approx E\big[\psi_i(\tau, \eta_0, \varphi_0)\big] \\
&+ (\Delta\eta)^\top \left.\frac{\partial E[\psi_i(\tau, \eta, \varphi)]}{\partial \eta}\right|_{(\eta_0, \varphi_0)} \underbrace{= 0}_{\text{by RC6}} \\
&+ (\Delta\varphi)^\top \left.\frac{\partial E[\psi_i(\tau, \eta, \varphi)]}{\partial \varphi}\right|_{(\eta_0, \varphi_0)} \underbrace{\approx 0}_{\text{by RC6}} \\
&+ \frac{1}{2} \begin{pmatrix} \Delta\eta \\ \Delta\varphi \end{pmatrix}^\top \left.\frac{\partial^2 E[\psi_i(\tau, \eta, \varphi)]}{\partial (\eta, \varphi)^2}\right|_{(\eta_0, \varphi_0)} \begin{pmatrix} \Delta\eta \\ \Delta\varphi \end{pmatrix} + \text{ higher-order terms.}
\end{aligned}$$

The first-order terms vanish due to RC6, which guarantees that

$$\left.\frac{\partial E[\psi_i(\tau, \eta, \varphi)]}{\partial \eta}\right|_{(\eta_0, \varphi_0)} = 0 \quad \text{and} \quad \left.\frac{\partial E[\psi_i(\tau, \eta, \varphi)]}{\partial \varphi}\right|_{(\eta_0, \varphi_0)} \approx 0.$$

This vanishing implies that small perturbations in η and φ do not affect the expectation of the moment function to first order. □

B.6 Proof of Theorem 3

Define the sample moment

$$M_n(\tau, \eta) = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\tau, \eta).$$

Fix τ and let $h = (\hat{q} - q,\ \hat{m} - m)$. By a two-term Gateaux-Taylor expansion in the nuisance directions (Lemma 1 and Lemma 2),

$$M_n(\tau, \hat{\eta}) = M_n(\tau, \eta_0) + D_q M_n(\tau, \eta_0)[\, \hat{q} - q\,] + D_m M_n(\tau, \eta_0)[\, \hat{m} - m\,] + R_n,$$

where $R_n = o_p(n^{-1/2})$.

But by definition of the true nuisances,

$$M_n(\tau, \eta_0) = \frac{1}{n} \sum (Z_i - \pi(W_i))\big[Y_i - \tau D_i - m(W_i) + \tau q(Z_i, W_i)\big] = 0,$$

while

$$D_q M_n(\tau,\eta_0)[\hat{q}-q] = \tau\,\frac{1}{n}\sum (Z_i - \pi(W_i))\big[\hat{q}(Z_i,W_i) - q(Z_i,W_i)\big],$$

which is $o_p(n^{-1/2})$ if $\hat{q} \to q$ in L^2, and

$$D_m M_n(\tau,\eta_0)[\hat{m}-m] = -\frac{1}{n}\sum (Z_i - \pi(W_i))\big[\hat{m}(W_i) - m(W_i)\big],$$

which is $o_p(n^{-1/2})$ if $\hat{m} \to m$ in L^2. Hence in either case $M_n(\tau,\hat{\eta}) = o_p(n^{-1/2})$. Any $\sqrt{n}$–consistent root $\hat{\tau}$ of the equation $M_n(\hat{\tau},\hat{\eta}) = 0$ must then satisfy $\hat{\tau} \xrightarrow{p} \tau$, establishing double robustness. □

B.7 Proof of Theorem 4

By Theorem 3, the estimator $\hat{\tau}_{\text{ML-IV}}$ remains consistent under either a correct treatment or a correct outcome nuisance model. Theorems 2 and 3 (together with Lemmas 1 and 2) further ensure that errors in nuisance-function estimation affect $\hat{\tau}_{\text{ML-IV}}$ only at higher order, so that $\hat{\tau}_{\text{ML-IV}}$ is asymptotically unbiased. Under these properties (and standard rate assumptions), a Taylor expansion plus the Central Limit Theorem yields consistency and asymptotic normality with variance V.

More concretely, using the Law of Large Numbers (LLN) and consistency of the nuisance estimators:

$$\frac{1}{n}\sum_{i=1}^n \hat{w}(Z_i,W_i)\,Y_i \;\xrightarrow{p}\; E[w(Z_i,W_i)\,Y_i], \quad \frac{1}{n}\sum_{i=1}^n \hat{w}(Z_i,W_i)\,D_i \;\xrightarrow{p}\; E[w(Z_i,W_i)\,D_i].$$

Given the moment condition $E[w(Z_i,W_i)(Y_i - \tau D_i)] = 0$, I obtain

$$E\big[w(Z_i,W_i)\,Y_i\big] \;=\; \tau\,E\big[w(Z_i,W_i)\,D_i\big],$$

which implies

$$\hat{\tau}_{\text{ML-IV}} \;=\; \frac{\frac{1}{n}\sum_{i=1}^n \hat{w}(Z_i,W_i)\,Y_i}{\frac{1}{n}\sum_{i=1}^n \hat{w}(Z_i,W_i)\,D_i} \;\xrightarrow{p}\; \tau.$$

Thus, consistency is established.

For asymptotic normality, let $\psi_i(\tau,\eta)$ be the orthogonal score. A first-order Taylor expansion of the sample moment condition around τ gives

$$0 \;=\; \frac{1}{n}\sum_{i=1}^{n}\psi_i\big(\hat\tau_{\text{ML-IV}},\hat\eta\big) \;\approx\; \frac{1}{n}\sum_{i=1}^{n}\psi_i(\tau,\eta) \;+\; \Big(\tfrac{1}{n}\sum_{i=1}^{n} w(Z_i,W_i)\,D_i\Big)\,\big(\hat\tau_{\text{ML-IV}}-\tau\big)+o_p(n^{-1/2}).$$

Rearranging terms yields

$$\hat\tau_{\text{ML-IV}}-\tau \;=\; \frac{\frac{1}{n}\sum_{i=1}^{n} w(Z_i,W_i)\,\big(\delta A_i+\phi_i\big)}{\frac{1}{n}\sum_{i=1}^{n} w(Z_i,W_i)\,D_i} \;+\; o_p\big(n^{-1/2}\big).$$

By the Central Limit Theorem (CLT),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(Z_i,W_i)\,\big(\delta A_i+\phi_i\big) \;\xrightarrow{d}\; \mathcal{N}(0,\,\sigma^2),$$

with

$$\sigma^2 \;=\; \operatorname{Var}\Big(w(Z_i,W_i)\,\big(\delta A_i+\phi_i\big)\Big),$$

and the denominator converges in probability to $E[w(Z_i,W_i)\,D_i]$. Therefore,

$$\begin{aligned}\sqrt{n}\,\big(\hat\tau_{\text{ML-IV}}-\tau\big) &= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w(Z_i,W_i)\,\big(\delta A_i+\phi_i\big)}{E\big[w(Z_i,W_i)\,D_i\big]} \;+\; o_p(1),\\ &\xrightarrow{d}\; \mathcal{N}\Big(0,\;\frac{\sigma^2}{\big(E[w(Z_i,W_i)\,D_i]\big)^2}\Big).\end{aligned} \tag{14}$$

Hence, $\hat\tau_{\text{ML-IV}}$ is asymptotically normal with variance $V=\frac{\sigma^2}{\big(E[w(Z_i,W_i)\,D_i]\big)^2}$ as given in (13).

□

B.8 Proof of Lemma 6

Define the empirical score

$$\hat\psi_i(\tau)=\big(Z_i-\hat\pi(W_i)\big)\Big[Y_i-\hat m(W_i)-\tau\big(D_i-\hat q(Z_i,W_i)\big)\Big].$$

The cross-fitted estimator $\hat{\tau}$ solves $n^{-1}\sum_{i=1}^{n}\hat{\psi}_i(\hat{\tau}) = 0$. A first-order Taylor expansion around the true parameter gives

$$0 = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i(\tau) + (\hat{\tau}-\tau)\,\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{\psi}_i(\tau)}{\partial\tau} + o_p\big(n^{-1/2}\big),$$

because $|\hat{\tau}-\tau| = O_p(n^{-1/2})$ by Theorem 4. The empirical derivative satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{\psi}_i(\tau)}{\partial\tau} = -\frac{1}{n}\sum_{i=1}^{n}\big(Z_i-\hat{\pi}(W_i)\big)\big(D_i-\hat{q}(Z_i,W_i)\big) \xrightarrow{p} -S,$$

by consistency of $\hat{\pi}$ and $\hat{q}$ and the law of large numbers.

Decompose the sample mean of the score at τ as

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i(\tau) = \frac{1}{n}\sum_{i=1}^{n}\psi_i(\tau) + R_n,$$

where R_n denotes the plug-in error. Lemma 5 shows $R_n = o_p(n^{-1/2})$. Solving for $\hat{\tau}-\tau$ therefore yields

$$\hat{\tau}-\tau = \frac{\frac{1}{n}\sum_{i=1}^{n}\psi_i(\tau)}{S} + o_p\big(n^{-1/2}\big), \qquad \psi_i(\tau) = \big(Z_i-\pi(W_i)\big)\big(Y_i-\tau D_i-m(W_i)+\tau q(Z_i,W_i)\big).$$

Because $m(W_i) = E[Y_i-\tau D_i \mid W_i]$ and $q(Z_i,W_i) = E[D_i \mid Z_i,W_i]$, $\psi_i(\tau)$ simplifies to $w(Z_i,W_i)\big(Y_i-\tau D_i\big)$. The central limit theorem then gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\tau) \xrightarrow{d} \mathcal{N}\big(0,\sigma^2\big), \quad \sigma^2 = \mathrm{Var}\big(w(Z_i,W_i)(\delta A_i+\phi_i)\big),$$

so $\sqrt{n}(\hat{\tau}-\tau)$ converges to $\mathcal{N}\big(0,\sigma^2/S^2\big)$, which equals V.

Lemma 4 places Ψ_i in the closure of the tangent space, whence no regular estimator can have asymptotic variance smaller than $\mathrm{Var}(\Psi_i)$. Consequently V is the semiparametric efficiency bound. □No data related to finr˙ a

Appendix C Incorporating Tertiary Parameters $\varphi = (\lambda, \theta)$ into Neyman Orthogonality

In this Appendix I show that the Neyman-orthogonal score

$$\psi_i(\tau, \eta, \varphi) \;=\; \big(Z_i - \pi(W_i;\varphi)\big)\,\big[Y_i - \tau\, D_i - m(W_i;\varphi) + \tau\, q(Z_i, W_i;\varphi)\big]$$

remains orthogonal not only to the usual nuisance functions $\eta = (\pi, q, m)$, but also to any data-driven choice of tertiary parameters φ. Here $\varphi = (\lambda, \theta)$ is allowed to index regularization parameters (e.g. Lasso/Ridge penalty λ, tree-depth penalty, network-dropout rate), feature importance–based dimension-reduction thresholds (e.g. a cutoff on an importance score that prunes low-weight covariates), and automated model-selection decisions (e.g. a discrete choice of "use Random Forest vs. XGBoost vs. neural net," plus their associated tuning parameters).

In other words, a single φ-vector can encode (i) which learner is chosen, (ii) which subset of features were retained based on importance scores, and (iii) the associated regularization/hyperparameter choices for that learner. The goal is to prove that, as long as the final $\hat{\varphi}$ converges to some φ_0 at a sufficiently fast rate and each corresponding nuisance estimate converges at $o_p(n^{-1/4})$, then

$$\frac{\partial}{\partial \varphi}\, E\big[\psi_i(\tau, \eta_0, \varphi)\big]\Big|_{\varphi=\varphi_0} \;=\; 0,$$

and the mixed second derivative $\partial^2 E[\psi]/(\partial\varphi\,\partial\eta)$ remains bounded. This guarantees that no matter how I built the ML pipeline—regularization, feature-importance pruning, or a cross-validated model-search tree—the first-order bias from tuning/model-selection disappears. The cross-fitting in the main text then ensures that hyperparameter search and automated model selection enter only at $o_p(n^{-1/2})$.

Assumption A.1 (Uniform Convergence & Cross-Fitting Independence). There exist deterministic sequences $\delta_n = o(n^{-1/4})$ and $\epsilon_n = o(n^{-1/2})$ such that, uniformly over a neighborhood of φ_0,

$$\|\hat{q} - q_0\|_{L^2},\ \|\hat{\pi} - \pi_0\|_{L^2},\ \|\hat{\mu} - \mu_0\|_{L^2},\ \|\hat{r} - r_0\|_{L^2} \;\le\; \delta_n,$$

$$\|\hat{q} - q_0\|_{\infty},\ \|\hat{\pi} - \pi_0\|_{\infty},\ \|\hat{\mu} - \mu_0\|_{\infty},\ \|\hat{r} - r_0\|_{\infty} \;\le\; \epsilon_n,$$

and $\hat{\varphi} \xrightarrow{p} \varphi_0$. Moreover, for each observation i, the nuisance estimators $\hat{\eta}_i$ and the chosen hyperparameters $\hat{\varphi}_i$ (trained on data excluding i) are independent of (Y_i, D_i, Z_i, W_i).

Assumption A.2 (Parametric Path & Gateaux Differentiability in φ). The hyperparameter φ lies in an open, convex subset $\Phi \subset \mathbb{R}^p$. For each $\varphi \in \Phi$, define nuisance maps

$$\pi(\,\cdot\,;\varphi), \quad q(\,\cdot\,,\cdot\,;\varphi), \quad \mu(\,\cdot\,;\varphi), \quad r(\,\cdot\,;\varphi),$$

so that at the true φ_0,

$$\pi_0(W) = \pi(W;\varphi_0) = E[\,Z \mid W], \qquad q_0(Z,W) = q(Z,W;\varphi_0) = E[\,D \mid Z,W],$$

$$\mu_0(W) = \mu(W;\varphi_0) = E[\,Y \mid W], \qquad r_0(W) = r(W;\varphi_0) = E[\,D \mid W], \qquad m_0(W) = \mu_0(W) - \tau\, r_0(W).$$

For each fixed (z,w), the functions $\varphi \mapsto \pi(w;\varphi)$, $q(z,w;\varphi)$, $\mu(w;\varphi)$, $r(w;\varphi)$ are Gateaux-differentiable at φ_0. Denote their directional derivatives by

$$D_\varphi \pi(w;\varphi_0)[h], \quad D_\varphi q(z,w;\varphi_0)[h], \quad D_\varphi \mu(w;\varphi_0)[h], \quad D_\varphi r(w;\varphi_0)[h],$$

for any direction $h \in \mathbb{R}^p$. Moreover, there exists a finite constant C such that for every unit direction h,

$$\sup_w |D_\varphi \pi(w;\varphi_0)[h]| + \sup_{(z,w)} |D_\varphi q(z,w;\varphi_0)[h]| + \sup_w |D_\varphi \mu(w;\varphi_0)[h]| + \sup_w |D_\varphi r(w;\varphi_0)[h]| \;\le\; C.$$

In particular, this covers (a) any continuous regularization parameter λ; (b) any continuous feature-importance threshold or feature-pruning decision that can be approximated by a smooth path; and (c) any smooth weight assignment among a small discrete set of candidate learners, provided it can be embedded in a continuous parameterization.

Assumption A.3 (Second-Order Remainder Control). For any direction $h \in \mathbb{R}^p$, define the perturbed nuisance functions

$$\pi_t(w) = \pi\big(w;\varphi_0 + t\,h\big), \quad q_t(z,w) = q\big(z,w;\varphi_0 + t\,h\big), \quad \mu_t(w) = \mu\big(w;\varphi_0 + t\,h\big), \quad r_t(w) = r\big(w;\varphi_0 + t\,h\big),$$

and set $m_t(w) = \mu_t(w) - \tau\, r_t(w)$. There exists some $\eta > 0$ and a function $R(t)$ such that

$\lim_{t\to 0} R(t)/t = 0$ and, for all sufficiently small $|t|$,

$$\sup_{(z,w)} \left| q_t(z,w) - q_0(z,w) - D_\varphi q(z,w;\varphi_0)[t\,h] \right| \;\le\; R(t),$$

and similarly for π_t, μ_t, r_t. Equivalently, each Gateaux expansion has a second-order remainder $o(t)$.

Lemma 7 (Neyman Orthogonality in φ). Under Assumptions A.1–A.3, define the population moment

$$M(\tau,\varphi) \;=\; E\Big[\psi_i\big(\tau,\eta_0,\varphi\big)\Big] \;=\; E\Big[\big(Z-\pi(W;\varphi)\big)\,\big\{Y-\tau D - m(W;\varphi) + \tau\, q(Z,W;\varphi)\big\}\Big].$$

Then for any direction h,

$$\frac{\partial}{\partial t} E\Big[\psi_i\big(\tau,\eta_0,\varphi_0 + t\,h\big)\Big]\Big|_{t=0} \;=\; 0,$$

and the mixed derivative $\partial^2 E[\psi]/(\partial\varphi\,\partial\eta)\big|_{(\eta_0,\varphi_0)}$ is bounded. Consequently, $\psi_i(\tau,\eta,\varphi)$ is Neyman orthogonal in φ at φ_0.

Proof of Lemma 7. Fix a direction $h \in \mathbb{R}^p$. For small t, let

$$\pi_t(w) = \pi\big(w;\varphi_0{+}t\,h\big), \quad q_t(z,w) = q\big(z,w;\varphi_0{+}t\,h\big), \quad \mu_t(w) = \mu\big(w;\varphi_0{+}t\,h\big), \quad r_t(w) = r\big(w;\varphi_0{+}t\,h\big),$$

and $m_t(w) = \mu_t(w) - \tau\, r_t(w)$. Define

$$G(t) = \; E\Big[\big(Z-\pi_t(W)\big)\,\big\{Y-\tau D - m_t(W) + \tau\, q_t(Z,W)\big\}\Big].$$

Icompute $\frac{dG(t)}{dt}\big|_{t=0}$. Observe that

$$Y-\tau D - m_t(W) + \tau\, q_t(Z,W) = \;\{Y-\tau D - \mu_t(W) + \tau\, r_t(W)\} + \tau\,\{q_t(Z,W) - r_t(W)\}.$$

Hence

$$\begin{aligned}\frac{dG(t)}{dt}\Big|_{t=0} \;=\;& E\Big[-D_\varphi \pi(W;\varphi_0)[h]\;\{Y-\tau D - m_0(W) + \tau\, q_0(Z,W)\}\Big]\\ &+\; E\Big[\big(Z-\pi_0(W)\big)\,\big\{-D_\varphi \mu(W;\varphi_0)[h] + \tau\, D_\varphi r(W;\varphi_0)[h] + \tau\, D_\varphi q(Z,W;\varphi_0)[h]\big\}\Big],\end{aligned}$$

where

$$m_0(W) = \mu_0(W) - \tau\, r_0(W), \quad q_0(Z,W) = E[D \mid Z,W], \quad \pi_0(W) = E[Z \mid W].$$

Let

$$R = Y - \tau D - \mu_0(W) + \tau\, r_0(W), \quad U = q_0(Z,W) - r_0(W).$$

By definition of the true nuisances, $E[R \mid W] = 0$ and $E[U \mid W] = 0$. Therefore,

$$E\big[D_\varphi \pi(W;\varphi_0)[h]\ R\big] = E\big[E[\ D_\varphi \pi(W;\varphi_0)[h]\ R \mid W]\big] = 0, \quad E\big[D_\varphi \pi(W;\varphi_0)[h]\ U\big] = 0.$$

Thus the first term vanishes. For the second term,

$$E\big[(Z - \pi_0(W))\, D_\varphi \mu(W;\varphi_0)[h]\big] = E\big[E[\ (Z - \pi_0(W))\, D_\varphi \mu(W;\varphi_0)[h] \mid W]\big] = 0,$$

and similarly $E[(Z - \pi_0(W))\, D_\varphi r(W;\varphi_0)[h]] = 0$. Finally,

$$\begin{aligned} E\big[(Z - \pi_0(W))\, D_\varphi q(Z,W;\varphi_0)[h]\big] &= E\big[E[\ (Z - \pi_0(W))\, D_\varphi q(Z,W;\varphi_0)[h] \mid W]\big] \\ &= E\big[D_\varphi r(W;\varphi_0)[h] - \pi_0(W)\, D_\varphi r(W;\varphi_0)[h]\big] = 0, \end{aligned}$$

since $E[D_\varphi q(Z,W;\varphi_0)[h] \mid W] = D_\varphi E[D \mid W][h] = D_\varphi r(W;\varphi_0)[h]$. This shows $\frac{dG(t)}{dt}\big|_{t=0} = 0$. Combined with the facts that $\frac{\partial}{\partial \eta} E[\psi_i(\tau,\eta,\varphi_0)]\Big|_{\eta_0} = 0$ and that the mixed second derivatives remain bounded under A.2–A.3, Iconclude that $\psi_i(\tau,\eta,\varphi)$ is orthogonal in φ at φ_0. □

By merging the standard Gateaux-orthogonality in η (e.g. Chernozhukov et al., 2018) with Lemma A.1 above, I obtain

$$\frac{\partial}{\partial \eta} E[\psi_i(\tau,\eta,\varphi)]\Big|_{(\eta_0,\varphi_0)} = \frac{\partial}{\partial \varphi} E[\psi_i(\tau,\eta,\varphi)]\Big|_{(\eta_0,\varphi_0)} = 0,$$

and the mixed partial $\partial^2 E[\psi]/(\partial\varphi\, \partial\eta)\big|_{(\eta_0,\varphi_0)}$ is bounded. Under cross-fitting (Assumption A.1), any first-order error in $\hat\eta$ or $\hat\varphi$ enters the sample analog of the moment only at $o_p(n^{-1/2})$. Therefore, a fully data-driven pipeline—incorporating robust regularization, feature-importance–based dimension reduction, and automated model selection (all subsumed under φ)—still yields a $\sqrt{n}$-consistent, asymptotically normal, and semiparametrically efficient DML–IV estimator. □

Appendix D Joint Identification of (τ, β)

Under the IV-exogeneity and completeness assumptions of Section 2.2, the pair (τ, β) is uniquely identified by the moment condition

$$E\Big[(Z_i - \pi(W_i))\,(Y_i - \tau D_i - \beta^\top W_{i,\mathrm{ld}} - m(W_i))\Big] = 0.$$

In particular, if (τ, β) and (τ', β') both satisfy

$$E\Big[(Z_i - \pi(W_i))\big((\tau - \tau')D_i + (\beta - \beta')^\top W_{i,\mathrm{ld}}\big)\Big] = 0,$$

then completeness of the conditional law of $(D_i, W_{i,\mathrm{ld}})$ given (Z_i, W_i) implies $\tau = \tau'$ and $\beta = \beta'$.

*Proof of **Appendix D**.* Define the deviation

$$h(D_i, W_{i,\mathrm{ld}}) \;=\; \Delta\tau\, D_i \;+\; \Delta\beta^\top W_{i,\mathrm{ld}},$$

where $\Delta\tau = \tau - \tau'$ and $\Delta\beta = \beta - \beta'$. From the two solutions (τ, β) and (τ', β') both satisfying

$$E\Big[(Z_i - \pi(W_i))\,(Y_i - \tau D_i - \beta^\top W_{i,\mathrm{ld}} - m(W_i))\Big] \;=\; 0,$$

I subtract and obtain

$$E\big[(Z_i - \pi(W_i))\, h(D_i, W_{i,\mathrm{ld}})\big] \;=\; 0.$$

Since $\pi(W_i) = E[Z_i \mid W_i]$, an application of the law of iterated expectations gives

$$0 \;=\; E\Big[\,E\big[(Z_i - \pi(W_i))\, h(D_i, W_{i,\mathrm{ld}}) \,\big|\, W_i\big]\Big] \;=\; E\big[\mathrm{Cov}\big(Z_i,\, h(D_i, W_{i,\mathrm{ld}}) \,\big|\, W_i\big)\big].$$

But observe that

$$\mathrm{Cov}\big(Z_i,\, h(D_i, W_{i,\mathrm{ld}}) \,\big|\, W_i = w\big) \;=\; E\big[(Z_i - \pi(w))\, h(D_i, W_{i,\mathrm{ld}}) \,\big|\, W_i = w\big] \;=\; \big\langle\, h(\cdot),\, k_w(\cdot)\big\rangle_{L^2(D, W_{ld}|w)},$$

where

$$k_w(d, w_{ld}) \;=\; f_{D, W_{ld}|Z,W}(d, w_{ld} \mid z, w)\,(z - \pi(w))$$

is the kernel of the linear operator $T_w : L^2\big(D, W_{ld} \mid w\big) \to L^2(Z \mid w)$,

$$(T_w f)(z) \;=\; \int f(d, w_{ld})\, f_{D,W_{ld}|Z,W}(d, w_{ld} \mid z, w)\, d(d, w_{ld}).$$

Completeness of the conditional law of $(D_i, W_{i,\text{ld}})$ given (Z_i, W_i) is exactly the statement that for each w, the operator T_w is injective:

$$T_w f = 0 \quad \Longrightarrow \quad f = 0 \quad \text{a.s.}$$

Since $\text{Cov}(Z, h \mid W = w) = 0$ for almost every w, I have

$$\big\langle\, h(\cdot),\, k_w(\cdot) \big\rangle = 0 \quad \Longrightarrow \quad h(d, w_{ld}) = 0 \quad \text{for almost every } (d, w_{ld}),$$

by injectivity of T_w. Hence $\Delta\tau\, D + \Delta\beta^\top W_{ld} = 0$ almost surely. Finally, because the vector (D, W_{ld}) has nondegenerate support, the only linear combination that vanishes a.s. is the trivial one: $\Delta\tau = 0$ and $\Delta\beta = 0$. This proves uniqueness of (τ, β). □

Remark 2. If one does not require explicit interpretation of β, simply redefine

$$\widetilde{m}(W_i) \;=\; m(W_i) + \beta^\top W_{i,\text{ld}},$$

drop the linear term from the structural equation, and estimate τ from

$$E\Big[(Z_i - \pi(W_i))\,(Y_i - \tau D_i - \widetilde{m}(W_i))\Big] = 0.$$

The same completeness argument then identifies τ without reference to β. Thus β is included only for interpretability and does not affect identification of τ.

□

Appendix E Supplementary Monte Carlo Results

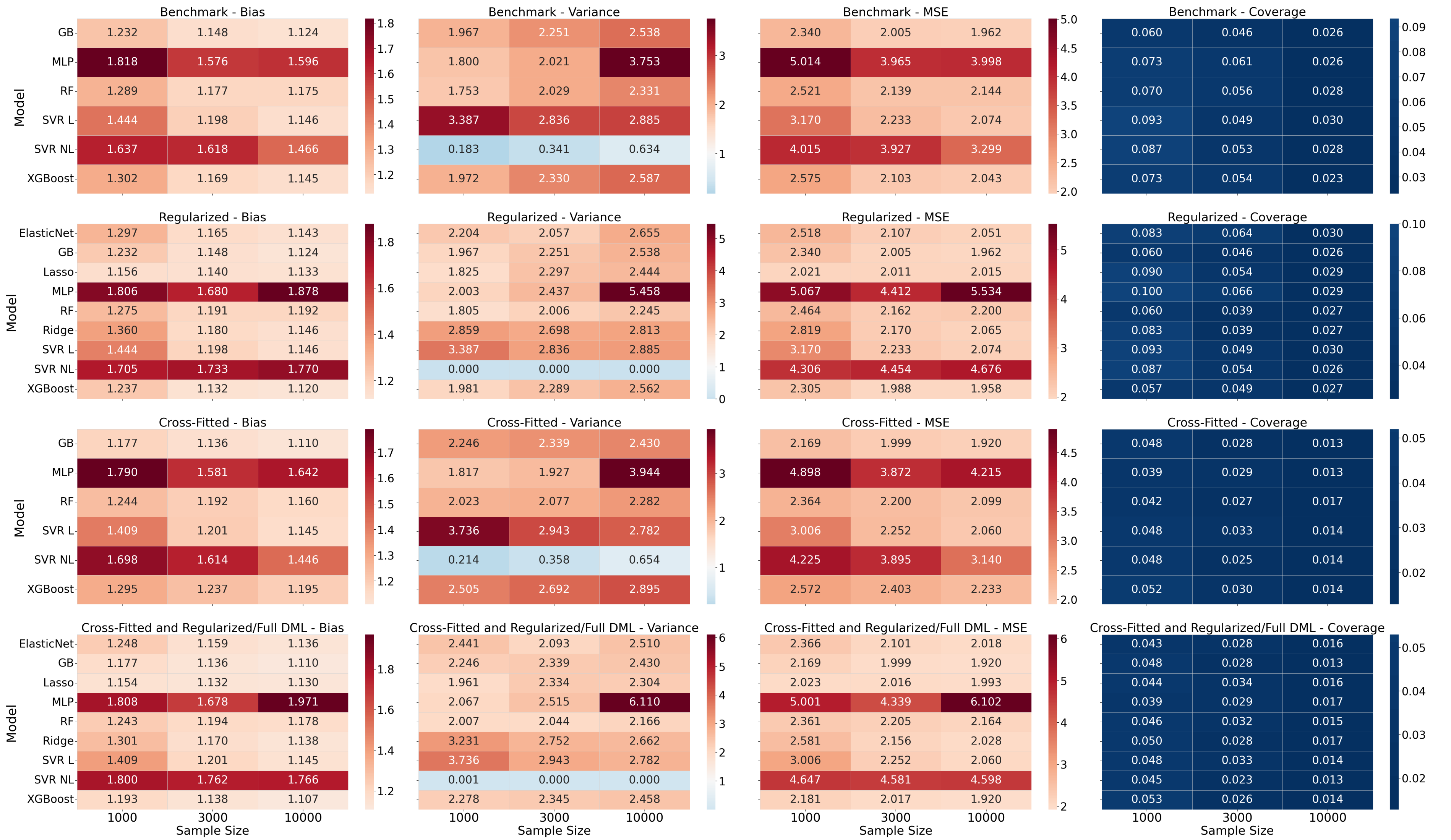

Figure 3: Metrics from the First-Stage Predictions

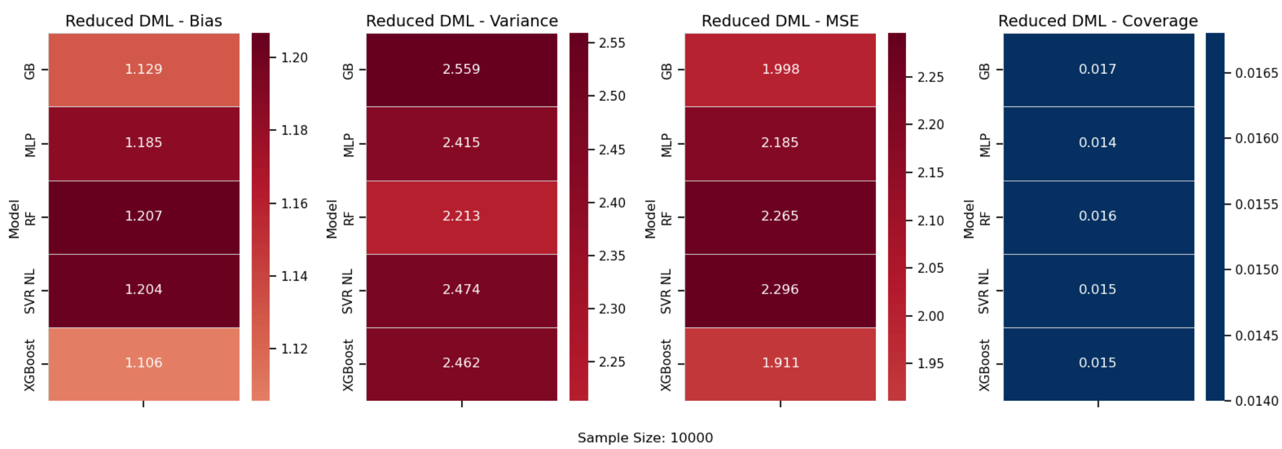

Figure 4: Metrics from the First-Stage Predictions (Reduced DML)